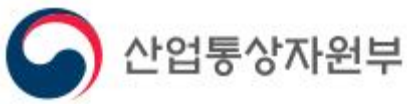
공공데이터 활용 빅데이터 분석

한국수출액 예측을 통한 수출 유망국가와 품목 추천

앙상블 예측 모델 개발과
군집화(Clustering) 기반 추천



김성범, 배준호

2021. 08. 10.

목 차

1. 분석 개요

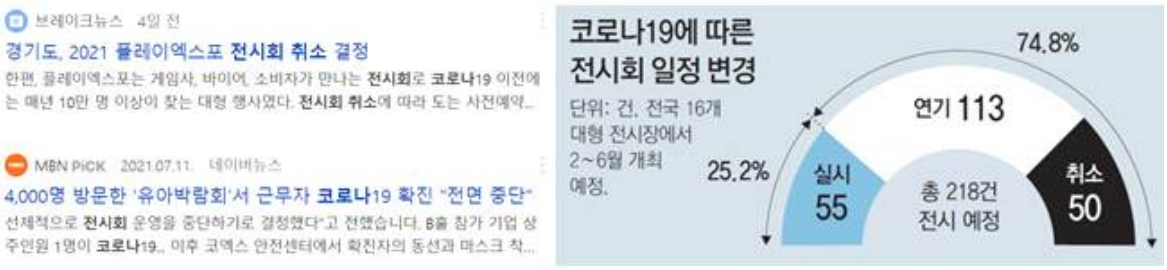
가. 분석 배경 및 문제 정의

1) 대한무역투자진흥공사(KOTRA) 활동 현황

□ 코로나 여파로 각종 전시회, 상담회가 취소되어 수출 기업의 고충 가중

□ KOTRA는 이를 극복하기 위해 디지털 서비스 기반의 수출 기업 지원 도모

□ 현재 KOTRA 빅데이터 플랫폼에서는 품목별 유망시장을 서비스 中



[그림 1-1] 오프라인 전시회 취소 관련 기사

2) 수출 기업의 해외 시장 정보획득 한계 발생

□ 코로나로 인해 해외에서의 입국과 정상적인 전시회 진행이 어려운 상황

□ 현재 여건상 수출 상담, 해외 바이어 미팅 등 해외 시장 정보를 얻을 수 있는 창구가 대부분 단절

□ 해외에 공식 대리점을 보유하고 있는 대기업과 달리, 적극적인 영업 활동이 필요한 중소·중견 기업은 오프라인 전시회 취소에 치명적 영향

3) 한국수출액 예측을 통한 수출 유망국가 및 품목 추천

□ 국내 중소·중견 기업의 수출을 빅데이터 분석을 통해 지원하는 것이 목적

□ 2018년 데이터로 특정 국가의 특정 품목을 한국에서 얼마나 수입할지 예측

□ 예측된 데이터를 통해 수출 기업이 수출할 시장을 탐색에 도움 제공

[그림 1-2] 수출 관련 빅데이터 분석 관련 기사 (자료: 뉴스원)

나. 분석 프로세스 제시

[그림 1-3] 앙상블 예측 모델 개발 프로세스

1) 디지털 무역투자 생태계를 위한 빅데이터 분석

다음 프로세스를 기반으로 기존 데이터와 추가 데이터를 병합하고 데이터 탐색과 전처리의 단계를 거쳐 앙상블 예측 모델을 개발하고자 한다. 결론을 통해 모델 해석과 시사점을 확인하고, 활용 방안과 기대효과, 한계점 및 제언을 하고자 한다. 이러한 빅데이터 기반의 모델 구현을 통해 디지털 무역투자 비즈니스 생태계 구축을 이루고자 한다.

2) 한국수출액 예측을 통한 수출 유망국가 및 품목 추천분석

[그림 1-4] 수출 유망국가 및 품목 추천 프로세스

개발한 앙상블 예측 모델을 통해 다음연도의 한국수출액을 예측한 뒤, 유사한 데이터끼리 같은 그룹에 속하게 묶는 군집화(Clustering)를 통해 수출 유망국가 및 품목을 탐색하고자 한다. 이를 통해 해외 무역에 어려움을 겪는 중소·중견 기업에 수출 확대와 기업 성장의 기회를 제공하고자 한다.

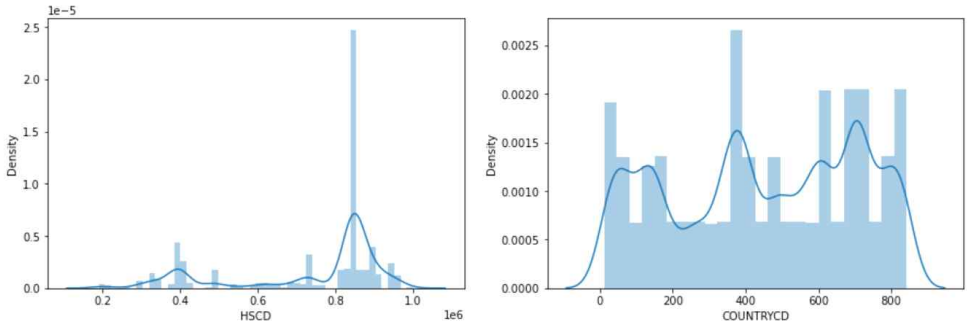
2. 데이터 탐색 및 전처리

가. 데이터 이해와 탐색

1) 데이터 파악

□ 2017년, 2018년 수출 데이터가 각각 학습용, 예측용 데이터로 구성

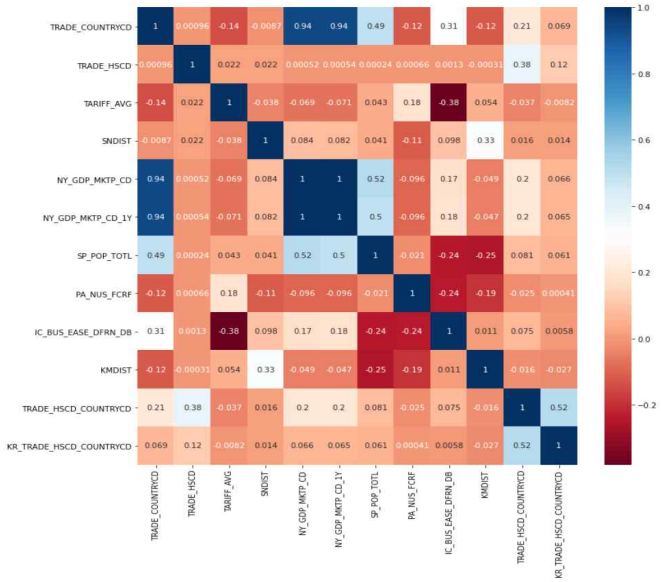
□ 두 데이터의 구조가 상당히 유사하여, 학습데이터의 이해만 서술



[그림 2-1] 데이터 분포의 이해

학습 데이터의 품목(HSCD) 개수는 총 500개이며, 국가의 개수는 총 43개로 확인하였으며, 특정 품목 및 국가에 대한 편향된 분포는 없는 것으로 확인하였다.

2) 변수 간 상관관계 파악



[그림 2-2] 변수 간 상관계수 확인

해당 국가의 해당 품목 수입금액을 나타내는 TRADE_COUNTRYCD 변수와 GDP 기반의 변수인 NY_GDP_MKTP_CD, NY_GDP_MKTP_CD_1Y와의 양의 상관관계가 높게 나타나는 것을 확인하였다.

3) 결측치 (Missing Value) 파악

□ 학습 데이터와 예측 데이터의 결측치 확인

결측치 칼럼	칼럼 설명	결측치 개수	결측치 비율
TARIFF_AVG	해당 국가에서 해당 품목에 적용되는 평균 관세율	129	0.61%
SNDIST	해당 국가와 수입 국가 간 평균 거리	22	0.1%
PA_NUS_FCRF	미국 달러에 대한 현지 통화 단위, 월평균을 기준으로 한 연평균	3488	16.46%
TRADE_HSCD_COUNTRYCD	해당 연도 해당 국가의 해당 품목 수입금액	21	0.1%

[표 1-1] 학습 데이터의 결측치 개수와 비율

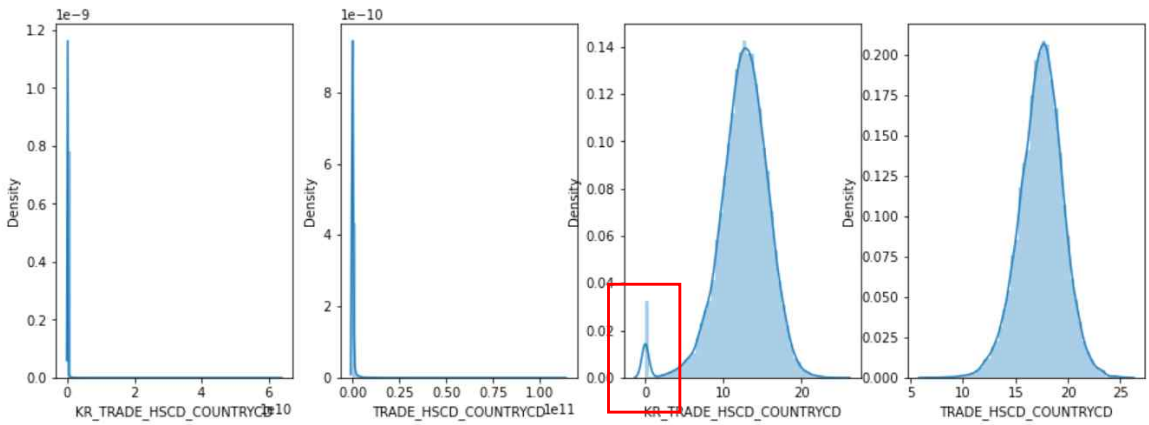
결측치 칼럼	칼럼 설명	결측치 개수	결측치 비율
TARIFF_AVG	해당 국가에서 해당 품목에 적용되는 평균 관세율	625	2.95%
SNDIST	해당 국가와 수입 국가 간 평균 거리	3	0.01%
PA_NUS_FCRF	미국 달러에 대한 현지 통화 단위, 월평균을 기준으로 한 연평균	3488	16.46%
TRADE_HSCD_COUNTRYCD	해당 연도 해당 국가의 해당 품목 수입금액	3	0.01%

[표 1-2] 예측 데이터의 결측치(R_TRADE_HSCD_COUNTRYCD 결측치 제외)

PA_NUS_FCRF는 3,488건의 결측의 개수와 16%가 넘는 비율로 결측치를 대체해야 할 필요성이 있다. 그 외 변수의 경우 결측 비율이 전체의 최대 3%를 넘지 않지만, 모델의 범용성을 위해 다른 변수의 관계를 이용하는 함수 위주로 결측값을 대체하고자 한다.

