산업통상자원부 공공데이터 활용 비즈니스아이디어 공모전 분석 결과 제출 양식

1 명칭

뉴스 기사 기반의 텍스트 모델과 자기회귀모형을 활용한 유망 국가-품목 그룹 추천 시스템

2 제안배경

1) 한국의 무역과 경제, 높은 대외 의존도로 인한 문제 상황

한국은 개발 초기부터 수출 주도형 경제 성장을 추진해 왔고, 그 결과로 수출과 수입이 크게 확대되었다. 수출 상품을 생산하기 위해서는 원자재와 자본재가 필요하므로 수입도 함께 증가하였다. 또한, 국외 요소소득도 수출입액의 약 5%를 차지하며 수출입액과 비슷한 속도로 확대되었다. 아래 그래프는 한국의 수출액과 수입액, 국외 수취 요소소득과 국외지급요소소득을 합한 총액의 GDP 대비 비율을 나타낸다.

출처:
한국은행, 「국민계정」

한국의 GDP 대비 수출입 비율은 1990년 53.0%에서 2022년 102.0%로 상당히 증가했다. 한국의 GNI 대비 수출입 비율은 2020년 기준 72.3%로, 이는 미국의 31%와 일본의 37%에 비해 상당히 높다. 현재 GDP 대비 수출액 비율은 약 100%로 경제의 큰 부분을 차지하고 있다. 즉, 한국의 경제 성장에 있어서 수출입은 매우 중요한 요소로 볼 수 있다.

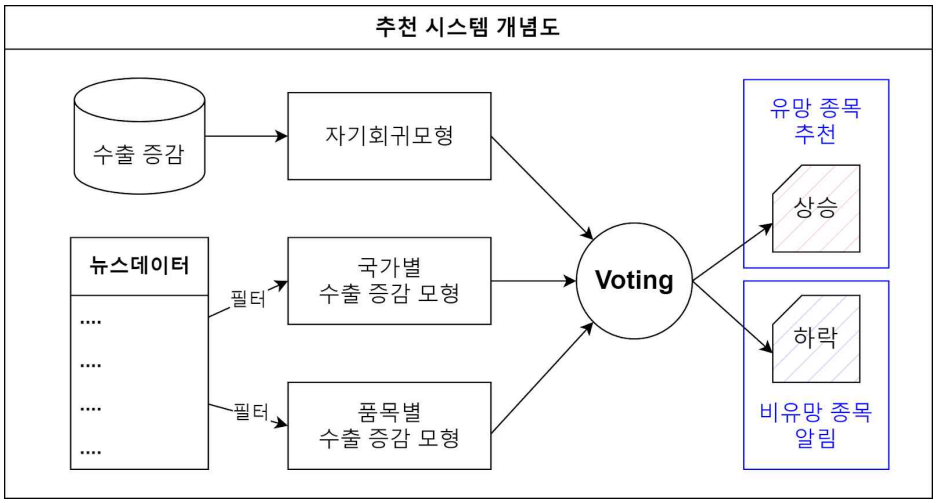
최근 한국의 세계 수출시장 점유율은 2.7%로 2년 연속으로 하락하며 무역적자 비중이 증가하고 있다. 미·중 무역전쟁으로 세계적으로 자국 중심주의와 보호 무역이 확산하고, 러시아 우크라이나 전쟁 기조가 두드러지면서 한국의 수출액에 악영향을 미쳤기 때문이다. 또한, 한국은 메모리 반도체 호황에 의존하며 수출 주력 업종 변화를 소홀히 하여 경쟁력과 역동성을 잃어가는 문제도 있다. 이렇듯 국제적인 이슈들과 국가들의 상황에 따라서 수출입에 영향을 받는 것을 확인할 수 있다.

한국무역협회에 따르면, 한국의 수출시장 점유율이 크게 하락한 이유는 미·중 무역 전쟁과 러우 전쟁으로 인해 세계적으로 자국 중심주의와 보호무역이 확산하여 수출의존도가 높은 한국이 악영향을 받았기 때문이다. 수출 점유율이 0.1% 포인트 하락하면 약 14만 개의 일자리가 감소하는 효과가 나타나며 경제 전반에도 부정적인 영향을 미치게 된다.

이렇듯 국제 금리, 환율, 국제시장 경제 성장률, G7 국가의 유가 상한제, 러시아의 무역 규제 등과 같은 대외적 변화는 한국의 수출입에 영향을 미치며, 이를 예측하는 것은 국가 경제 성장에 있어서 중요한 과제이다. 그러므로 다양한 변수를 고려하여 국가별로 수출액 증가가 예상되는 품목을 선정한 후 이를 활용하여 국가별 유망 품목을 추천해 주는 전략이 필요하다.

2) 구현하고자 하는 시스템

이러한 배경 하에서 '자기회귀모형과 뉴스 데이터를 활용한 모형이 국가-품목별 수출액을 월 스케일에서 예측하는 데 사용될 수 있을 것이다'라는 가정을 두고, 아래와 같은 품목 추천 시스템을 구현하는 분석을 진행하고자 한다.



해당 추천 시스템에서는 지난 시기의 수출 증감 여부를 기반으로 하는 자기회귀모형과, 뉴스 데이터를 필터링한 뒤 이에 머신러닝 모형을 적합시켜 국가별/품목별로 수출 금액 증감 여부를 예측하는 모형을 바탕으로 소프트 보팅(Soft Voting)을 실시하여 해당 국가-품목에 대해 향후 수출액이 상승할지 하락할 지를 도출한다. 이렇게 도출한 상승/하락 예측치는 향후 수출에 있어 규모가 커질 종목을 추천하는 데 사용될 수 있다.

경제 변수들을 활용하여 모형을 적합시키는 기존의 사례들에서는 HS코드 4자리 기준 품목과 같이 세부적인 품목들에 연결시킬 경제 변수를 선정하기 어렵기에 이러한 단계에 대한 분석을 실행하기 어렵다.

본 분석은 이러한 부분을 보완하여 HS코드 4자리의 비교적 세분화 된 품목에도 적용할 수 있는 독립변수를 확보하기 위해, 뉴스 데이터를 수집하고 이를 키워드를 기준으로 필터링하여 사용하였다. 필터링한 뉴스 데이터는 Konlpy의 Okt 형태소 분석기를 활용하여 명사 단위로 단어가 분리되며, 이를 월별로 취합하여 열은 고유한 명사로, 행에는 월별로 각 명사의 등장 횟수를 나열하여 BOW(Bag of Words) 형태로 저장한다. 해당 데이터를 사용하여 국가별, 품목별 수출금액 상승 여부를 예측해보고자 한다.

3 분석 내용 및 분석 결과

목차

---

1) 분석 계획과 가설

1.1) 분석 계획

본 분석은 크게 데이터 파악, 예측 모형 적합, 추천 시스템 구현의 3단계로 진행된다.

데이터 파악 단계에서는 전 세계와 대한민국의 무역 동향을 파악하고, 주된 수출 품목을 확인한다. 이에 더해 경제 변수와 수출 변동률간의 관계를 확인하고, 수집한 뉴스 데이터의 단어 분포에 대해서도 간단한 파악하고자 한다.

예측 모형 적합 단계에서는 수집한 데이터들을 기반으로 3가지 모형을 훈련시키고, 평가한다. 3가지 모형은 이전 12개월동안의 수출액 변동률을 가지고 현 시점의 수출액 변동을 예측하는 자기회귀모형과, 뉴스 데이터를 기반으로 국가별로 대한민국의 수출액 변동 방향을 예측하는 모형, 마찬가지로 뉴스 데이터를 기반으로 하지만 품목별로 수출액 변동 방향을 예측하는 모형으로 구성된다.

마지막으로 추천 시스템을 구현하는데, 적합한 예측 모형들의 결과를 가지고, 2022년부터 2023년 5월까지의 수출액 변동을 예측하는 과제에, 앞서 제사한 개념도의 추천 시스템을 적용해보려 한다. 3개의 예측 모형에 대해 가중치를 두고, 해당 가중치를 기반으로 Soft Voting을 실행한 결과를 기반으로 추천 시스템을 평가할 것이다.

다만 추천 시스템의 평가에 사용되는 기간이 2022년 이후로, 대내외적으로 경제가 불안정한 시기이므로 평가 지표를 해석하는 데 있어 이 점을 고려하려 한다.

1.2) 가설

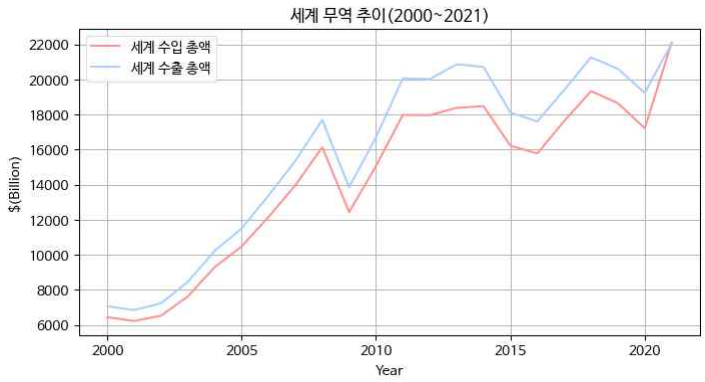
본 분석은 '자기회귀모형과 뉴스 데이터를 활용한 모형이 국가-품목별 수출액을 월 스케일에서 예측하는 데 사용될 수 있을 것이다'라는 가정 하에 실행될 것이다.

과거의 데이터로 미래의 데이터를 예측하는 모형은 일반적인 경제 상황에서 트렌드가 유지될 때에는 높은 예측력을 보이나, 충격이 발생하는 상황에서 이에 적절히 대응하기 어렵다는 단점이 있다. 이러한 단점을 뉴스 데이터 기반의 모형들이 보완하여 정확도를 높일 것으로 기대된다.

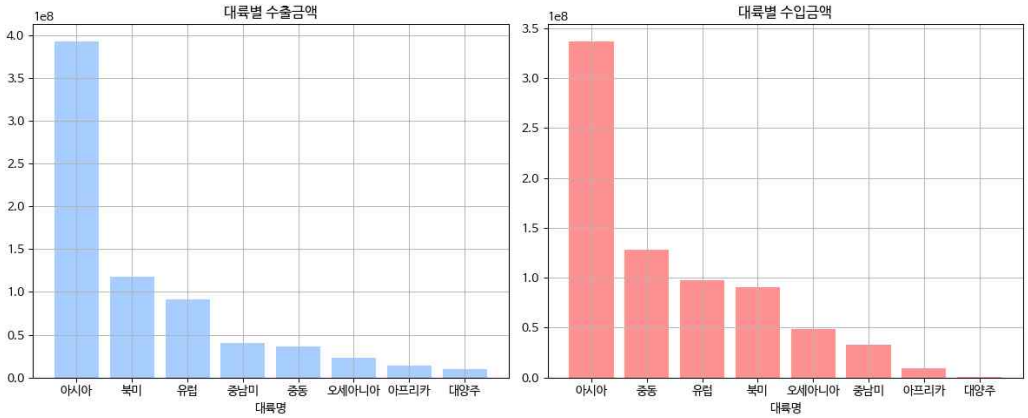
2) 데이터 파악

2.1) 세계 무역 동향

World Bank의 WITS에서 2000년~2021년까지의 세계 수출액 및 수입액 데이터를 가져와서 전세계 수출입 등 무역 동향을 확인해보았다. 그 후 관세청의 수출입무역통계에서 2000년~2023년까지의 대륙별 수출입 데이터를 가져와서 대륙별 수출입 수출입 그래프를 그려서 대륙별 무역 동향을 확인해 보았다.