4) 이상치 (Oultier Value) 파악

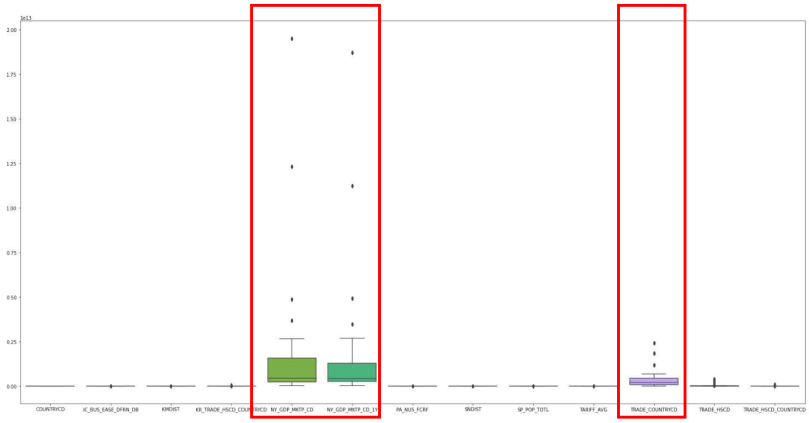
□ 로그(Log) 변환을 통한 이상치 확인



[그림 2-3] 이상치(Outlier) 확인 그래프

종속 변수 KR_TRADE_HSCD_COUNTRYCD와 TRADE_HSCD_COUNTRYCD에 대해 심한 왜도(Skewness) 형태를 확인하였다. 로그 변환 이후 정규분포의 값을 보이나, 로그 값이 0인 부분을 확인할 수 있다. 이는 향후 데이터 모델링 부분에서 전처리하고자 한다.

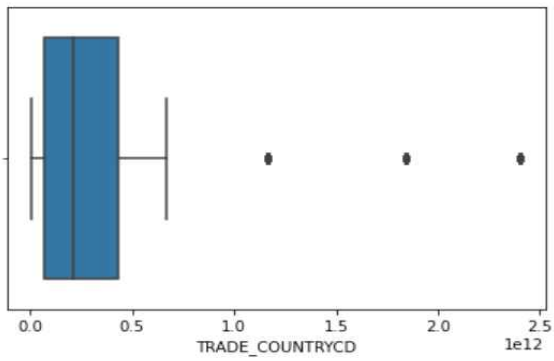
□ 사분위수 범위(IQR)를 통한 이상치 확인



[그림 2-4] IQR를 활용한 이상치 확인

Q1-(1.5*IQR)미만 Q3+(1.5*IQR) 초과하는 값을 Turkey Fences 함수를 이용해 파악한 결과, 총 3개의 변수에서 이상치가 검출되었으며, Boxplot을 통해 시각적으로 이상치를 확인하였다. NY_GDP_MKTP_CD와 NY_GDP_MKTP_CD_1Y의 경우 실질적으로 국가별 경제 규모 차이가 있기에 이상치로 판단하기 어렵고, TRADE_HSCD_COUNTRYCD의 변수에 이상치 확인 작업하고자 한다.

□ Z-Score를 통한 이상치 확인



[그림 2-5] Z-Score을 활용한 이상치 검출

TRADE_HSCD_COUNTRYCD에 대한 Z-Score 이상치 검출 결과, 총 994개가 검출되었다. 상세 확인 결과 '1,843,792,938,795', '2,405,276,626,657' 두 값이 반복적으로 검출됨을 확인하였으며, 향후 분석용 데이터 전처리에서 처리하고자 한다.

나. 결측치 처리

변수	데이터 설명	활용 데이터 및 처리 방식
PA_NUS_FCRF	국가별 1년 환율 평균	Exchange Rates Domestic Currency per U.S. Dollar, Period Average
TRADE_HSCD_COUNTRYCD	HS코드별 1년 총 수입액	Yearly Time Series, Imported Value by HS Code
TARIFF_AVG	관세	다른 변수를 이용하여 처리
SNDIST	품목을 수입한 국가 간 거리	

[표 2-1] 결측값 처리 활용 데이터 및 처리 방법

1) PA_NUS_FCRF 결측값 처리

□ 결측값 확인

- 국가별 결측치 -

Germany	500
Netherlands	500
Austria	499
France	499
Italy	499
Spain	498
Belgium	493

결측치 국가는 모두 유럽연합(EU)국가 중 유로(EURO)화를 사용한다는 공통점 파악하였다.

	Concept	2017	2018
0	National Currency per SDR,...	1.19000	1.21000
1	National Currency per SDR,...	1.23000	1.20000
2	Domestic Currency per U.S....	0.83000	0.87000
3	Domestic Currency per U.S....	0.88521	0.84677
4	Nominal Effective Exchange...	102.65000	108.02000

[그림 2-6] IMF 환율 데이터

□ 결측값 처리

IMF에서 제공하는 데이터를 통해 각 EU 국가의 US$ 환율 연평균 데이터(Exchange Rates Domestic Currency per U.S. Dollar, Period Average)를 추출하였고, 학습 데이터와 예측 데이터의 PA_NUS_FCRF 결측값에 각각 대체 수행하였다. (2017년 = 0.88521, 2018년 = 0.84677)

2) TRADE_HSCD_COUNTRYCD 결측값 처리

□ 결측값 확인

학습 데이터의 경우 10개 국가의 다양한 HS Code에서 결측치가 검출되었고, 예측 데이터의 경우 HS Code '999999' 에서만 3개의 결측치가 발생하였다. 해당 품목의 경우 종류가 정해지지 않은 상품을 의미하며, 일반적이지 않은 품목이기 때문에 파생변수 과정 이후에 처리하고자 한다.

□ 결측값 대체

ITC Trade Map 데이터를 통해 HS코드 별 1년 총 수입액 데이터를 확보하였다. 해당 데이터는 조회한 모든 연도에 품목에 대한 수입 기록이 없을 시 해당 품목 자체를 표시하지 않는 데이터를 보여주며, 조회 기간 중 한 번이라도 데이터가 존재할 경우 품목을 표시하되 0의 값을 표출하는 것을 확인하였다.

따라서 실제 수입 기록이 동시에 없는 경우에도 데이터가 없는 Null상태와 0의 값이 혼용되어 기재되는 특징을 가지는 것이다. 학습 데이터의 결측치 국가인 Kuwait, Saudi Arabia, Philippines, Mexico, Singapore, Egypt, Viet Nam, Iran, Hong Kong 10개국에 대해 실제 데이터와 대조하는 방식으로 결측치를 검정하였다.

그 결과 Kuwait를 제외한 9개 국가에서 발생한 결측값은 결측된 것이 아니라 실제 수입 기록이 없는 0의 값임을 확인하였고, Kuwait의 경우 품목에 따른 실제 수입액을 결측치에 대입하여 처리하였다.

3) TARIFF_AVG 결측값 처리

□ 결측값 확인

- HS Code별 결측치 -

852852	43
382499	43
999999	36
:	

- 예측 데이터 국가별 결측치 -

USA	500
Iran	5
United Kingdom	4
:	

TARIFF_AVG 결측값의 경우, 학습 데이터와 예측 데이터 모두 특정 3개의 품목에서 결측값이 주로 발생하였으며 예측 데이터는 USA에서 결측치가 500건 발생했음을 확인하였다.

□ 결측값 대체

$$\frac{\text{해당 품목을 구매한 국가들의 평균 관세} + \text{해당 국가의 평균 관세}}{2}$$

해당 변수에 대한 결측치는 추가 데이터가 아닌 기존 데이터를 이용하여 다음과 같은 식을 통하여 값을 대체하였다.

4) SNDIST 결측값 처리

□ 결측값 확인

- 학습 데이터 결측값 -	- 예측 데이터 결측값 -
결측 품목코드 개수 : 22	결측 품목코드 개수 : 3
결측 국가명 개수 : 11	결측 국가명 개수 : 3
해당 품목의 타겟 국가 개수 : 43	해당 품목의 타겟 국가 개수 : 36

SNDIST 변수의 경우 해당 품목에 대한 수입 국가 간 거리의 평균을 의미한다. 예를 들어, 태국이 Z품목을 미국, 프랑스, 인도에서만 수입했다면, 태국과 미국과의 거리 = A, 태국과 프랑스와의 거리 = B, 태국과 인도와의 거리 = C로 두고 SNDIST 변수의 식은 (A+B+C)/3으로 도출하고자 한다.