2000년부터 2021년까지의 세계 수출액 및 수입액의 추이를 나타낸 그래프이다. 2000년대부터 꾸준히 증가하다 2008년 금융위기 여파로 수출입이 모두 크게 감소하였다가 다시 회복하는 추세를 보였다. 2016년에는 석유화학 이슈로 인해 수출액이 감소하는 추세를 보였다 회복했다. 2019년부터는 코로나 팬데믹으로 인해 수출입이 20년만에 가장 큰 폭으로 감소하였다. 2021년에는 전년도에 비해 수출입이 상승하는 모습을 보였다.
금융 이슈가 생기기 전에 수출입은 최고액을 보였다가 이슈가 발생하면서 급감하게 되고 이슈가 완화되면 다시 수출, 수입액을 회복하는 경향이 있다. 이를 무역시장의 사이클로 확인할 수 있다.

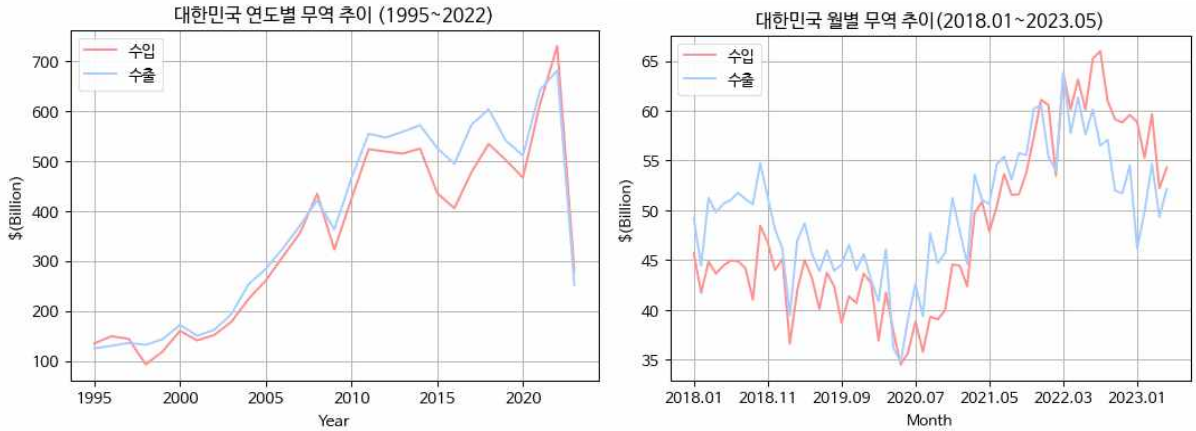


2000년부터 2023년까지의 대륙별 수출금액과 수입금액 총액이다. 아시아의 수출수입액이 큰 비중을 차지하는 것을 확인할 수 있다. 수출수입액은 아시아, 북미, 유럽, 중남미, 중동, 오세아니아, 아프리카, 대양주 순으로 높다.

2.2) 대한민국 무역 동향

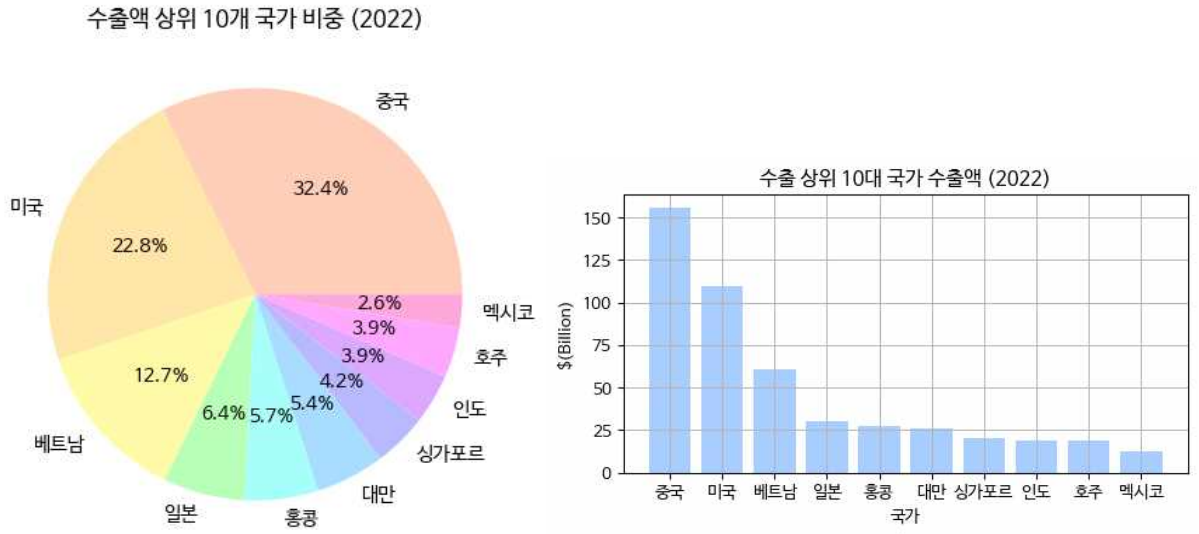
2.2.1) 수출입 동향

관세청에서 공개하는 api를 이용해서 수집한 품목별 수출입 데이터로 한국의 수출입 동향을 연도별로 확인해 보았다. 그 후 수출액 상위 10개 국가의 우리나라 수출금액 비중과 상위 10개 국가의 수출입 동향을 확인해보았다.

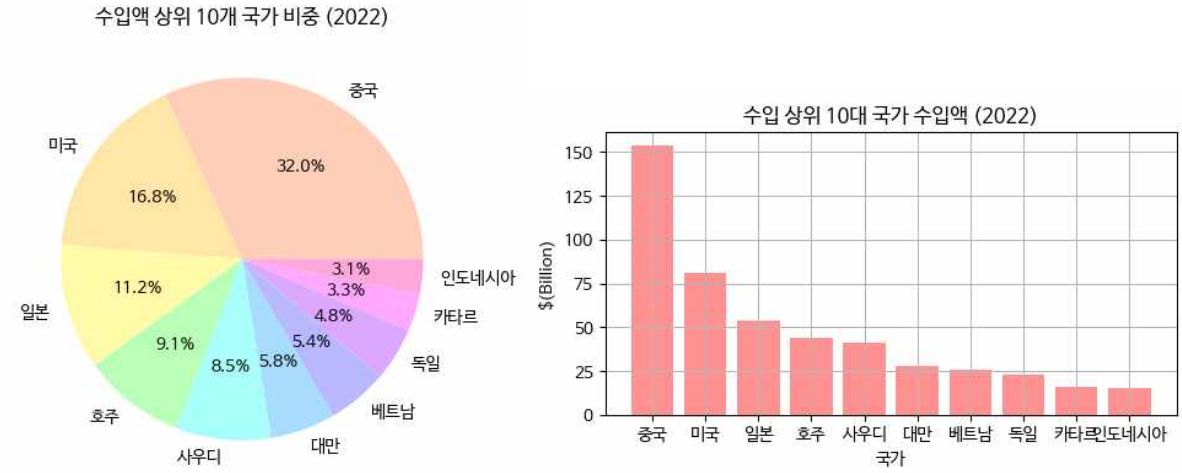


한국의 수출입은 1995년부터 2020년까지 빠르게 성장하였다. 코로나 펜데믹이 발생한 2019년에 수출입이 크게 감소하였고 2020년 하반기부터 점차 회복하는 모습을 보였다.

2.2.2) 국가별 수출입 동향 (대외 수출입)

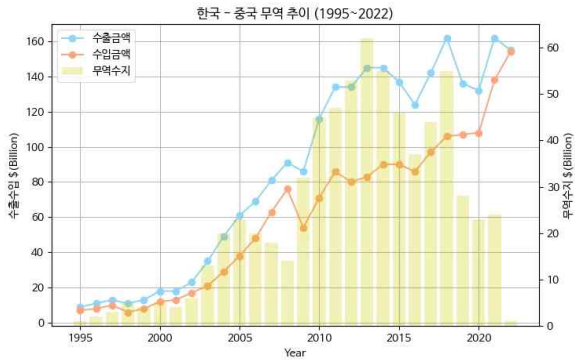


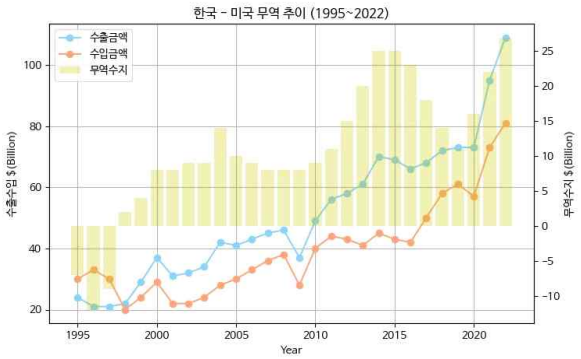
2022년 기준 우리나라 수출액 상위 10개 국가에 대한 수출액 비중과 수출금액을 나타냈다. 전체 수출액 중 중국이 32.4%, 미국이 22.8%로 전체 수출액의 절반 이상을 차지한다.



2022년 기준 우리나라 수입액 상위 10개 국가에 대한 수입액 비중과 수입금액을 나타냈다. 전체 수입액 중 중국이 36.4%, 미국이 25.7%로 전체 수입액의 절반 이상을 차지한다.

(1, 2) 중국, 미국

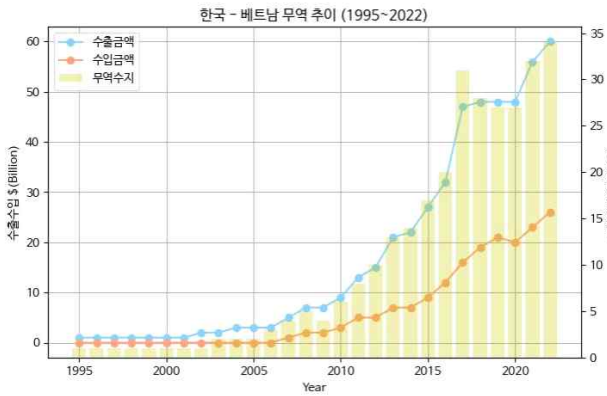


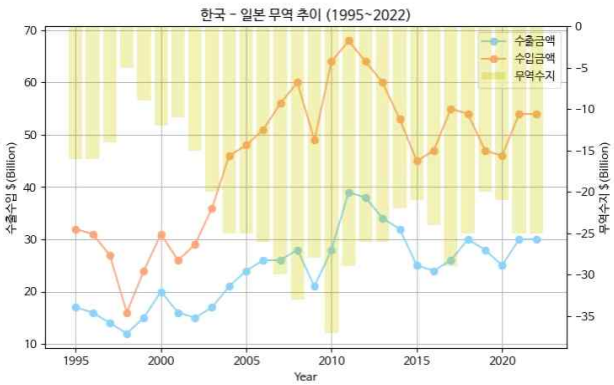


좌측은 1995년부터 2022년까지의 연도별 중국의 수출액, 수입액, 무역수지를 함께 그린 그래프이다. 중국의 수출입은 전반적으로 증가하는 경향을 보이지만 세계 금융위기가 발생했던 2008년, 코로나 팬데믹이 발생했던 2020년에는 수출입이 급격하게 감소하는 모습을 보였다.

우측은 1995년부터 2022년까지의 연도별 미국의 수출액, 수입액, 무역수지를 함께 그린 그래프이다. 미국의 수출입은 전반적으로 증가하는 경향을 보인다. IMF 외환위기가 발생했던 90년대 후반에는 무역수지가 급격이 악화되었고 세계 금융위기가 발생했던 2008년, 코로나 팬데믹이 발생했던 2020년에는 수출입이 감소하는 모습을 보였다.

(3, 4) 베트남, 일본

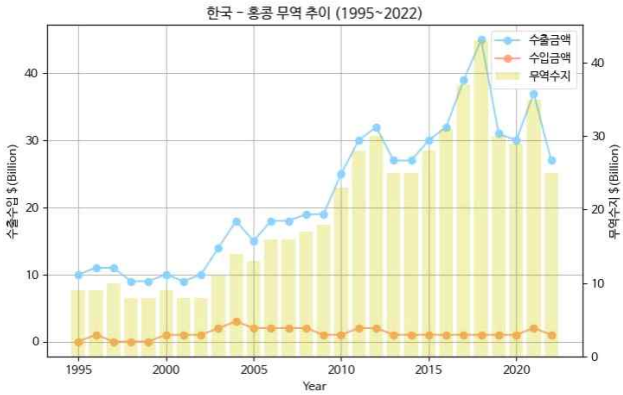


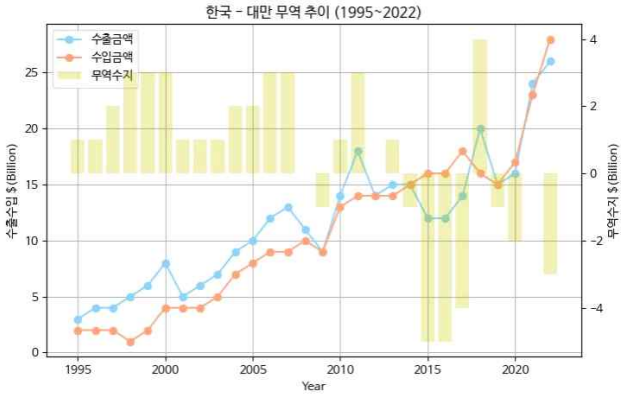


베트남의 수출입은 전반적으로 증가하는 경향을 보이다 2010년대부터 급격하게 증가하였다. 수출입이 2008년 세계금융위기, 코로나 펜데믹 등의 세계 경제 위기에 크게 영향을 받지않은 것으로 보인다.

한국-일본의 무역의 경우 전 기간동안 수입이 수출보다 높은 모습을 보였고 무역적자 폭이 증가하고 있다. IMF 외환위기, 2008년 세계 금융위기, 코로나 펜데믹 기간 모두 수출입이 감소하며 영향을 받는 모습을 보인다.

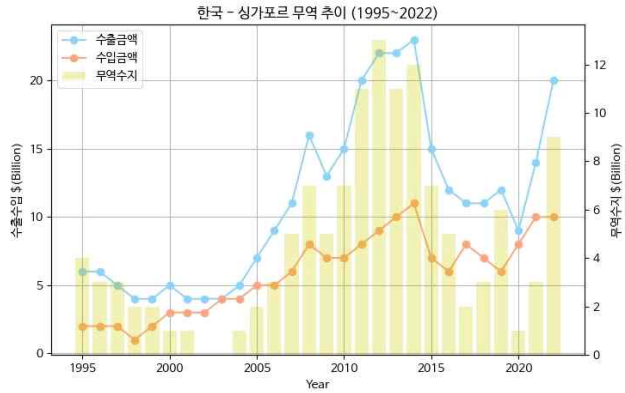
(5, 6) 홍콩, 대만

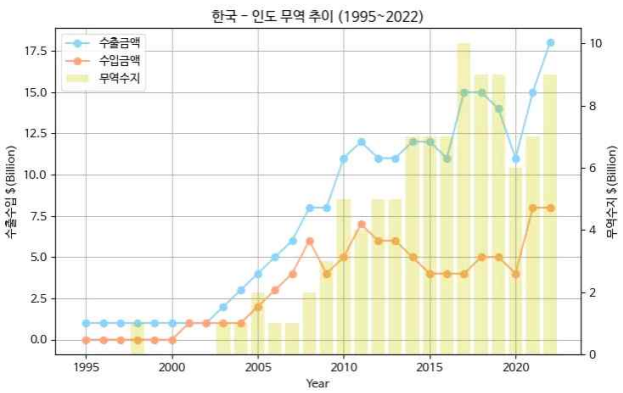




홍콩은 제조업보다 서비스산업에 중점을 둔 경제 구조로 수입액 대비 수출액이 큰 비중을 차지하고 있다. 따라서 수입액은 전기간 동안 낮은 수준으로 유지되는 반면 경제가 개방되며 수출액은 꾸준히 증가하며 무역수지가 증가되고 있다.
대만과의 무역은 수출입 모두 전반적으로 증가하는 모습을 보이지만 2015년 수입액 대비 수출입이 감소하며 큰 폭의 무역 적자를 보였다. 코로나 펜데믹 이후 수출입 모두 개선되는 모습을 보였다.

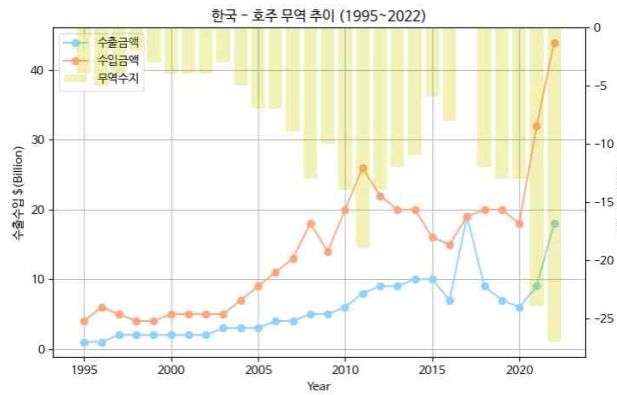
(7, 8) 싱가포르, 인도

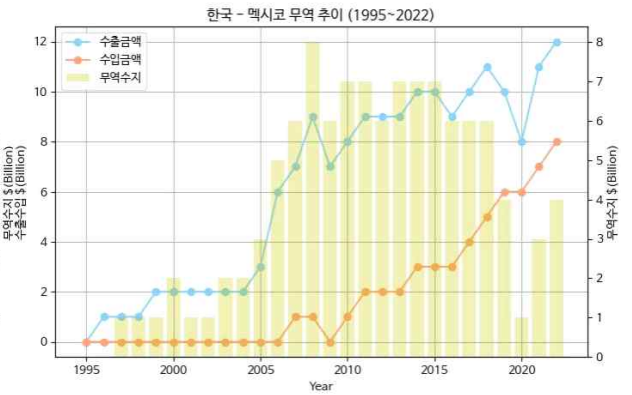




싱가포르의 경우 수출입 모두 전반적으로 증가하는 경향을 보이며 2000년부터 2015년까지 수입액 대비 수출액이 크게 증가하며 무역 흑자를 나타냈다. 코로나 팬데믹 이후 수출액이 크게 증가하며 무역수지가 증가되었다. 인도는 수출입이 꾸준하게 증가하고 있으며 무역수지 또한 꾸준하게 증가되고 있다.

(9, 10) 호주, 멕시코





호주와의 무역은 일본과 마찬가지로 전 기간동안 수입이 수출보다 높은 모습을 보이며 무역적자 폭이 증가하고 있다. 2008년 세계 금융위기, 코로나 펜데믹 기간 모두 수출입이 감소하며 영향을 받는 모습을 보인다.
멕시코의 경우 수출입이 모두 꾸준하게 증가하고 있고 수입액 증가 속도에 비해서 수출액이 급격하게 증가하며 무역수지 또한 증가하고 있다.

2.3) 대한민국 주요 무역 품목

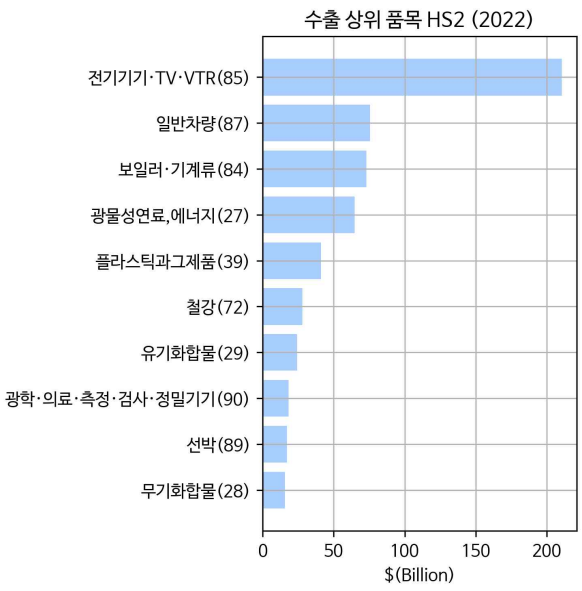
관세청에서 공개하는 api를 이용해서 수집한 품목별 수출입 데이터로 한국 수출품목 TOP10의 수출액과 수입품목 TOP10의 수입액을 확인해보았다. 2022년과 1995년~2022년 평균으로 기간을 나누어서 품목과 수출입액을 확인하였다.

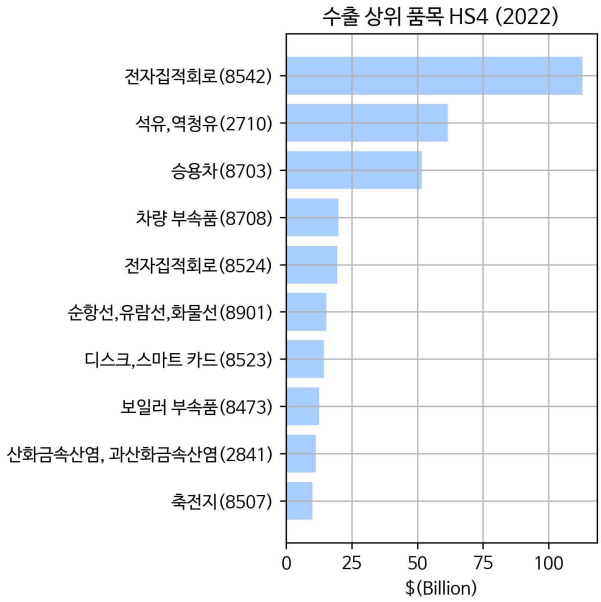
2.3.1) 수출 상위 10개 품목

HS2	수출액*	품목명(HS2)	HS4	수출액*	품목명(HS4)
85	210.4346	전기기기·TV·VTR(85)	8542	112.8472	전자집적회로(8542)
87	75.47448	일반차량(87)	2710	61.56026	석유,역청유(2710)
84	73.03319	보일러·기계류(84)	8703	51.68033	승용차(8703)
27	64.75133	광물성연료,에너지(27)	8708	19.9219	차량 부속품(8708)
39	41.15808	플라스틱과그제품(39)	8524	19.44906	전자집적회로(8524)
72	28.10846	철강(72)	8901	15.25926	순항선,유람선,화물선(8901)
29	24.30409	유기화합물(29)	8523	14.29477	디스크,스마트 카드(8523)
90	18.20857	광학·의료·측정·검사·정밀기기(90)	8473	12.54094	보일러 부속품(8473)
89	17.13771	선박(89)	2841	11.26826	산화금속산염, 과산화금속산염(2841)
28	15.6432	무기화합물(28)	8507	9.981342	축전지(8507)

*단위는 10조 원

2022년을 기준으로 한국의 수출액 상위 10개 품목과 수출액을 나타낸 표이다. HSCODE 2자리, HSCODE 4자리를 기준으로 품목을 나눠서 수출액을 내림차순 정렬하였다.