□ 결측값 대체

해당 변수에 대한 결측치는 실제 SNDIST 변수를 만들 때 사용되었던 CEPII의 데이터를 확보하여 해당 국가가 해당 품목을 수입한 국가 간 평균 거리를 구하여 대체하고자 하였다.

CEPII 추가 데이터에서 결측값이 있는 HSCD 품목별 국가에 대한 거리를 구하고, 해당 품목을 수입한 국가들 간의 평균 거리를 계산하는 SNDIST_missing 사용자 함수를 생성하여 학습 데이터와 예측 데이터에 각각 적용하여 결측값을 처리하였다.

다. 파생변수 정의 및 생성

파생변수	데이터 설명	활용 데이터
IMPORT_COST	OPEC 대표 7개 유종의 가중평균 가격	OPEC Basket Price
	각 국가의 1인당 국내총생산	World Development Indicators
TOB_HSCD	각 국가들의 AG2 별 전세계 대상 수출액, 수입액	Export, Import by HS Commodity Codes
OVERSEA_DPNDCY, IMPORT_TENDENCY	국가별 소비재, 원자재, 자본재의 수출액, 수입액	WITS Product Exports, Imports by Country
	HS코드, 신성질 분류 연계	HS - 신성질별 연계표
INCREASE_RATES	해당 국가의 대출 금리	IMF Lending Intreres Rates
VISA_DAYS	해당 국가의 무비자 일수	외교부

[표 2-2] 추가 활용 데이터 목록

1) IMPORT_COST (수입 비용 부담도) 파생변수

□ 변수 설명

IMPORT_COST = KMDIST * (Oil Basket Price/GDP per Capita)

세계 유가(Oil Basket Price)를 해당 국가의 1인당 GDP (GDP per Capita)로 각 국가의 실질 유가 부담 지수(OBP_Capita)를 도출하고, 이를 기반으로 한국과 거리(KMDIST)에 따른 원유 사용량을 통해 수입 비용을 추출하였다.

□ OBP (Oil Basket Price)

OPEC 회원국들이 생산하는 7개 대표 유종(油種)의 가격을 가중평균한 원유가격인 Oil Basket Price(OBP) 데이터 확보하였고, 배럴당 단위 가격을 리터 단위로 변환하여 진행하였다.

□ GDP per Capita (1인당 국내 총생산)

Word Bank에서 제공하는 World Development Indicators 지표에서 각 국가의 1인당 국내 총 생산 데이터를 확보하였고, 이를 OBP로 나누어 실질 유가 부담 지수 (OBP_Capita)를 도출하였다.

2) TOB_HSCD (HSCD별 무역수지) 파생변수

□ 변수 설명

Trade of Balance = (Export Value - Import Value) by AG2

Trade of Balance (TOB)는 무역 수지를 나타내며 수출액과 수입액의 차이로 계산한다. TOB_HSCD 변수의 경우 HS Code 2품목 (AG2)별로 구분된 국가별 무역 수지를 의미한다.

□ AG2 (All-2-digit HS commodities)

UN Comtrade에서 각 국가들의 AG2 별 전세계 대상 수출액, 수입액 데이터를 확보하였다. AG2는 HS Code의 앞 2자리를 의미하며, 품목의 부(Section)을 나타내는 코드이다. AG2로 구분된 각 국가별 수출액에서 수입액을 뺀 TOB를 학습 데이터와 예측 데이터의 HSCD 앞 2자리에 맞게 결합하여 파생변수를 추출하였다.

3) OVERSEA_DPNDCY (해외 의존도) 파생변수

□ 변수 설명

OVERSEA_DPENDCY = (Import Value - Export Value) by 신성질별 분류

해외 의존도는 해당 국가에서 해당 품목을 얼마나 해외에 의존하고 있는지를 의미하는 파생변수이며, 신성질 별 분류 체계로 구분된 각 분류의 수입액에서 수출액을 빼는 것으로 계산한다.

□ 신성질 별 분류

신성질 별 분류는 관세청이 HS 품목 분류를 기준으로 소비재 (Consumer Goods), 원자재 (Raw Materials), 자본재 (Capital Goods)를 대분류로 하여 제공하는 무역 통계를 의미한다. 관세청에서 확보한 HS 신성질별 연계표에 따르면 각 HS 품목은 대분류, 중분류, 소분류로 구분되어 있다.

	HSCD	Category_L	Category_M	Category_S
0	010121	소비재	직접소비재	축산물
2	010129	소비재	직접소비재	축산물
4	010130	소비재	직접소비재	축산물
6	010190	소비재	직접소비재	축산물
7	010221	소비재	직접소비재	축산물

[그림 2-7] 신성질 별 분류 예시

□ DEPENDENCY

World Integrated Trade Solution(WITS)의 신성질 분류 별 수출입 데이터를 확보하였고, 국가별 Consumer Goods, Raw Materials, Capital Goods의 수입액에서 수출액을 뺀 값으로 각 품목 분류의 의존도(DEPENDENCY)를 추출하였다.

□ Export Share & Import Share

WITS 데이터에서 각 국가의 신성질 분류 별 총 상품 무역 (수출, 수입)의 점유율 데이터를 확보하여 부가적으로 파생변수에 추가하였다.

4) IMPORT_TENDENCY (수입 증감도) 파생변수

□ 변수 설명

IMPORT_TENDENCY = Import Value (N) year - Import Value (N-1) year

수입 증감도는 같은 신성질 분류 별 품목에 대해 수입액이 얼마나 증감했는지를 알아보는 파생변수이며, 해당연도 수입액에서 전년도 수입액을 차감하여 계산한다.

□ TENDENCY

WITS의 신성질 분류 별 수출입 데이터에서 학습 데이터 기준연도인 2017년 수입 데이터와 그 전년도 2016년 수입 데이터를 추출하여 증감도(TENDENCY)를 계산하였고, 같은 방식으로 예측 데이터 기준연도인 2018년과 2017년도 데이터를 기반으로 증감도를 계산하였다.

5) INCREASE_RATES(대출 금리) 파생변수

□ 변수 설명

INCREASE_RATES= Lending Interest Rates로 해당 국가의 대출금리 의미

특정 국가의 대출금리는 현지 자금 조달과 수출 대금 회수에 영향을 주는 것을 확인하였다. 이에 따라 해당 국가의 대출금리가 낮을수록 국내 수출 기업의 사업이 용이하다고 판단하였다.

6) VISA_DAYS(비자 일수) 파생변수

□ 변수 설명

VISA_DAYS = 각 국가별 0 ~ 180으로 이루어진 비자 일수

비자의 경우 해외 시장 진출 시 인력 교류에 제약 조건으로 작용한다. 0은 무비자를 지원하지 않는 국가이며, 90은 90일 동안 무비자 체류를 지원한다는 의미이다.

7) 특성 공학(Feature Engineering)을 통한 설명 변수 추출

구분	변수설명	변수 개수
국내총생산(GDP)	1인당 실질 GDP, 경제성장률 등 GDP 변수를 이용하여 국가의 재정적인 능력을 확인하고자 함.	4
품목·국가별 수입액	해당 품목 및 국가에 관한 수입액을 통해 전세계 대비 특정 품목과 국가에서 한국으로부터 수입액을 확인하고자 함.	8
관세	품목에 대한 수입을 결정짓는 관세를 이용하여, 특정 품목 및 국가 간 용이성을 확인하고자 함.	4
환율	해당 국가의 환율을 고려한 변수로, 해당 국가의 무역 용이성 및 GDP 및 수입액과 연관을 확인하고자 함.	9

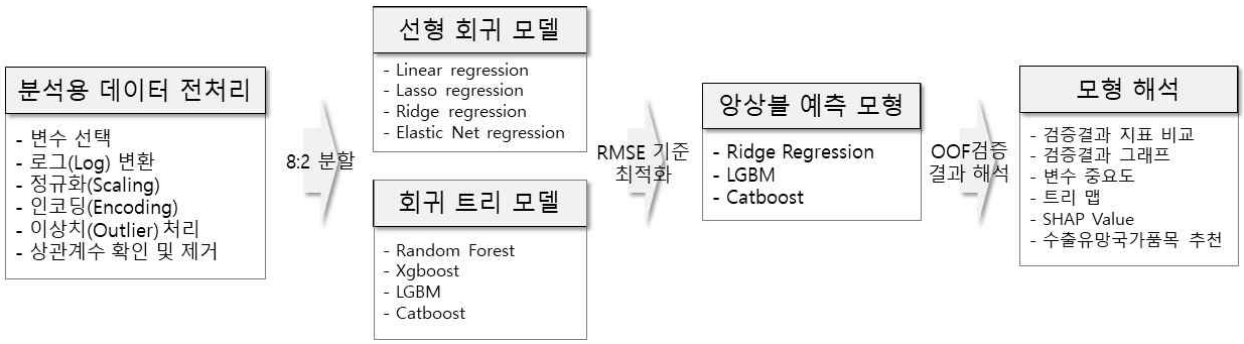
[표 2-3] 4가지로 구분한 설명 변수 추출

본 예측 앙상블 모델링을 위해 데이터 모델의 복잡도와 최적화를 통해 성능을 높이는 것을 지양하였다. 대신 도메인에 대한 이해를 높이고 예측모형에 있어 이를 잘 설명할 수 있는 특성 공학(Feature Engineering)에 초점을 맞추어 다음과 같이 파생변수를 정의하고 추출하였다.