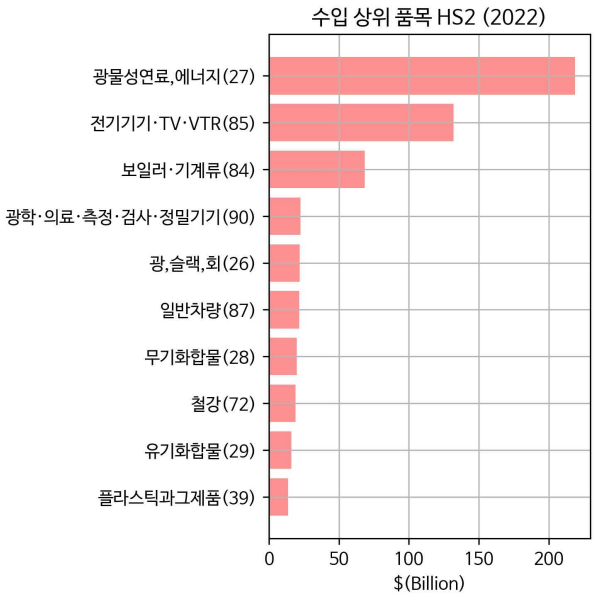


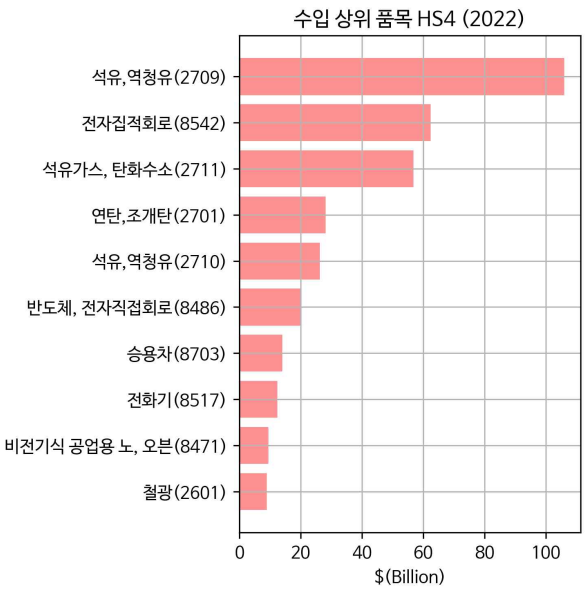
HSCODE 2자리, HSCODE 4자리 기준으로 2022년 수출액 상위 품목을 나타낸 그래프이다.
HSCODE2자리를 기준으로 보면 전자기기 · TV · VTR이 수출액의 큰 비중을 차지하고 있고 HSCODE4 자리를 기준으로 보면 전자집적회로가 큰 비중을 차지하고 있다.

2.3.2) 수입 상위 10개 품목

HS2	수입액	품목명(HS2)	HS4	수입액	품목명(HS4)
27	218.679	광물성연료,에너지(27)	2709	105.9635	석유,역청유(2709)
85	131.7612	전기기기·TV·VTR(85)	8542	62.40026	전자집적회로(8542)
84	68.42358	보일러·기계류(84)	2711	56.74928	석유가스, 탄화수소(2711)
90	22.67749	광학·의료·측정·검사·정밀기기(90)	2701	28.15428	연탄,조개탄(2701)
26	21.84794	광,슬랙,회(26)	2710	26.32512	석유,역청유(2710)
87	21.65207	일반차량(87)	8486	20.14234	반도체,전자직접회로(8486)
28	19.99641	무기화합물(28)	8703	14.04894	승용차(8703)
72	18.9438	철강(72)	8517	12.38908	전화기(8517)
29	16.10429	유기화합물(29)	8471	9.500684	비전기식 공업용 노, 오븐(8471)
39	13.81509	플라스틱과그제품(39)	2601	8.92045	철강(2601)

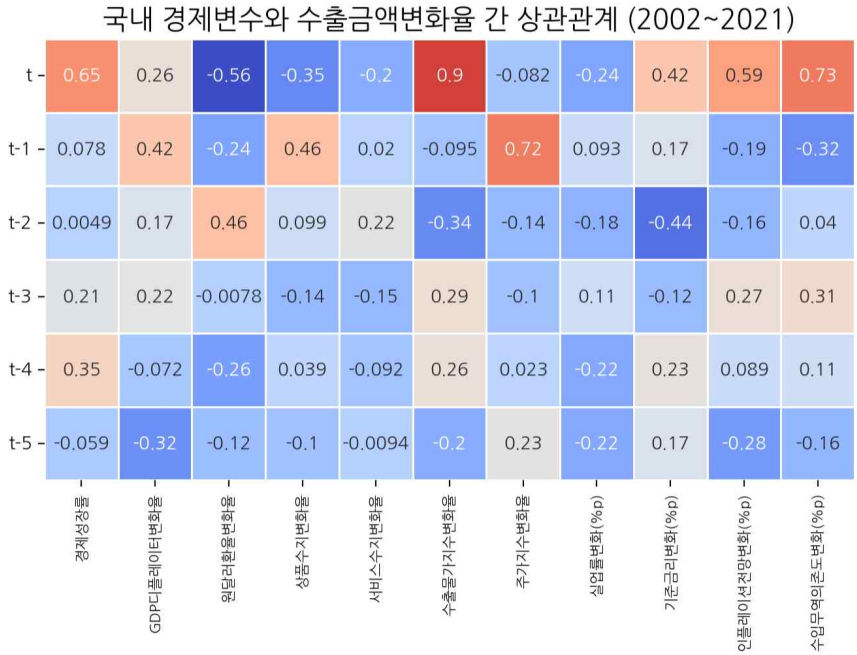
2022년을 기준으로 한국의 수입액 상위 10개 품목과 수출액을 나타낸 표이다. HSCODE 2자리, HSCODE 4자리를 기준으로 품목을 나눠서 수입액을 내림차순 정렬하였다.





HSCODE 2자리, HSCODE 4자리 기준으로 2022년 수입액 상위 품목을 나타낸 그래프이다.
HSCODE2자리를 기준으로 보면 광물성연료, 에너지가 수입액의 큰 비중을 차지하고 있고 HSCODE4 자리를 기준으로 보면 석유, 역청유가 큰 비중을 차지하고 있다.

2.4) 경제변수와 수출 간의 관계



수출금액 변화율에 가장 큰 양의 상관관계를 가지는 변수는 당해 년도의 수출물가지수 변화율이었다. 수출품의 가격이 상승하면 경쟁국들과의 가격 경쟁력이 떨어져 수출이 감소한다는 경제학적 이론과는 다른 결과이다. 이는 우리나라가 반도체나 차량용 부속품 등등 주로 가격 비탄력적인 품목을 수출하고 있는 것으로 분석할 수 있다.

그 다음으로 큰 양의 상관관계를 가지는 변수는 당해년도의 수입무역의존도 변화율이었다. 원자재·중간재 의존도가 높은 우리나라의 특성상 자연스러운 현상으로 보인다. 이전 연도들의 수입무역의존도 변화율은 당해 수출금액변화율과 상관관계가 거의 없었다.

전년도 주가지수변화율 또한 수출금액변화율과 높은 양의 상관관계를 보였다. 시장전반에 대한 기대와 정보를 반영한 기대이론에 따라 당해년도보다는 전년도 수치와 높은 관계를 맺는것으로 보인다. 이에 따라, 수출 금액을 예측하기 위해 전년도 주가지수 변화율을 고려할 필요가 있을것으로 보인다.

당해년도의 경제성장률은 수출금액변화율과 0.65의 양의 상관관계를 보였다. 이는 국가 단위의 생산성의 변화가 수출액에도 영향을 미친것으로 보인다. 다만, 수출금액 변화율이 경제성장률에 영향을 미칠 가능성 또한 다분히 높다. 이전 연도들의 경제성장률은 당해 수출금액변화율과 상관관계가 거의 없었다.

전년도 GDP디플레이터는 당해년도의 수출금액변화율과 0.42의 양의 상관관계를 보였다. 전년도 GDP 디플레이터의 상승이 수출물가의 상승을 야기시키고, 이에따라 수출금액이 증가하는 것으로 보인다.

당해년도의 원/달러환율 변화율은 당해년도의 수출금액변화율과 -0.56의 음의 상관관계를 보였다. 당해년도의 수출금액의 증가가 곧바로 원/달러환율 감소에 영향을 미쳤을 것으로 보인다 . t-2기의 원/달러환율 변화율은 수출금액변화율과 0.46의 양의 상관관계를 보였다. 이는 원/달러환율의 상승이 수출금액 증가에 영향을 미치는 데에 2년 정도의 시차가 있음으로 보인다.

당해년도의 기준금리변화는 수출금액변화율과 0.42의 양의 상관관계를 보였다. 이는 수출금액 증가로 인한 통화량 증가가 인플레이션을 야기시키고 이것이 정부의 기준금리 인상 요인에 영향을 미친것으로 보인다. t-2기의 기준금리 변화는 수출금액변화율과 -0.44의 음의 상관관계를 보였다. 기준금리 하락이 국내 기업들의 수출금액 상승에 영향을 미치는데에 2년 정도의 시차가 있음으로 보인다.

당해년도의 인플레이션전망 변화는 수출금액변화율과 0.59의 양의 상관관계를 보였다. 인플레이션전망의 상승은 수출물가의 상승을 야기시키고, 이에따라 수출금액이 증가하는 것으로 보인다.

상품수지변화율, 서비스수지변화율, 실업률 변화 등은 수출금액변화율과 낮은 상관관계를 보였다.

2.5) 뉴스 데이터 확인

경제 섹션에 해당하는 뉴스 데이터를 전부 수집하며 수집한 뉴스 데이터는 키워드를 통해 필터링한 뒤, 해당 뉴스 텍스트에 대해 토큰화와 취합 과정을 거쳐 모델이나 시각화가 가능한 형태로 가공한다.

두 국가를 합쳐 대한민국 수출액의 절반 이상(2022년 기준)을 차지하는 중국과 미국을 키워드로 하여 기사를 필터링하고 이를 통해 워드클라우드를 글리면 위와 같다. 불용어는 제거하였고, 각각 '미국', '중국'과 '수출'이라는 키워드가 포함된 기사를 선정하여 명사만을 추출 후 Word Count를 실행하여 생성하였다.

'수출', '미국', '중국'의 경우 키워드이므로 높은 빈도를 보인 것은 당연하며, 두 국가 모두에서 나타나는 주요 단어가 유사한 것을 확인할 수 있다. 이는 무역 분쟁으로 인하여 두 국가가 동시에 등장하는 기사의 빈도가 크게 늘어난 것이 영향을 주었기 때문으로 해석할 수 있다.

뉴스 기사 기반의 모델에 적용할 시에는 위와 같은 방식으로 필터링 후 명사만을 추출한 데이터에 대해, Word Count를 실시하고 해당 결과물을 월별로 합산하여 Bag of Word를 만드는 형태로 이뤄진다.

3) 예측 모형
3.1) 예측 모형 개요

3.1.1) 예측 모형
본 보고서의 추천 시스템은 3개의 예측 모형의 결과물을 가중평균하여 타깃 시점의 수출액 변동 방향을 예측하는 방식으로 운영된다. 따라서 이에 해당되는 예측 모형들은 t기 시점 이전의 데이터를 사용하여 t기 시점의 종속변수를 예측하는 것으로 정의된다. 각각의 예측 모형은 다음과 같다.
1. 과거의 수출액 변동을 기반으로 미래의 국가-품목 조합별 수출액 변동을 예측하는 자기회귀모형
2. 과거의 뉴스 데이터를 기반으로 미래의 국가별 수출액 변동을 예측하는 자연어 모델
3. 과거의 뉴스 데이터를 기반으로 미래의 품목별 수출액 변동을 예측하는 자연어 모델

가장 첫 번째 모델은 이전 시기의 종속변수를 사용하여 t기의 종속변수 변동을 예측하는 모형이다. 의사결정나무 알고리즘을 사용한다. 일반적으로 안정적인 기간에서는 과거의 종속변수로 미래의 종속변수를 설명하는 것이 유의할 가능성이 높다. 가장 단순하지만 안정적인 시기의 트렌드를 예측할 수 있을 것으로 기대되어 모형에 포함하였다.

두 번째 모형은 필터링한 뉴스 데이터를 기반으로 각각 국가별 수출 증감 여부를 예측하는 모형이다. 랜덤포레스트 알고리즘을 사용한다. 구체적으로는 한국에서 해당 국가로의 수출 금액이 전년동월대비 증감하였는지 여부를 예측한다. 국가 명칭과 '수출'을 키워드로 하여 두 키워드가 모두 포함된 기사 중, 1000자 이상의 길이를 가진 것들을 데이터로 사용한다.
수집한 전체 기사를 필터링한 뒤, 걸러진 기사에서 명사만을 추출하고 이를 월별로 병합하여 단어와 등장 횟수를 기반으로 BOW(Bag of Words: 각 열이 하나의 단어를 의미하고, 각 행의 값들은 해당하는 열의 단어가 문서 내에서 몇 번 등장하였는지를 나타내는 데이터)를 생성하여 이를 통해 분석을 실행한다.

세 번째 모형은 두 번째 모형과 유사한 방식으로 작동하지만, 키워드와 종속변수가 변경된 모형이다. 랜덤포레스트 알고리즘을 사용한다. 필터링한 뉴스 데이터를 기반으로 HS코드 4자리 기준(이하 HS4) 한국의 해당 품목의 상품 수출액이 전년동월대비 증감하였는지 여부를 예측한다. 품목별 예측 모형의 경우 HS4 코드의 대표적 상품명을 GPT를 통해 추출한 뒤 이를 가공하여 4~6개의 키워드가 선정된다.

분석에 사용되는 모든 랜덤 모듈에는 반복 수행에서의 결과 변동을 방지하기 위해 Random Seed를 0으로 할당하여 실행하였다.

3.1.2) 예측 대상(국가-품목) 선정
최종 예측 대상은 국가와 품목의 그룹에 따라 구분된다. 예를 들어 중국에 대한 HS코드 4자리 기준 8473에 대항하는 품목의 수출금액 증감을 예측하는 식이다.

자기회귀 모형은 모든 그룹에 대해 각각 개별적으로 예측을 진행한다. 국가별 모형은 해당 국가로의 수출금액 증감 여부만을 예측하며 품목별 구분은 이뤄지지 않는다. 품목별 모형은 해당 품목의 수출금액 증감 여부만을 예측하여 국가별 구분은 이뤄지지 않다.
이는 뉴스 데이터를 필터링 할 때, 유의미한 독립변수를 제작하기 위한 충분한 수의 뉴스 데이터를 확보하기 위해 그룹별 하나의 모형이 아닌 2개의 모형으로 분리하여 학습한다.

모든 HS 4자리 코드의 품목과 국가에 대해 분석을 실행할 수 있다면 이상적이겠으나, 현실적으로 키워드와 뉴스 기사를 추출해야 하기 때문에, 본 분석에서는 예측 모형을 훈련할 대상 국가와 품목을 지정하고 모형 훈련을 진행하고자 한다.

분석 대상이 되는 국가는 2022년 수출금액 기준 상위 10개국이며, 분석 대상 품목은 이들 국가로 2013~2022년의 10년간 수출되었던 품목(HS코드 기준 4자리) 중 수출액수 상위 5개 품목을 선정하여

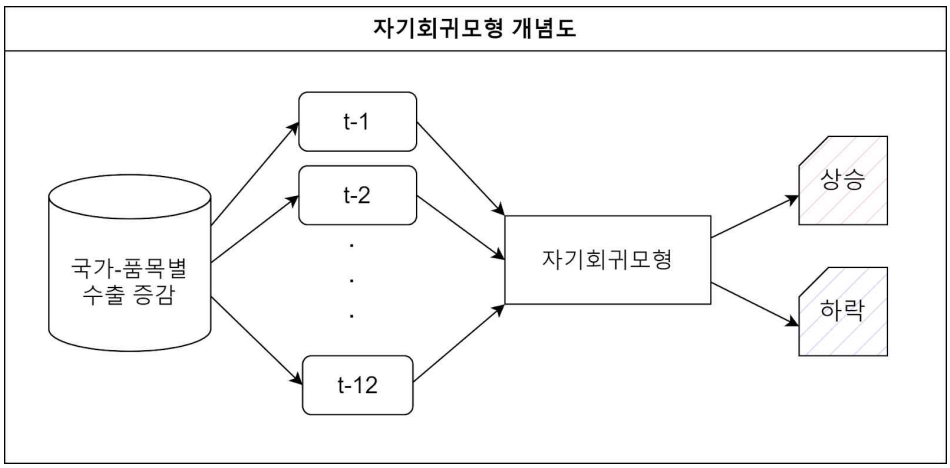
진행한다. 즉 총 분석 대상이 되는 그룹은 5*10=50, 50 가지가 된다.
해당하는 국가와 품목은 아래 표와 같다.

순위	중국	미국	베트남	일본	홍콩
1	전자집적회로(8542)	승용차(8703)	전자집적회로(8542)	석유,역청유(2710)	전자집적회로(8542)
2	레이저기기(9013)	차량 부속품(8708)	전화기(8517)	전자집적회로(8542)	전화기(8517)
3	환식탄화수소(2902)	전화기(8517)	석유,역청유(2710)	전화기(8517)	보일러 부속품(8473)
4	석유,역청유(2710)	보일러 부속품(8473)	전자기기 부속품(8529)	은(7106)	석유,역청유(2710)
5	반도체, 전자직접회로_8486	석유,역청유(2710)	인쇄회로(8534)	비도금 강판(포일)(7208)	순항선,유람선,화물선(8901)

순위	호주	대만	인도	싱가포르	멕시코
1	석유,역청유_2710	전자집적회로(8542)	차량 부속품(8708)	석유,역청유(2710)	차량 부속품(8708)
2	승용차_8703	석유,역청유(2710)	전자집적회로(8542)	전자집적회로(8542)	전자기기 부속품(8529)
3	특수 선박_8905	환식탄화수소(2902)	전화기(8517)	순항선,유람선,화물선(8901)	레이저기기(9013)
4	철강 구조물_7308	반도체, 전자직접회로(8486)	석유,역청유(2710)	특수 선박(8905)	도금 강판(포일)(7210)
5	차량 부속품-8708	인쇄회로(8534)	비도금 강판(포일)(7208)	반도체, 전자직접회로(8486)	승용차(8703)

3.2) 자기 회귀 모형

3.2.1) 모형 개요



자기회귀 모형의 경우 T-1 시점부터 T-12 시점까지의 월별 종속변수(수출액 변화율)를 독립변수로 하고 T시점의 종속변수를 예측하는 모형이다. Decision Tree 알고리즘을 사용하여 생성하였고, max_depth를 4로 제한하여 pre-pruning 작업을 수행하였다.
1995~2021년까지의 데이터를 훈련용으로 사용하였고, 이 중 20%를 사전에 랜덤으로 추출하여 Accuacy와 ROC AUC Score의 평가지표를 생성하는 데 사용하였다. 추천 시스템에서는 2022년 1월부터 2023년 5월까지의 예측 값을 생성하여 사용하였다.

3.2.2) 모형 적합 결과

국가별 수출액 상위 5개 품목에 대한 그래프를 나타냈다. 품목명은 HSCODE 4자리를 기준으로 지정해줬고 수출액 기준으로 내림차순 정렬 후 표를 생성했다.
국가별 첫번째 표는 품목별 t-1시점부터 t-12시점까지의 feature importance를 나타냈고 두번째 표

는 품목별 Train data수, Test data수, accuaracy, roc_auc score를 나타냈다.

(1) 중국

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	전자집적회로(8542)	**0.731**	0.055	0	0	0.033	0	0.023	0	0	0.047	0	0.111
2	레이저기기(9013)	**0.758**	0.05	0.088	0	0.007	0	0	0	0	0	0.02	0.08
3	환식탄화수소(2902)	**0.609**	0.026	0.08	0.02	0.025	0	0.06	0	0.024	0	0.11	0.044
4	석유,역청유(2710)	**0.652**	0.15	0	0	0	0	0	0.051	0	0.026	0	0.121
5	반도체, 전자직접회로(8486)	**0.375**	0	0.055	0.05	0.038	0.1	0	0	0	0.066	0	0.287

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	전자집적회로(8542)	240	60	0.917	0.856
2	레이저기기(9013)	240	60	0.8	0.844
3	환식탄화수소(2902)	240	60	0.7	0.616
4	석유,역청유(2710)	240	60	0.817	0.872
5	반도체, 전자직접회로(8486)	124	32	0.688	0.698

대중 상위 5개의 품목의 feature importance는 t-1기에서 가장 높게 나왔다. 다만, 반도체의 경우 t-12기의 feature importance가 t-1기의 수치보다 약간 낮은 수준으로 나왔다. 정확도는 대체적으로 0.7~0.8 정도도의 수준을 보이고 있고, 특히 전자집적회로의 경우 0.9이상의 높은 정확도를 보인다. roc_auc의 경우, 0.6~0.9 정도의 수준을 보였다.

(2) 미국

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	승용차(8708)	240	60	0.683	0.642
2	차량 부속품(8529)	240	60	0.75	0.662
3	전화기(9013)	240	60	0.683	0.686
4	보일러 부속품(7210)	240	60	0.867	0.89
5	석유,역청유(8703)	240	60	0.6	0.662

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	승용차(8708)	240	60	0.683	0.642
2	차량 부속품(8529)	240	60	0.75	0.662
3	전화기(9013)	240	60	0.683	0.686
4	보일러 부속품(7210)	240	60	0.867	0.89
5	석유,역청유(8703)	240	60	0.6	0.662

대미 상위 5개의 품목의 feature importance는 t-1기에서 가장 높게 나왔다. 다만, 석유, 역청유유의 경우 t-12기의 feature importance가 t-1기의 수치보다 약간 낮은 수준으로 나왔다. 정확도는 0.6~0.8 정도의 수준을 보이고 있다. roc_auc 의 경우 0.6~0.7 정도의 수준을 보이고 있다. 보일러 부속품이 가장 높은 정확도와 roc_auc 점수를 보였다.

(3) 베트남

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	전자집적회로(8542)	224	56	0.821	0.8
2	전화기(8517)	239	60	0.733	0.7
3	석유,역청유(2710)	240	60	0.65	0.8
4	전자기기 부속품(8529)	239	60	0.767	0.8
5	인쇄회로(8534)	190	48	0.75	0.7

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	전자집적회로(8542)	**0.394**	0.03	0.125	0	0.179	0	0	0.054	0	0	0	0.219
2	전화기(8517)	**0.34**	0	0	0.11	0	0.1	0	0	0	0.134	0.1	0.257
3	석유,역청유(2710)	0.111	0	0.034	0.03	0.152	0.1	0.101	0	0.067	0.051	0.02	**0.377**
4	전자기기 부속품(8529)	**0.579**	0	0	0.07	0	0	0.039	0	0	0.097	0.08	0.108
5	인쇄회로(8534)	0.097	**0.29**	0.026	0	0.112	0.1	0.057	0	0.04	0.071	0.08	0.162

대베트남 상위 5개 품목 중 석유와 역청유, 인쇄회로를 제외한 품목들의 feature_importance는 t-1기에서 가장 높게 나왔다. 석유와 역청유의 경우 t-12기, 인쇄회로의 경우 t-2기의 feature_importance가 가장 높았다. 대체적으로 모든 품목의 feature_importance가 낮았다. 정확도는 0.6~0.8 수준을 보였다. roc_auc는 0.7~0.8 정도의 수준을 보였다.