3. 앙상블 예측 모델링

가. 분석용 데이터 전처리

1) 앙상블 예측 모델 개발 흐름도



[그림 3-1] 앙상블 예측 모델 개발 프로세스

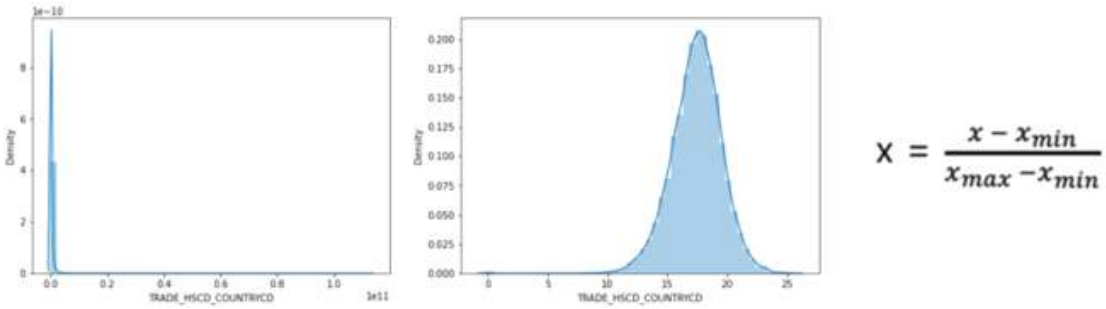
앙상블 예측 모델을 개발하기 위해 다음과 같은 프로세스를 진행하였다. 먼저 분석용 데이터 전처리를 통해 모델에 투입될 데이터를 정제하였으며, 평가 척도는 모델이 예측한 값과 실제 관찰되는 값의 차이를 다룰 때 사용하는 평균 제곱근 오차(Root Mean Square Error; RMSE)를 사용하였다.

비대칭 분포를 이루는 변수에 대해 로그(Log) 변환, 수치형 변수에 대해 정규화(Normalization), 범주형 변수에 대한 인코딩(Encoding), 학습에 부정적인 영향을 끼치는 이상치(Outlier) 처리, 강한 상관성을 보이는 변수를 제거하며 분석용 데이터 전처리 과정을 진행하였다.

예측 모델을 개발하기 위해 훈련데이터와 테스트 데이터는 무작위로 8:2 비율로 나누었으며, 훈련데이터는 K-중첩 교차분석(K-Fold cross validation)을 통해 모델 평가 및 검증하였다.

총 8개의 회귀 알고리즘의 RMSE 값을 비교하여 최종적으로 선정된 모형은 Ridge regression, LGBM, Catboost의 3가지 모형을 선정하였다. 하이퍼 파라미터(Hyper Parameter) 최적화를 통해 최종적으로 앙상블(Ensemble) 모델을 개발하였다. 이를 통해 한국으로부터 수입한 금액을 예측하고, 어떤 변수가 수출 금액을 예측하는 데 있어 중요한 요인인지에 관한 모형을 해석하고 결과를 확인하고자 한다. 아울러 특정 국가의 품목에 따른 군집화(Clustering) 기법을 통해 수출 유망 품목과 국가를 탐색하고자 한다.

2) 로그(Log) 변환 및 최소최대정규화(Min-Max Normalization)



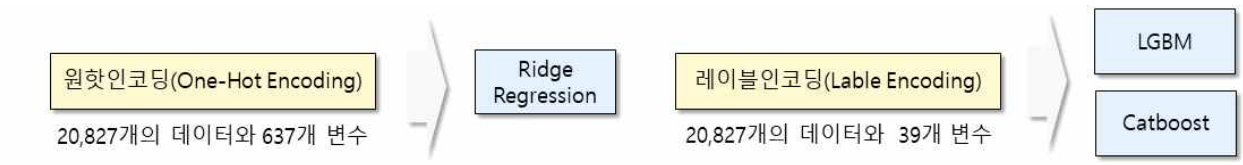
$$X = \frac{x - x_{min}}{x_{max} - x_{min}}$$

[그림 3-2] 로그(Log) 변환과 최소최대정규화(Min-Max Normalization)

좌측 그림과 같이 한쪽으로 치우쳐진 형태를 왜도(Skewness) 존재한다고 한다. 이는 정규분포를 이루지 않아 올바른 학습이 되지 않아 예측 성능에 부정적인 영향을 미칠 수 있다. 따라서 왜도 값이 1이 넘는 변수들에 대하여 로그(Log) 변환을 통해 우측 그림으로 변경하였다.

아울러 올바른 학습을 위해 일정한 범위의 값으로 수렴하는 것이 중요하다. 따라서 수치형 변수들에 대해 최솟값이 0, 최댓값이 1인 사이로 값을 변환하는 최소최대정규화(Min-Max Normalization)를 적용하였다.

3) 원핫인코딩(One-Hot Encoding)과 레이블 인코딩(Label Encoding)

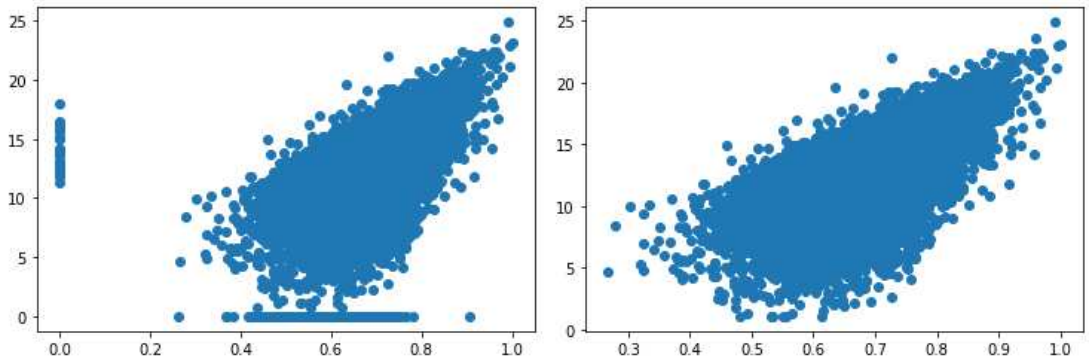


[그림 3-3] 인코딩에 따른 예측모형 선정

Ridge Regression 모형의 경우 품목코드(HSCD)와 국가코드(COUNTRYCD), 품목에 대한 대분류(Category_L), 중분류(Category_M), 소분류(Category_S)의 범주형 변수에 대해 원핫인코딩(One-hot Encoding)을 통해 0과 1의 더미 변수로 변환하였다.

이외 LGBM과 Catboost의 트리 기반 부스팅 모델은 범주형 변수 자체를 카테고리로 인식하기 때문에, 레이블 인코딩(Label encoding)을 적용하였다.

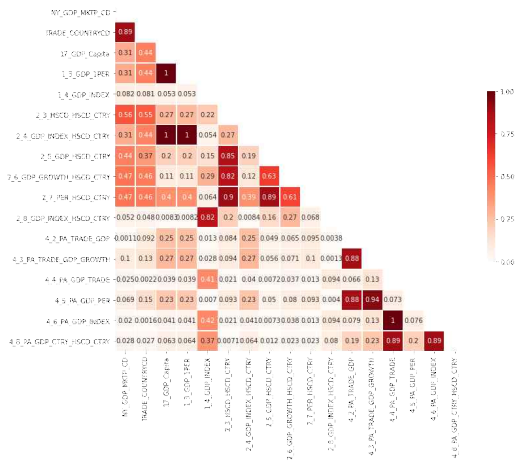
4) 이상치(Outlier) 처리 및 상관계수 확인



[그림 3-4] 설명 변수와 종속 변수의 이상치 제거

회귀모델은 기본적으로 종속 변수와 설명 변수 간 선형관계를 가정한다. 따라서 회귀 문제를 해결하기 위해 종속 변수와 설명 변수 간 선형관계를 확인하였다. 다음 산점도를 통해 이상치(Outlier)를 정의하고 이를 제거하였다.

TRADE_HSCD_COUNTRYCD 와 KR_TRADE_HSCD_COUNTRYCD 의 변수에 대한 분포 확인하면, 정규분포가 아닌 좌측으로 심하게 치우쳐 있다는 것을 확인하였다. 이를 도메인 관점에서 해석하였을 때, 해당연도 해당 국가의 해당 품목 수입금액(TRADE_HSCD_COUNTRYCD)은 있으나 다음연도 해당 국가는 한국으로부터 수입한 금액(KR_TRADE_HSCD_COUNTRYCD)은 0, 즉 한국으로부터 수입하지 않았다는 것으로 이해할 수 있다. 본 수출액예측 모델을 위해 다음연도의 수출액을 정확히 예측하기 위해서 왜도 분포를 이상치라고 정의하였다.

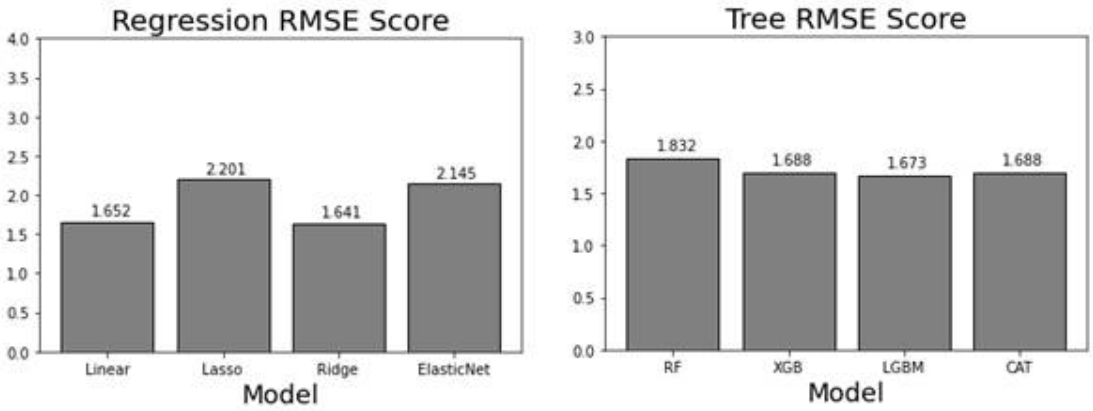


[그림 3-5] 상관성이 높은 변수 제거

설명 변수 간 상관계수를 확인하기 위해 히트맵(Heatmap)을 통하여 강한 상관성을 보이는 0.8 이상의 변수를 확인하였다. 이를 통해 최종적으로 학습 모형에 필요하지 않은 변수를 일부 선택하여 제거하였다.

나. 앙상블 예측 모델 개발

1) 예측모형 제안 및 비교선정

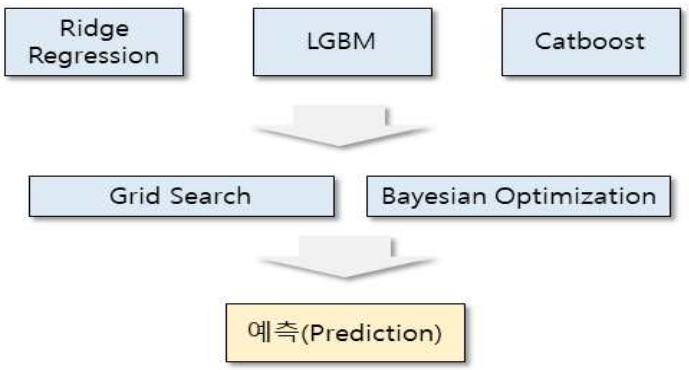


[그림 3-6] 각 예측모형에 대한 RMSE 비교

총 7개의 회귀 모형을 비교하여 최종적으로 선정된 모형은 평가지표인 RMSE 값이 가장 낮은 Ridge regression, LGBM, Catboost 3가지 모형을 선정하였다.

원핫 인코딩(One-Hot Encoding)을 적용한 데이터의 경우 강한 희소성을 보이기 때문에 과적합(Overfitting)의 위험이 많다. 이에 대해 변수의 가중치를 조절하여 과적합을 방지하는 L2 규제의 Ridge Regression을 선정하였다. 레이블 인코딩(Label Encoding)을 적용한 데이터의 경우 의사결정(Decision Tree)의 앙상블 모형으로 속도가 빠르며 적은 메모리 사용량을 가진 고효율 알고리즘이자 성능이 뛰어난 LGBMRegressor 모형을 선정하였다. 아울러 데이터 특성상 500개의 품목코드와 43개 국가, 품목코드에 대한 대·중·소 분류에 대한 카테고리형의 데이터 특성을 고려하여 카테고리형 예측에 적합한 CatboostRegressor 모형을 사용하였다.

2) 예측모형의 하이퍼파라미터(Hyper-Parameter) 최적화

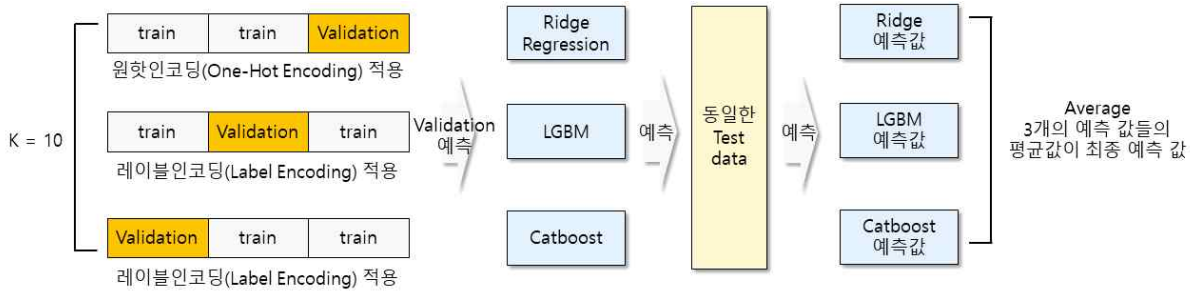


[그림 3-7] 하이퍼파라미터(Hyper-parameter) 조정

각 예측모형은 Python의 머신러닝 모듈인 Scikit-learn을 사용하였고, Grid Search와 Bayesian Optimization을 통해 조정할 수 있는 각각의 하이퍼 파라미터를 탐색하여 모델의 최적화를 진행하였다. 테스트 결과, Ridge Model의 경우 alpha에 대한 파라미터를 조정하였을 때, 10의 값에서 최적의 성능을 확인하였다.

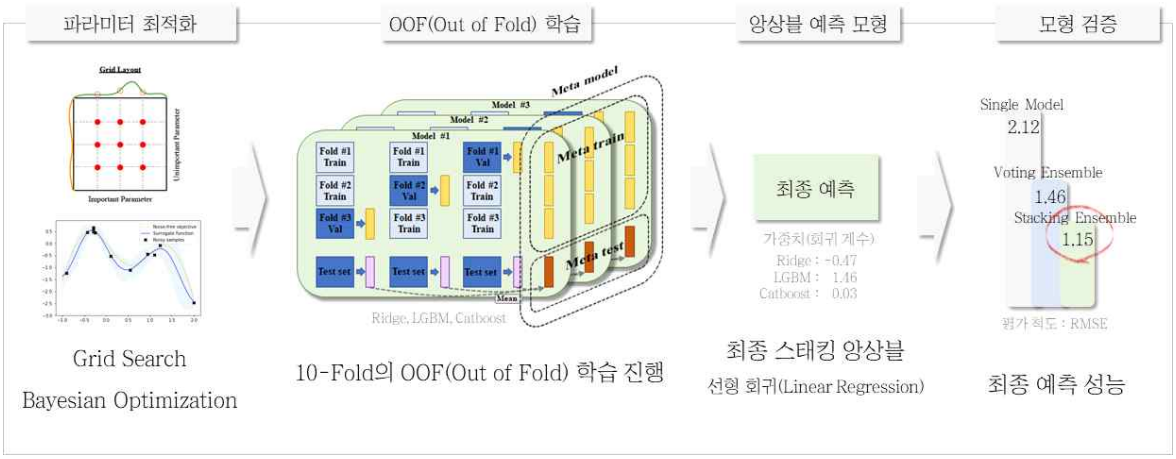
LGBM의 경우 n_estimators = 1000, learning_rate = 0.05, max_depth = 16, num_leaves = 150, min_child_samples = 15, colsample_bytree = 0.65, subsample = 0.65의 값에서 최적의 성능을 확인하였다. Catboost의 경우 파라미터 조정하지 않는 경우가 가장 높은 성능을 확인하였다.

3) 최종 앙상블 예측 모델 개발



[그림 3-8] OOF(Out of Fold) 방식을 이용한 앙상블 모델 개발

앙상블(Ensemble)이란 여러 모델을 이용하여 데이터를 학습하고, 모든 모델의 예측 결과를 평균하여 예측하는 기법이다. 과적합 가능성을 예방하기 위해 자주 사용되는 교차검증 방법인 K-중첩 교차분석(K-fold cross validation)을 이용하였다.



[그림 3-9] 스태킹 앙상블(Stacking Ensemble) 예측

본 앙상블 모델은 10번의 교차검증을 이용하여 예측값을 계산하는 OOF(Out of Fold) 방식을 통해 과적합의 위험을 낮추고자 하였다. 최종적으로 3개 모델의 예측

값들의 평균값을 취하여 테스트 데이터에 대한 예측값을 계산하였다. 최종적으로 선형 회귀(Linear Regression) 모형을 통한 스태킹 앙상블(Stacking Ensemble)을 이용하여 최종 예측하였다. 동일한 테스트 데이터에 대한 성능은 1.15의 RMSE 값으로, 기존 3개의 모델링과 비교하여 우수한 성능을 확인하였다.

이러한 과정을 통해 2017년 세계 수출액 데이터와 2018년 한국수출액에 대해 학습하였고, 2018년 세계 수출액 데이터를 통해 2019년 한국수출액을 예측하였다. 이후 각 예측 모델에 대한 해석을 진행하고, 2019년 한국수출액을 이용하여 수출 유망국가 및 품목을 탐색하고자 한다.