(4) 일본

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	석유,역청유(2710)	0.2	**0.56**	0.04	0.08	0.058	0	0	0.015	0	0.026	0	0.028
2	전자집적회로(8542)	**0.732**	0.04	0.041	0	0	0	0	0.025	0.106	0	0	0.061
3	전화기(8517)	**0.568**	0.11	0	0	0	0	0.065	0	0.117	0	0.08	0.063
4	은(7106)	0.203	**0.52**	0	0.02	0	0.1	0	0.134	0	0	0.02	0.048
5	비도금 강판(포일)(7208)	**0.859**	0	0	0.02	0.038	0	0	0	0	0.02	0	0.068

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	석유,역청유(2710)	240	60	0.7	0.748
2	전자집적회로(8542)	240	60	0.833	0.869
3	전화기(8517)	240	60	0.783	0.82
4	은(7106)	240	60	0.65	0.654
5	비도금 강판(포일)(7208)	240	60	0.733	0.801

대일 상위 5개 품목 중 석유와 역청유, 은을 제외한 품목들의 feature_importance는 t-1기에서 가장 높게 나왔다.
석유와 역청유, 인쇄회로의 경우 t-2기의 feature_importance가 가장 높았다. 정확도는 0.7~0.8 정도의 수준을 보였다. roc_auc는 0.7~0.8정도의 수준을 보였다.

(5) 홍콩

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	전자집적회로(8542)	**0.723**	0	0	0.021	0.024	0	0	0	0.081	0.063	0	0.058
2	전화기(8517)	**0.652**	0.045	0.038	0.03	0	0	0.11	0	0	0	0	0.078
3	보일러 부속품(8473)	**0.515**	0.083	0.052	0	0.095	0	0	0	0.095	0.067	0	0.093
4	석유,역청유(2710)	**0.326**	0.08	0	0	0	0	0.103	0	0.072	0.072	0.13	0.217
5	순항선,유람선,화물선(8901)	0.094	0.035	0.237	0.155	0	0	0	0	0	0.189	0	**0.291**

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	전자집적회로(8542)	240	60	0.617	0.491
2	전화기(8517)	240	60	0.767	0.797
3	보일러 부속품(8473)	240	60	0.75	0.784
4	석유,역청유(2710)	240	60	0.65	0.667
5	순항선,유람선,화물선(8901)	148	37	0.541	0.653

홍콩 수출 상위 5개 품목 중 순항선, 유람선, 화물선을 제외한 품목들의 feature_importance는 t-1기에서 가장 높게 나왔다. 순항선, 유람선, 화물선의 경우 t-12기의 feature_importance가 가장 높았다. 정확도는 0.6~0.8 정도의 수준을 보였다. roc_auc는 0.6~0.8 정도의 수준을 보였다. 다만 전자집적회로의 경우 roc_auc 점수가 0.5 이하로 정확도와 비교해 낮은 수치를 보였다.

(6) 대만

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	전자집적회로(8542)	**0.831**	0	0	0.024	0	0	0.024	0	0.016	0.025	0.08	0
2	석유,역청유(2710)	**0.293**	0.049	0	0.056	0	0	0.177	0	0.068	0.078	0.05	0.233
3	환식탄화수소(2902)	**0.356**	0.057	0.098	0.057	0	0	0.13	0	0	0.075	0	0.227
4	반도체, 전자직접회로(8486)	0.048	0.18	0.081	**0.338**	0	0	0.091	0	0	0.051	0	0.211
5	인쇄회로(8534)	**0.696**	0	0	0.047	0.028	0	0.055	0.049	0	0	0.03	0.059

순위	품목명	N_Train	N_Test	accuracy	roc_auc
1	전자집적회로(8542)	240	60	0.817	0.925
2	석유,역청유(2710)	240	60	0.517	0.428
3	환식탄화수소(2902)	240	60	0.683	0.728
4	반도체, 전자직접회로(8486)	124	32	0.688	0.721
5	인쇄회로(8534)	240	60	0.633	0.576

상위 5개 품목 중 반도체와 전자직접회로를 제외한 품목들의 feature_importance는 t-1기에서 가장 높게 나왔다. 반도체와 전자집적회로의 경우 t-4기의 feature_importance가 가장 높았다. 정확도는 0.6~0.7 정도의 수준을 보였다. 다만, 석유와 역청유의 경우 0.5 정도의 비교적 낮은 정확도를, 전자집적회로의 경우 0.8 정도의 높은 정확도를 보였다. roc_auc는 0.6~0.7 정도의 수준을 보였지만 이 또한 석유와 역청유는 0.4 정도의 비교적 낮은 수치를, 전자집적회로의 경우 0.9 정도의 높은 수치를 보였다.

(7) 싱가포르

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	석유,역청유(2710)	0.245	0.007	0.205	0	0	0	0.038	0	0	0.228	0	**0.276**
2	전자집적회로(8542)	**0.693**	0.063	0.019	0.043	0	0	0	0.071	0	0	0	0.08
3	순항선,유람선,화물선(8901)	0	0	0	0.089	0.235	0	0.141	0	0.052	0.122	0	**0.361**
4	특수 선박(8905)	0	0	0	0	0.192	0	0	0	0.147	0	0	**0.662**
5	반도체, 전자직접회로(8486)	**0.352**	0	0	0.131	0	0	0.121	0.091	0.048	0	0	0.257

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	석유,역청유(2710)	240	60	0.667	0.734
2	전자집적회로(8542)	240	60	0.733	0.784
3	순항선,유람선,화물선(8901)	201	51	0.647	0.735
4	특수 선박(8905)	23	6	0.667	0.75
5	반도체, 전자직접회로(8486)	124	32	0.656	0.491

상위 5개 품목 중 전자집적회로, 반도체와 전자직접회로는 t-1기에서 그 외의 품목들은 t-12기에서 feature_importance가 가장 높았다. 다만, 대체적으로 낮은 수치를 보였다. 정확도는 0.6~0.7 정도의 수준을 보였다. roc_auc는 0.7~0.8 정도의 수준을 보였다. 반도체와 전자직접회로의 경우 정확도에 비해 0.5 이하의 낮은 roc_auc 수치를 보였다.

(8) 인도

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	차량 부속품(8708)	**0.667**	0	0.071	0	0.027	0	0.054	0	0	0.075	0.02	0.053
2	전자집적회로(8542)	**0.53**	0	0.271	0	0	0	0.044	0.026	0	0.052	0	0.077
3	전화기(8517)	**0.447**	0.151	0	0.037	0.048	0	0.083	0	0.149	0	0.06	0.026
4	석유,역청유(2710)	**0.543**	0.23	0	0.088	0	0.1	0	0	0	0	0	0.064
5	비도금 강판(포일)(7208)	**0.393**	0.217	0	0.049	0	0	0	0.11	0.029	0.062	0.09	0

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	차량 부속품(8708)	240	60	0.667	0.608
2	전자집적회로(8542)	240	60	0.733	0.734
3	전화기(8517)	240	60	0.6	0.632
4	석유,역청유(2710)	240	60	0.7	0.748
5	비도금 강판(포일)(7208)	240	60	0.65	0.716

인도의 경우 상위 5개의 품목의 feature importance는 t-1기에서 가장 높게 나왔다. 정확도는 0.6~0.7 정도의 수준을 보였다. roc_auc 또한 0.6~0.7정도의 수준을 보였다.

(9) 호주

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	연탄,조개탄(2710)	0.203	0.133	0	0	0.047	0.3	0	0.054	0	0	0	**0.241**
2	승용차(8703)	**0.447**	0.185	0.015	0.053	0.052	0	0	0.029	0	0.064	0.06	0.091
3	철강 구조물(7308)	0.077	0.141	0.06	0	0.253	0	0	0	0.044	0	0	**0.426**
4	차량 부속품(8708)	0.067	**0.406**	0.096	0	0.017	0	0	0.04	0.088	0	0.06	0.224

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	연탄,조개탄(2710)	239	60	0.65	0.661
2	승용차(8703)	240	60	0.6	0.705
3	철강 구조물(7308)	236	59	0.576	0.544
4	차량 부속품(8708)	240	60	0.8	0.809

상위 4개의 품목 중 연탄과 조개탄, 철강 구조물 등은 t-12기에서, 승용차는 t-1기에서, 차량 부속품은 t-2기에서 feature_importance가 가장 높게 나왔다. 다만, 연탄과 조개탄의 경우 0.2 정도의 매우 낮은 수치를 보였다. 정확도는 0.6~0.7 정도의 수준을 보였다. roc_auc는 0.6~0.7 정도의 수준을 보였다. 차량 부속품의 경우 0.8의 비교적 높은 정확도와 roc_auc 수치를 보였다.

(10) 멕시코

순위	품목명(HS4)	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
1	차량 부속품(8708)	**0.425**	0.166	0.034	0.11	0.033	0	0	0	0	0.135	0.03	0.065
2	전자기기 부속품(8529)	**0.729**	0.051	0.058	0	0	0	0	0	0.019	0.019	0	0.091
3	레이저기기(9013)	**0.671**	0.051	0.116	0	0	0	0.027	0.07	0	0	0	0.047
4	도금 강판(포일)(7210)	**0.164**	0.369	0.047	0	0.048	0	0.045	0	0.107	0	0.12	0.099
5	승용차(8703)	**0.451**	0	0.053	0	0	0	0.072	0	0.055	0.089	0.14	0.138

순위	품목명(HS4)	N_Train	N_Test	accuracy	roc_auc
1	차량 부속품(8708)	240	60	0.733	0.68
2	전자기기 부속품(8529)	240	60	0.85	0.88
3	레이저기기(9013)	229	58	0.793	0.85
4	도금 강판(포일)(7210)	234	59	0.729	0.71
5	승용차(8703)	208	52	0.712	0.73

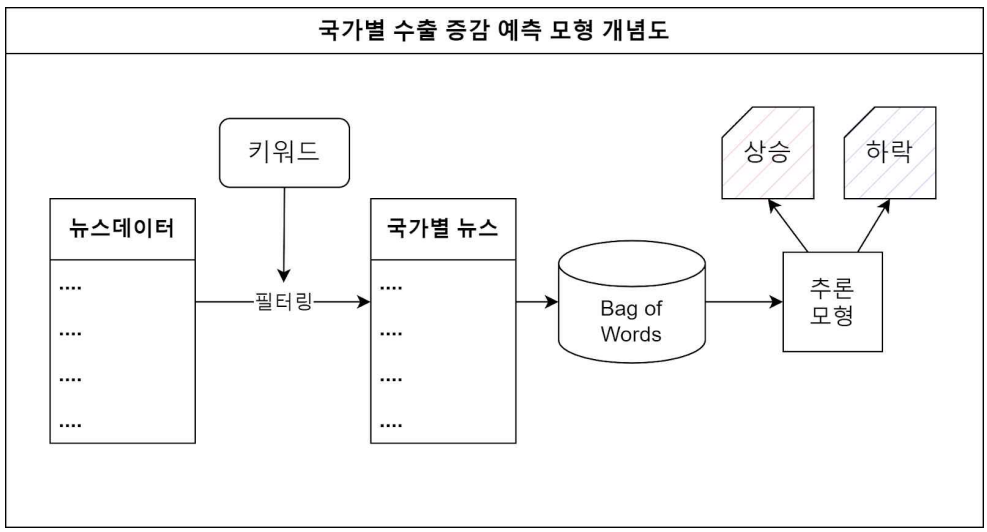
멕시코의 경우 상위 5개의 품목의 feature importance는 t-1기에서 가장 높게 나왔다. 다만, 도금 강판의 경우 0.2 이하의 매우 낮은 수치를 보였다. 정확도는 0.7~0.8 정도의 수준을 보였다. roc_auc는는0.7~0.9 정도의 수준을 보였다.