다. 모델 결과 및 시사점

1) 실제값과 앙상블 예측값 비교

Ridge	LGBM	Catboost	Actual_log	Stacking_log	Actual_value	Predict_value
14.36933	14.61473	14.52873	15.11129	14.75029	3653845.00000	2546656.93341
21.74515	21.83725	21.07773	22.12470	22.06328	4061035672.00000	3819106247.02447
9.74074	9.55222	10.02636	9.15377	9.36201	9449.00000	11636.80440
12.08858	12.63590	12.52811	12.35819	12.87323	232859.00000	389738.67147
14.49703	14.59196	14.91440	14.46261	14.65266	1909995.00000	2309778.96375

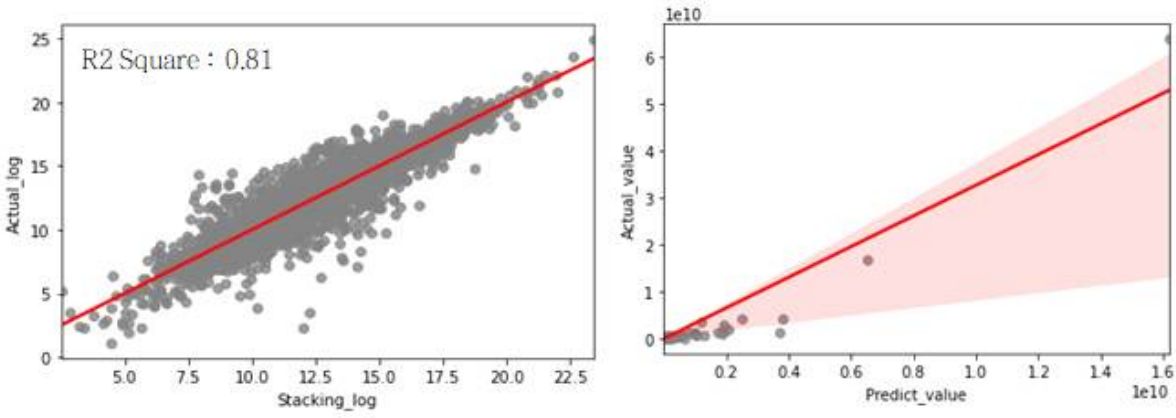
[그림 3-10] 실제 값과 앙상블 예측값의 비교

Ridge, LGBM, Catboost 개별 예측 모델과 이를 스태킹 앙상블 예측 모델에 대한 최종 결괏값은 다음과 같다. 로그(Log)값을 기준으로 확인하였을 때, 앙상블 예측 모형은 실제값과 상당히 유사하도록 예측하는 것을 확인하였다. 이를 실제 값과 비교하였을 때, 수출액의 규모나 추세를 예측할 수 있는 것으로 확인하였다.

대체로 Ridge 모형은 L2 규제를 통해 과적합을 방지하고, 실제로 낮은 수출액에 대해 LGBM, Catboost보다 나은 예측하는 것으로 이해할 수 있다. 아울러 LGBM과 Catboost는 범주형 변수에 대해 카테고리 자체로 인식하기에, 품목과 국가를 통한 수출액에 대한 경향을 보다 자세하게 예측하는 것으로 이해할 수 있다. 이러한 각 모델 특성을 고려한 스태킹 앙상블을 이용하여 최종 예측값을 도출하였다.

결론적으로 실제값과 앙상블 예측값의 RMSE 는 1.15으로 확인하였다. 이는 초기 모델 선형회귀(Linear Regression) 모형으로 예측하였을 때, RMSE 값이 2.12에 비하면 많은 예측 성능향상을 하였다고 할 수 있다.

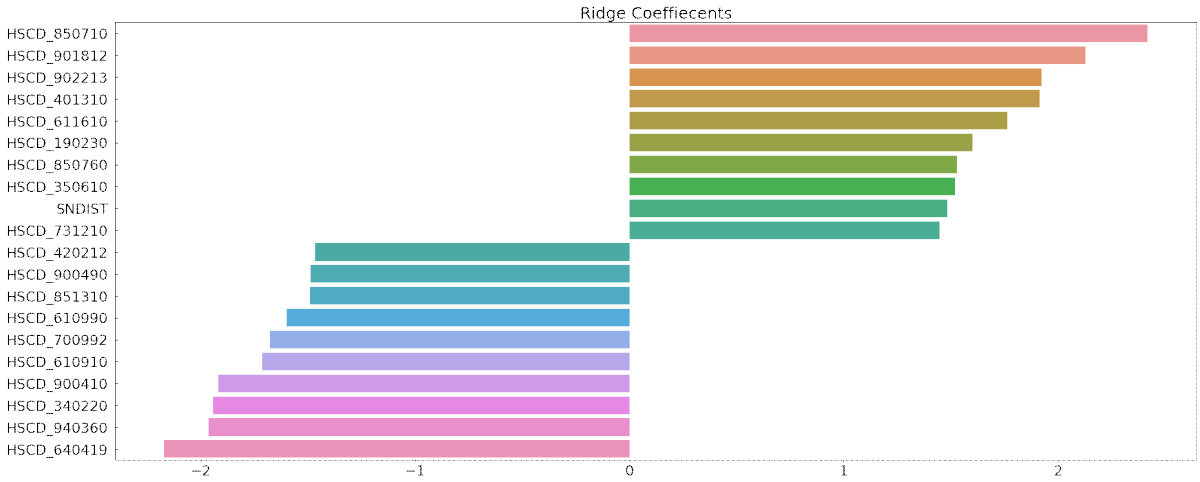
2) 실제값과 앙상블 예측값의 추세 반영 그래프



[그림 3-11] 실제값과 예측값의 선형추세

좌측 그래프는 로그를 기준으로 실제값과 앙상블 예측값의 선형관계를 출력한 것이다. R2 값이 0.81로 선형관계를 통해 실제값과 앙상블 예측값의 추세를 반영하고 있음을 의미한다. 아울러 우측 그래프를 통해 실제값과 예측값의 차이로, 일부를 제외한 데이터에서 규모와 추세를 반영함을 알 수 있다.

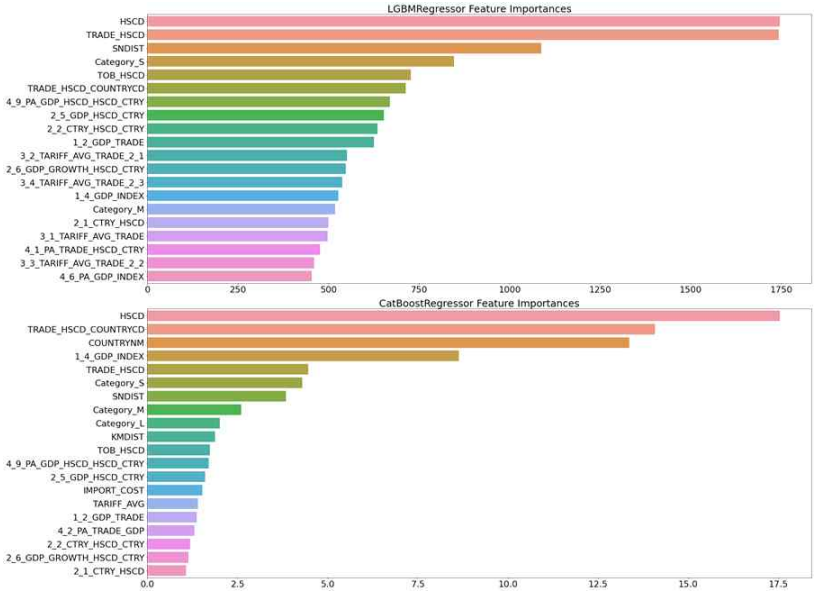
3) 회귀 계수(Coefficient) 그래프 확인



[그림 3-12] Ridge 모형의 회귀 계수(Coefficient)

다음 그래프는 Ridge 모형의 회귀 계수 중 상위 10개와 하위 10개를 출력하여, 어떠한 변수가 중요한지 확인하고자 하였다. 주로 중요한 변수는 품목코드(HSCD)임을 확인하였다. 이는 특정 품목에 따라 한국수출액에 영향을 미치는 것으로 해석할 수 있다.

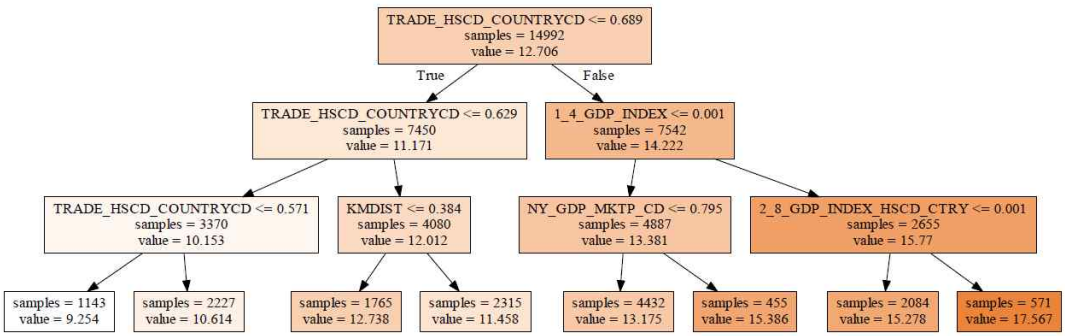
4) 변수 중요도(Feature Importance) 확인



[그림 3-13] 트리 기반 모델의 변수 중요도

해당 모델에서의 변수 중요도(Feature Importance)는 수출액을 예측할 때, 자주 사용되는 변수가 무엇인지 알아보기 위한 과정이다. LGBM과 Catboost의 변수 중요도를 확인하면 주로 품목, 국가, 수입액, 국내총생산(GDP)의 순으로 변수 중요도가 높게 나왔다. 이는 품목명, 국가, 해당 국가의 수입액과 국내총생산(GDP)에 따라 한국수출액을 예측하는 중요한 요인으로 해석할 수 있다.

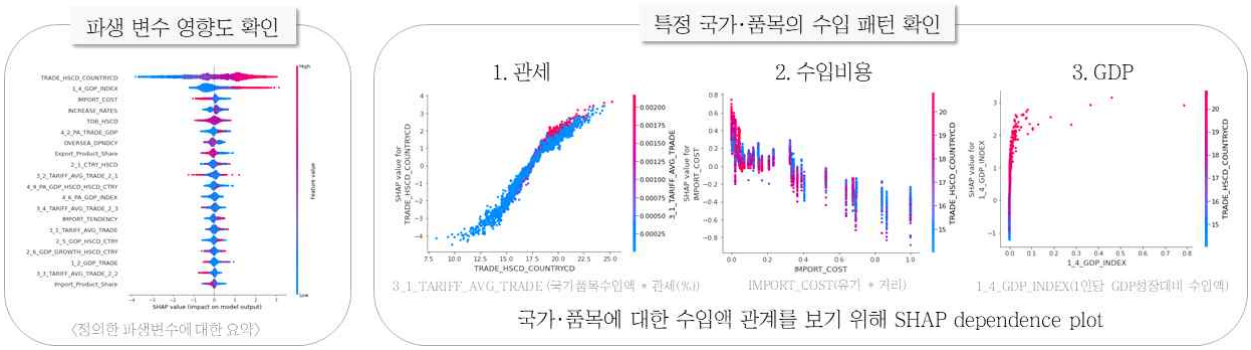
5) 수출예측 모형의 덴드로그램(Dendrogram) 시각화



[그림 3-14] 의사결정 나무(Decision Tree)의 분류 시각화

의사결정 나무(Decision Tree)모형을 이용하여 어떠한 변수를 통해 분류되는지 특성을 파악하고자 해석하였다. 먼저, TRADE_HSCD_COUNTRYCD의 변수를 통해 초기 분류되고, 국내총생산(GDP)을 이용한 변수를 중심으로 분류되는 것을 확인하였다. 이는 해당 품목의 국가별 품목에 대한 수입액과 국내총생산(GDP)에 관한 파생변수가 수출액예측에 중요한 요인으로 해석할 수 있다.

6) 설명가능한 인공지능 SHAP Value 확인



[그림 3-15] XAI(SHAP Value)를 적용한 모델 해석

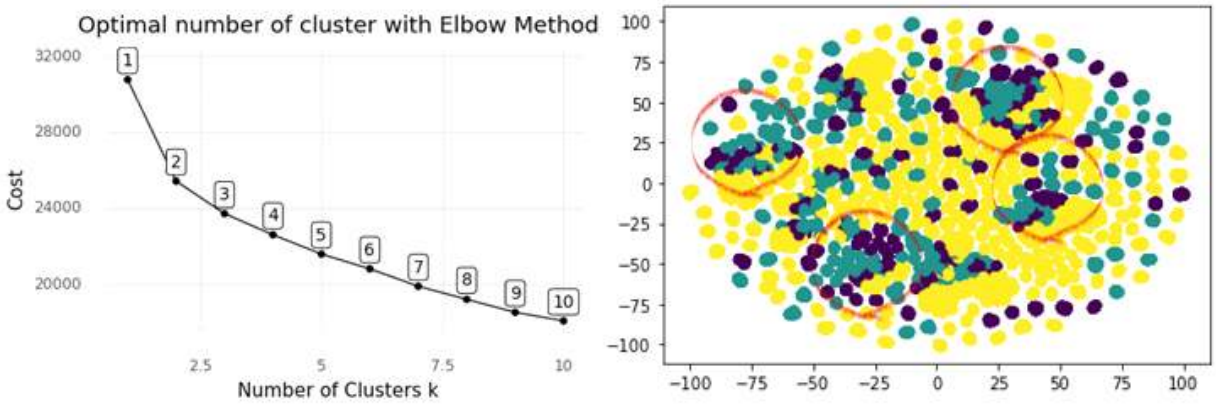
수출 관점에서 도메인 지식을 활용한 파생변수에 대한 특성을 확인하기 위해 설명가능한 인공지능의 SHAP value를 확인하였다. 먼저, 해당 국가의 낮은 관세 품목일수록 수입하는 경향이 높은 것을 확인하였다. 아울러 해당 국가의 낮은 수입 비용일수록 수입하는 경향 높은 것을 확인하고, 국내 총생산(GDP)을 통한 시장 성장 가능성이 클수록 국가의 수입액이 높음을 확인하였다.

이를 통해 수출입에 영향을 주는 요인은 주로 관세, 수입 비용, GDP의 성장성에서 국가의 수입액과 의존성을 확인할 수 있었다. 이러한 영향요인은 국내 수출기업이 수입 비용, 국가정보, 시장 성장성, 규모 등을 통해 의사결정에 영향을 줄 수 있다고 판단하였다. 따라서, 수출 유망국가와 품목을 탐색할 때, “종합적인 해외 시장 요인”을 제공하고자 하였으며, 클러스터링 기반의 기업 선호도 반영 서비스를 제안하고자 한다.

4. 결론

가. 활용 방안

1) 최적 K를 통한 군집화(Clustering) 구성 및 시각화



[그림 4-1] Elbow Method를 통한 최적 K 그룹의 시각화

2018년 데이터를 2019년의 한국으로부터 수입할 금액을 앙상블 모델을 통해 예측하였다. 이를 통해 군집화(Clustering) 기반의 수출 유망국가와 품목을 탐색 및 추천하고자 한다.

군집 알고리즘은 500개의 품목과 43개 국가가 있는 범주형 데이터와 수치형 데이터가 혼합된 특성을 고려하여, 가장 적합한 K-prototypes 알고리즘을 활용하였다. 군집화를 위해 데이터 전처리를 진행한 후 최적의 K를 찾기 위한 과정을 진행하였다. 군집의 최적 개수는 비용이 가장 완만하게 감소하는 구간인 3개의 군집으로 최적 K를 선정하였다. 우측 그래프는 모델용 고차원 학습데이터를 차원 축소를 통해 2차원의 데이터 형태로 함축하였으며, 군집이 생성된 것을 확인할 수 있다.

2) 군집화(Clustering) 기반 수출 유망 국가 및 품목 추천

cluster_id		COUNTRYNM	HSCD	Category_M	Category_S	Detail_Category	KR_TRADE_HSCD_COUNTRYCD	1_1_GDP_GROWTH_RATIO	SP_POP_TOTL
0	19816	Mexico	840820	기계류	일반기계	내연기관 및 터빈	388568936.99419	0.57362	0.60540
	19730	Mexico	840991	기계류	일반기계	내연기관 및 터빈	263080704.66474	0.57362	0.60540
	19859	Mexico	840734	기계류	일반기계	내연기관 및 터빈	232649409.27000	0.57362	0.60540
1	16447	Philippines	847990	기계류	일반기계	냉난방기계	142734560.00886	0.57810	0.57776
	15168	Viet Nam	850440	기계류	전기기계	전기변환장치	79466745.76488	0.73208	0.55968
	11201	Viet Nam	854449	기계류	전기기계	전선 및 케이블	75272978.53110	0.73208	0.55968
2	15143	China	850440	기계류	전기기계	전기변환장치	820444865.03291	0.85919	1.00000
	6225	China	903149	기계류	정밀기계	계측기	352686932.65761	0.85919	1.00000
	16213	China	848180	기계류	기계요소 및 공구류	기계요소	351896946.08704	0.85919	1.00000

[그림 4-2] 5개의 세그먼트 중 일부 데이터 목록

다음 표는 K-prototypes 알고리즘을 통해 도출한 품목 및 국가별 3개의 세그먼트이다. 이는 예측 모델에 사용한 다양한 변수를 기반으로 종합된 군집의 결과이다. 한국으로부터 수입한 금액(KR_TRADE_HSCD_COUNTRYCD)의 변수를 기준으로 내림차순 하여, 세그먼트 별 가장 유망한 품목과 국가를 확인할 수 있다.

아울러, 세그먼트 별 비즈니스 용이성 점수, 국내 총생산(GDP), 수출액, 관세, 환율, 거리, 파생변수 등 다양한 정보를 종합적으로 고려할 수 있다. 이는 수출 기업의 선호도를 반영하여 국가별 품목에 대한 수출 전망을 비교 분석할 수 있다.