3.2.3) 모형 평가

대부분의 국가-HS4품목 그룹에서 자기회귀모형은 t-1 시점의 종속변수를 분류에 유용하게 활용하였다. 평가지표의 경우 정확도와 ROC AUC Score는 그룹에 따라 다르게 나타났는데, 일반적으로 점수가 높은 그룹은 0.7후반 ~ 0.8 정도의 점수를 나타낸 것에 비해, 점수가 낮은 그룹은 0.6 내외의 평가지표를 나타내어 유의미하다고 보기 어려운 결과를 나타냈다. 이는 각 그룹이 포함하는 국가와 품목의 특징이 분명이 구분되는 고유한 것이므로, 해당 특성들이 자기회귀의 유의성에 반영되었다고 추론할 수 있다.

3.3) 뉴스 기반 국가별 수출 증감 예측 모형

3.3.1) 모형 개요

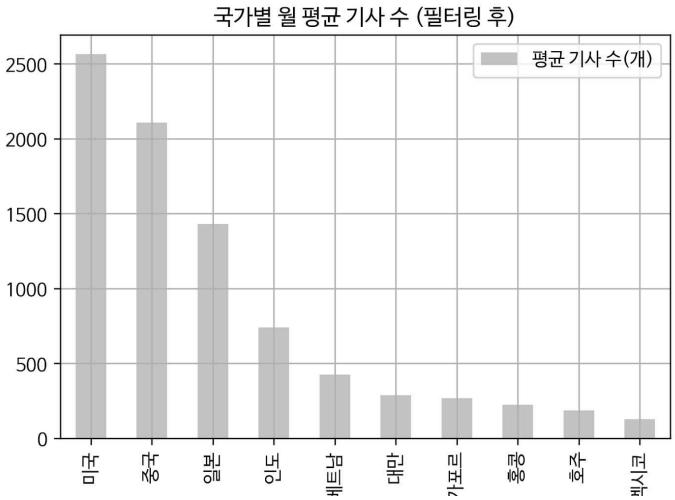


국가별 수출 증감 예측 모형의 경우, 뉴스 데이터를 키워드를 기준으로 필터하여 국가별로 별도의 뉴

스 데이터셋을 생성하고, 이를 기반으로 명사만을 추출 후 Word Count를 거쳐 Bag of Word를 생성한다. 이 데이터를 Random Forest 알고리즘에 적용하여 해당 시점의 수출액 변동 방향을 예측하는 형태로 구성된다. 예측 모형이기에 1기(1개월) 전의 뉴스 Bag of Words 벡터를 받아 해당 기의 수출액 변동 방향을 예측한다. Random Forest 모형은 estimator의 수를 500으로 설정하였으며, 내부적인 샘플링 시 사용하는 Randm seed는 0으로 설정하였다.

2018년부터 2023년 5월까지의 뉴스 데이터를 수집하였으며, 이 중 2018~2021년의 뉴스 데이터로 모델을 학습하고, 평가 지표를 생성하였다. 2022~2023년 5월의 데이터는 추천 시스템을 위한 예측값을 생성하는 데 사용하였다. 생성 시 사용하는 데이터의 개수가 제한적이었으므로, 결과로 도출된 예측 모형의 정확도가 상대적으로 낮게 나타날 수 있다.

3.3.2) 뉴스 데이터 필터링



국가별 뉴스의 경우 해당 국가의 정식 명칭과 '수출'이라는 단어를 모두 포함한 기사만을 필터링하여 사용한다. 위 그래프는 국가별 월 평균 기사 수를 나타낸 것이다.

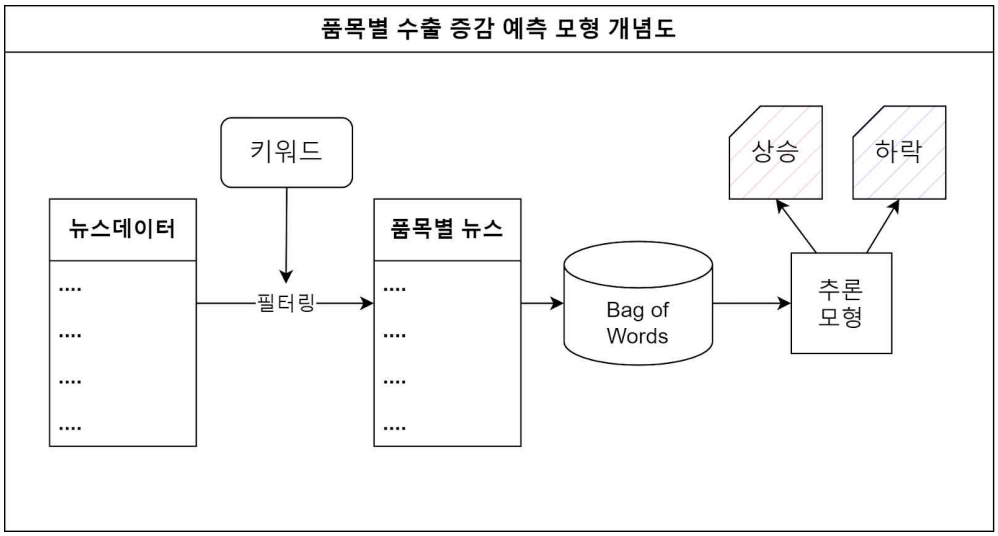
3.3.3) 모형 적합 결과

Country	N-Train	N-Test	True Ratio	Accuracy	ROC_AUC Score
중국	37	10	0.6	0.5	0.6
미국	37	10	0.6	0.6	0.7
베트남	37	10	0.5	0.6	0.8
일본	37	10	0.3	0.9	0.9
홍콩	37	10	0.4	0.7	0.8
대만	37	10	0.8	0.8	0.9
싱가포르	37	10	0.4	0.5	0.6
인도	37	10	0.6	0.8	0.9
호주	37	10	0.2	0.9	0.9
멕시코	37	10	0.5	0.5	0.8

수출액 상위 10개 국가에 대한 수출 증감 예측 모형의 결과 표이다. Train data의 수는 37개, Test data의 수는 10개로 지정하였다. 정확도는 평균 0.7정도의 수준을 보였고 roc_auc의 경우 평균 0.7정도의 수준을 보였다. 특히, 일본과 호주에서는 정확도와 roc_auc모두 0.9이상으로 높은 예측성능을 보였다.

3.4) 뉴스 기반 품목별 수출 증감 예측 모형

3.4.1) 모형 개요

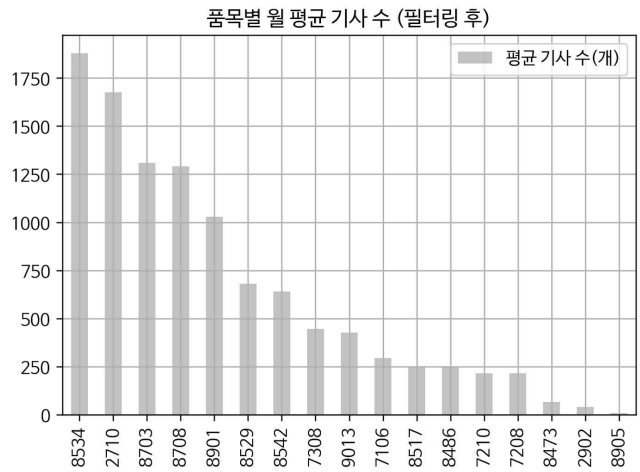


품목별 수출 증감 예측 모델의 경우, 기본적인 생성-평가-예측의 과정과 활용하는 데이터는 국가별 수출 예측 모형과 동일하지만, 기사를 필터링하는 과정에 있어 차이가 존재한다. 품목별 기사의 경우 HS 코드 4자리를 이용하여 GPT를 활용하여 생성한 키워드를 검색에 맞게 변형하여 사용하며, 키워드 중 하나라도 존재하는 기사를 필터링한다.

단 키워드가 일반적이기에 기사가 매우 많이 남게 되는 경우에는 '수출'을 무조건 포함하는 기사만을 남겨 분석을 용이하게 하였다. (해당 품목: 8703, 8542, 8517, 9013)

3.4.2) 뉴스 데이터 필터링

HS코드 4자리	키워드 (GPT 활용)
2710	"원유", "가솔린", "경유", "항공유", "중질유"
2902	"벤젠", "나프탈렌", "톨루엔", "사이클로헥세인", "아닐린"
7106	"실버", "은괴", "은박", "은사"
7208	"철판", "강판", "코일", "압연강판", "압연 강판"
7210	"철판", "강판", "코일", "도금 강판", "도금강판"
7308	"레일", "강철 레일", "강철 교량", "교량", "강철 틀"
8473	"디램", "비디오 카드", "사운드 카드", "통신 모듈", "디램 모듈"
8486	"연마기", "웨이퍼", "도포기", "포토레지스트", "성형기"
8517	"스마트폰", "휴대전화", "모뎀", "전화기", "라우터"
8529	"안테나", "오디오 케이블", "라디오", "튜너", "스피커"
8534	"인쇄회로", "인쇄회로 기판", "회로", "PCB", "기판"
8542	"프로세서", "메모리", "IC", "그래픽", "GPU"
8703	"승용차", "자동차", "SUV", "EV", "트럭"
8708	"엔진", "변속기", "자동차 내장", "내장재", "자동차 부품"
8901	"컨테이너선", "유조선", "선박", "화물선", "벌크선"
8905	"시추선", "구조선", "소방선", "기중기선", "채굴 선박"
9013	"액정", "디스플레이", "패널", "LCD", "OLED"



위 그래프는 키워드를 활용하여 필터링한 기사의 월 평균 숫자를 나타낸 것이다. 특수 선박(8905)이나 환식탄화수소(2902)와 같이 키워드 자체가 드물게 나타나는 품목의 경우 기사 수가 매우 부족한 것을 확인할 수 있다. 이는 모형의 정확도에도 영향을 미쳤을 것으로 예상된다.