세그먼트	국가	유망 세부품목	고려 변수
A	멕시코	전동기, 펌프, 전선 및 케이블, 전기공급 및 제어장치, 발전기 등	인구, 국내총생산(GDP), GDP 성장률, 대출금리, 수출액증감비율, 관세, 비자일수, 비즈니스 용이성
	터키	섬유 및 피혁가공기계, 기계요소, 전기변환장치 등	
B	필리핀	냉난방기계, 건설기계 등	
	베트남	전기변환장치, 전선 및 케이블, 일차전지 및 축전지 등	
	카자흐스탄	금속구조물, 기계요소	
	인도	기계요소, 금속구조물, 건설기계	
	인도네시아	건설기계	
C	중국	전기변환장치, 계측기, 기계요소, 운반하역기계 등	
	미국	전기변환장치, 기타 일반기계, 기계요소 등	
	UAE	금속구조물 등	
	독일	조명장치	

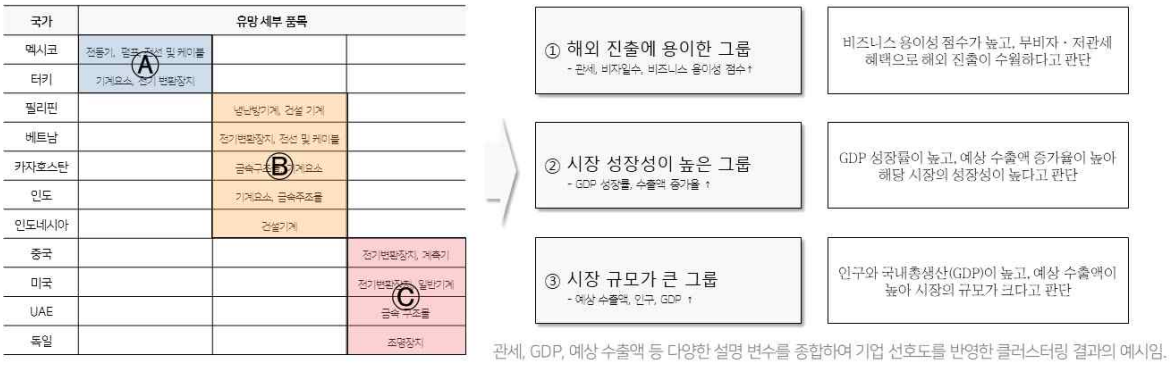
[표 4-1] 3개의 세그먼트 중 일부 수출 유망국가 및 품목 추천

다음의 표는 고도화된 빅데이터로 품목별 유망시장 제시의 예시이다. 다음과 같이 세그먼트 별 인구, 국내 총생산(GDP), 대출금리 등 수출 기업의 선호도를 반영하여 해외 시장 진출 요인을 고려할 수 있다.

이는 코트라에서 제공하는 무역투자빅데이터 플랫폼을 기반에서 조금 더 발전된 방향으로 제안하였다. 앙상블 예측 모델을 통해 다음 연도의 한국 수출 금액을 예측하고, 클러스터링 기반 수출 유망국가와 품목을 고려하여 사용자는 선제적 행동을 할 수 있다. 아울러 클러스터링으로 구별하여 같은 군집 내에 비교 분석을 통해 조금 더 세밀한 세그먼트를 통해 추천할 수 있다는 점에서 의의가 있다.

나. 기대효과

1) 중소·중견 기업의 차별화된 수출전략



[그림 4-3] 세그먼트 별 수출 유망 품목 및 국가 탐색과 추천

다음과 같은 세그먼트 별 수출 유망 품목 및 국가 추천 시뮬레이션을 통해 중소·중견 기업은 무분별한 품목과 국가에 대해 수출을 하는 비용을 감소시킬 수 있다. 아울러 국내 수출액에 영향을 주는 국내총생산(GDP), 관세, 환율, 국제 정세 등 다양한 요인을 종합적으로 비교 분석을 통해 해외 무역 위험에 상대적으로 취약한 중소·중견 기업의 어려움을 덜 수 있을 것이다. 이러한 빅데이터 기반의 서비스를 통해 기존의 코트라 빅데이터 플랫폼을 강화하고, 중소·중견 기업이 수출을 지원하고자 할 때 각 기업의 상황에 맞는 차별적 선택을 통한 기업 성장을 기대할 수 있다.

2) 수출입 관련 정보 획득 비용 절감

중소·중견 기업이 참여하는 수출무역 전시회의 부스 비용은 2칸 기준 약 800만원으로 알려져 있다. 실질적으로 행사를 위해 각종 부가 비용이 발생하기 때문에, 중소·중견 기업의 부담은 그 이상으로 생각할 수 있다. 아울러 코로나로 인한 사회적 거리두기로 인한 부정적인 영향을 간과할 수 없다는 것을 실무 관계자와의 인터뷰를 통해 확인하였다.

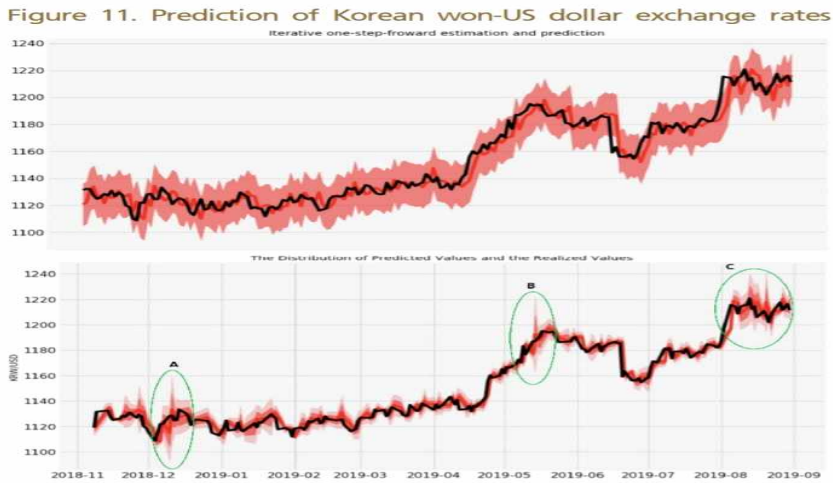
본 앙상블 예측 모델을 통해 한국수출액을 예측하고, 조금 더 세분화하여 수출 유망국가와 품목을 추천하는 서비스를 상용화한다면, 해외 시장 정보 획득을 위한 중소·중견 기업의 비용을 절감할 수 있을 것이다. 아울러 타겟 국가와 품목에 대해 선제적으로 대응할 수 있도록 하여 분석 목적인 중소·중견 무역기업이 수출할 시장을 탐색에 있어 도움을 제공할 수 있을 것으로 기대한다.

다. 한계점 및 제언

1) 국내·국제 정서 요인을 고려한 예측 모델의 필요성

본 앙상블 예측 모델을 통해 해외 요인을 이용한 품목 국가별 수출액 규모와 추세를 예측하였다. 하지만 한국으로부터 수입하지 않은 국가에 대해 예측하는 한계를 확인하였다. 이는 세계 수출액, 환율, 관세, 국가별 국내총생산(GDP)과 같은 해외 요인만을 통해 한국 수출에 큰 영향을 미치는 해외 경기 요인을 제대로 반영하지 못한다는 것으로 해석할 수 있다. 이를 해결하기 위해 수출에 영향을 미치는 선거, 파업, 외환위기, 주가, 기업 신뢰도 등 다양한 국내·국제정서·기업환경 등 다양한 요인을 고려해야 할 것이다.

2) 딥러닝(Deep Learning)을 활용한 비정형·정형 데이터 분석 제언



[그림 4-4] 딥러닝으로 환율·수출 예측 분석(한국은행)

한국 수출액예측 모델의 경우 해당 연도 데이터뿐만 아니라 과거 연도 데이터로 확장하여 시계열 딥러닝 모델인 순환신경망(Recurrent Neural Network)을 제안한다. 이는 시계열 속성을 반영하고, 거시경제와 금융시장 예측을 통해 수출, 환율, 국내총생산(GDP) 등 종합적인 전망을 예측하여 오차범위를 줄일 수 있을 것이다.

기존 정형 데이터를 이용하여 기계학습의 수출예측은 150~200억 달러의 오차 수준이었다, 하지만 국가 간 무역협상, 코로나 등 국제 정세와 같이 불확실성 사건을 반영하는 정형·비정형 데이터를 이용한다면, 딥러닝의 수출예측은 50~100억 달러 수준으로 오차범위를 줄일 수 있을 것으로 기대한다.

5. 부록

가. 참고문헌

◈ 신현수·배상근·최호상, 단기 수출입 예측모형 개발. 한국산업경제연구원, 1999

◈ “한은, 딥러닝으로 환율·수출 예측분석…‘오차범위 줄어‘” 한국은행, 2020

◈ 빅데이터로 본 내년 품목별 유망시장은…자동차는 미국, 반도체는 중국. 이투데이, 2020

◈ 무역보험공사 '수출 빅데이트 분석 리포트' 발간, 파이낸셜투데이, 2020

◈ 내년 수출 6~7%↑ 5400~5500억 달러…코트라 전망, 이투데이, 2020

◈ 윤상혁·최윤진·이소현·김희웅, 머신러닝 기반의 뷰티커머스 고객 세그먼트 분류 및 활용방안. Information System Review, 2020

◈ 무역투자 빅데이터 서비스. 코트라, 2021

◈ Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 1998

나. 활용 데이터 목록

활용 데이터	데이터 형태	목적	출처
Exchange Rates Domestic Currency per U.S. Dollar, Period Average	정형 데이터	결측값 처리	IMF
Yearly Time Series, Imported Value by HS Code	정형 데이터	결측값 처리	ITC TradeMap
HS 신성질별 연계표	정형 데이터	파생변수 생성	관세청 수출입무역통계
OPEC Basket Price	반정형 데이터	파생변수 생성	OPEC
World Development Indicators	정형 데이터	파생변수 생성	World Bank
Export, Import by HS Commodity Codes	정형 데이터	파생변수 생성	UN Comtrade
WITS Product Exports, Imports by Country	정형 데이터	파생변수 생성	World Integrated Trade Solution
IMF Lending Intrests Rates	정형 데이터	파생변수 생성	IMF
외교부 해외안전여행	정형 데이터	파생변수 생성	외교부

[표 5-1] 활용 데이터 목록