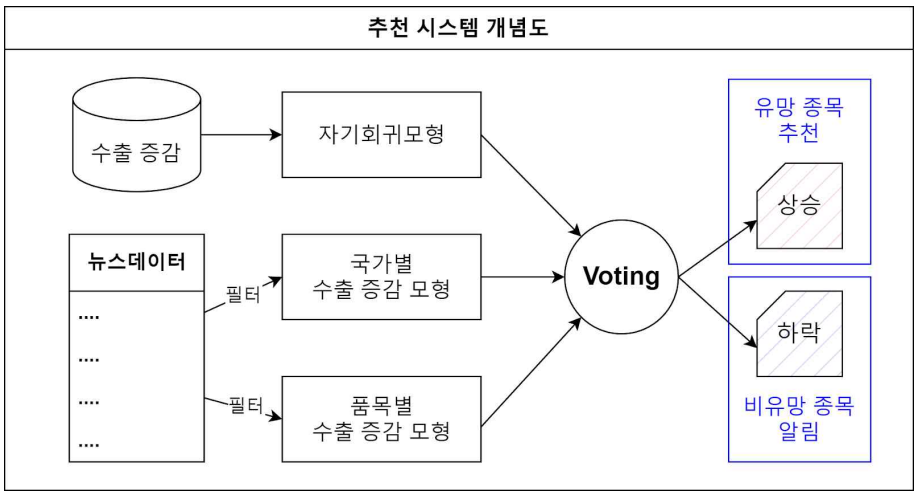
3.4.3) 모형 적합 결과

품목명(HS4)	N-Train	N-Test	True Ratio	Accuracy	ROC_AUC Score
석유,역청유(2710)	37	10	0.3	0.9	1
환식탄화수소(2902)	37	10	0.3	0.7	0.8
은(7106)	37	10	0.7	0.9	1
비도금 강판(포일)(7208)	37	10	0.4	0.6	0.8
도금 강판(포일)(7210)	37	10	0.4	0.7	0.5
철강 구조물(7308)	37	10	0.4	0.6	0.8
보일러 부속품(8473)	37	10	0.8	0.8	0.9
반도체, 전자직접회로(8486)	37	10	0.7	0.7	0.5
전화기(8517)	37	10	0.8	0.8	0.7
전자기기 부속품(8529)	37	10	0.4	0.7	0.5
인쇄회로(8534)	37	10	0.5	0.9	0.8
전자직접회로(8542)	37	10	0.5	0.6	0.8
승용차(8703)	37	10	0.7	0.5	0.5
차량 부속품(8708)	37	10	0.7	0.4	0.4
순항선,유람선,화물선(8901)	37	10	0.4	0.5	0.6
특수 선박(8905)	34	9	0.3333	0.6667	0.5
레이저기기(9013)	37	10	0.3	0.7	1

수출액 상위 10개 품목에 대한 수출 증감 예측 모형의 결과 표이다. Train data의 수는 37개, Test data의 수는 10개로 지정하였다. 정확도는 평균 0.7정도의 수준을 보였고 roc_auc의 경우 평균 0.7정도의 수준을 보였다. 특히, 석유,역청유와 은에서는 정확도와 roc_auc모두 0.9이상으로 높은 예측성능을 보였다.

4) 추천 시스템

4.1) 추천 시스템 개요



추천 시스템은 앞서 생성한 3개의 예측 모형을 앙상블 하는 형태로 구현된다. t+1기의 유망 종목을 알 아보기 위해선 해당 시스템에서는 t~t-11기의(12개월) 종목별 과거 수출액 변동률을 자기회귀모형에 적용하여 t기의 증감 여부를 확률로 도출하고, t기의 뉴스 데이터를 필터하여 국가별/품목별 수출액 증감 여부를 도출한다.

그리고 3개의 값을 가중평균한 확률값을 기준으로 t+1기에 해당 그룹(국가-종목의 조합)의 수출이 증가할지 여부를 판단한다. 증가할 것으로 예상되는 경우 유망 종목에 편입하고, 하락할 것으로 예상되면 비유망 종목으로 판단한다.

해당 시스템은 국가-품목 조합의 그룹별로 증감 여부를 예측하여 이를 기반으로 추천하는 시스템이므로, 3개 모형에 대해 적용하는 가중치는 자기회귀모형에 60%, 뉴스 기반의 두 모델에 각각 20%로 설정하였다. 뉴스 기반의 모형들은 데이터의 문제 때문에 매 그룹을 예측하기가 어렵고 이로 인하여 국가별/품목별 2개의 모형으로 나뉘었기에 가중치를 비교적 낮게 설정한 것이다.

4.2) 추천 시스템 평가

아래의 두 표는 추천 시스템으로 데이터가 결측되어 예측이 불가능한 4개를 제외한 46개 그룹(10개국가, 개별 상위 5개 종목)에 대해 2022~2023.05월 까지의 추천 시스템의 추천(예측) 결과와 실제 값을 기반으로 평가 지표를 생성한 것이다. 두 표는 각 46개 그룹을 국가별, 품목별로 평균낸 지표들이다.

순위	Country	true_ratio	acc	auc	acc_ar	acc_country	acc_hs4
1	중국	0.188	**0.800**	**0.855**	**0.765**	0.188	0.494
2	미국	0.600	0.671	0.639	0.647	0.600	0.588
3	베트남	0.341	**0.647**	**0.720**	**0.588**	0.341	0.424
4	일본	0.424	**0.682**	**0.520**	**0.553**	0.424	0.412
5	홍콩	0.250	0.706	0.534	0.721	0.250	0.250
6	호주	0.691	0.691	0.581	0.706	0.632	0.691
7	대만	0.435	**0.612**	**0.713**	**0.565**	0.435	0.471
8	인도	0.565	**0.624**	**0.645**	**0.588**	0.565	0.588
9	싱가포르	0.706	0.627	0.606	0.667	0.706	0.686
10	멕시코	0.435	0.729	0.702	0.729	0.435	0.553

평가지표 표의 변수 중 true_ratio는 예측할 대상 데이터 중 1의 값의 비율을, acc는 예측한 값의 정

확도를 auc는 ROC AUC Score를, acc_ar는 자기회귀모형의 정확도, acc_country는 국가별 모형의 정확도, acc_hs4는 품목별 모형의 정확도를 의미한다.

국가별로 정리된 지표를 볼 때 중국, 미국, 베트남, 일본, 대만, 인도의 경우 자기회귀모형을 사용할 때보다 평가지표가 개선되었다. 하지만 다른 국가들의 경우 자기회귀모형의 정확도를 앙상블을 거치며 떨어지게 되는 측면이 존재하였다.
이는 본 분석에서 사용한 텍스트 분석 방법으로 수출액과의 관계를 잡아낼 수 있는 그룹의 경우에는 국가별, 품목별 모형을 앙상블한 것이 평가지표의 개선으로 나타났지만, 반대의 경우 해당 그룹의 경우 필터링한 기사의 특성이 수출액을 예측하는 데 유의미한 관계를 찾기 어려웠을 것으로 예상된다.

순위	품목명(HS4)	true_ratio	acc	auc	acc_ar	acc_country	acc_hs4
1	석유,역청유(2710)	0.542	**0.601**	**0.559**	**0.595**	0.523	0.542
2	환식탄화수소(2902)	0.353	0.588	0.688	0.588	0.353	0.706
3	은(7106)	0.059	**0.824**	**0.094**	**0.706**	0.059	0.059
4	비도금 강판(포일)(7208)	0.559	**0.676**	**0.703**	**0.647**	0.559	0.559
5	도금 강판(포일)(7210)	0.353	0.882	0.985	0.882	0.353	0.412
6	철강 구조물(7308)	0.471	0.412	0.708	0.529	0.529	0.529
7	보일러 부속품(8473)	0.353	**0.706**	**0.774**	**0.676**	0.353	0.353
8	반도체, 전자직접회로(8486)	0.373	**0.686**	**0.731**	**0.667**	0.373	0.353
9	전화기(8517)	0.235	0.600	0.395	0.424	0.235	0.235
10	전자기기 부속품(8529)	0.029	0.794	0.563	0.794	0.029	0.441
11	인쇄회로(8534)	0.382	0.618	0.787	0.618	0.382	0.382
12	평판디스플레이 모듈(8542)	0.529	**0.756**	**0.791**	**0.731**	0.529	0.529
13	승용차(8703)	0.784	**0.804**	**0.833**	**0.784**	0.725	0.784
14	차량 부속품(8708)	0.735	**0.706**	**0.627**	**0.662**	0.750	0.706
15	레이저기기(9013)	0.206	0.706	0.500	0.765	0.206	0.794

품목별 평가의 경우에도 국가별과 마찬가지로 추천 시스템으로 Soft Voting 시 정확도가 자기회귀모형에 비해 개선되는 경우가 그렇지 않은 경우가 큰 차이를 보였다. 앞선 결과와 마찬가지로 해당 그룹에 대해 필터링되어 사용되는 기사의 특성이 수출액의 방향 예측에 유의미했는지 여부를 간접적으로 판단할 수 있을 것으로 보인다.

5) 결론

5.1) 최초 가설의 평가

최초에 설정했던 가설 '자기회귀모형과 뉴스 데이터를 활용한 모형이 국가-품목별 수출액을 월 스케일에서 예측하는 데 사용될 수 있을 것이다' 의 경우 부분 채택이 가능할 것으로보인다. 분석의 결과를 보면 예측하고자 하는 그룹이 어떻게 되는 것인지에 따라 극명하게 다른 결과를 나타내었다.

예를 들어 국가의 경우 중국, 베트남, 일본과 같은 국가는 자기회귀모형보다 텍스트 모형을 앙상블한 결과물이 더 높은 정확도를 나타낸 것에 비해 홍콩, 호주, 싱가포르의 경우 오히려 정확도가 악화되었다. 품목의 경우에도 이는 마찬가지로 은(7106)이나 승용차(8703)과 같은 품목은 평가지표가 개선되었지만, 자기회귀모형과 다를 바 없는 정확도를 가진 모형 또한 존재하였다.

이로부터 도출할 수 있는 결론은, Bag of Words를 생성하여 Random Forest 모형을 통해 적합시키는 예측 시스템으로 한정할 때. 뉴스 데이터를 도입하는 것이 자기회귀 모형과 비교하여 더 높은 정확도를 나타내는 그룹과 그렇지 않은 그룹이 존재하며. 뉴스 데이터를 적용하여 추론하고자 한다면 이 부분을 고려하여 적절한 그룹을 선정하는 것이 필요할 것으로 보인다는 것이다.

5.2) 한계점과 향후 개선 방향

본 분석의 가장 명확한 한계점은 뉴스 데이터의 부족과 고도화하지 못한 자연어처리 모델에 존재한다. 뉴스 데이터의 경우 2018~2023.05 까지의 뉴스를 수집하여 사용하였지만, 이는 경제적인 사이클로 볼 때 아주 짧은 트렌드만을 확인할 수 있는 기간이고 코로나19로 인한 팬데믹 상황과 러우 전쟁으로 인한 불확실한 세계 경제의 충격과 같은 변동성이 큰 기간을 포함하였기에. 대표성이 있는 데이터로 보기에 어려움이 존재한다.

또한 데이터 저장과 처리, 분석 환경의 현실적인 한계에 직면하여. BERT와 같은 고도화된 딥러닝 모델을 사용하여 자연어 처리를 실시하지 못한 것 또한 한계점이라 할 수 있다. 본 분석에서는 아주 단순한 자연어 처리 방법 중 하나인 Bag of Words를 생성하는 방법을 통해 행렬을 생성하고 이를 Random Forest 모형을 통해 단순 적합시키는 방법으로 자연어 처리를 수행하였다.

이는 단순히 단어의 등장 여부와 빈도만을 가지고 분류하여도 큰 정확도를 가지는 Task에 대해서는 문제 없는 접근이나. 본 분석의 종속변수가 국가/품목으로 세분화된 그룹의 수출액 변동률이라는 것을 감안할 때. 단순 빈도 차원의 접근보다 더욱 고도화된 자연어 처리가 필요했을 것으로 예상된다.

이러한 부분은 분석 과정에서 챙기지 못한 명확한 한계점이지만, 그럼에도 불구하고 특정 그룹에서 자기회귀 모형에 비해 약간 정확도가 개선된 케이스가 존재하였다.

따라서 다음 단계에서는 뉴스 데이터의 기간과 폭을 다양화하고, 기사의 필터링 방식을 고도화하며, 더욱 높은 수준의 자연어 모델을 도입하여 유사한 방식의 분석을 수행한다면, 뉴스 기사가 실제 특정 국가-품목 그룹의 수출액 변화를 예측하고 나아가 유망한 그룹을 찾아내는 데 유의미한 성능을 가지는지 좀 더 면밀히 파악할 수 있을 것으로 기대된다.

4 활용데이터

ㅇ 국제 무역
관세청, ECOS, WITS에서 csv 파일 형태로 데이터를 수집하였다.

변수명	출처	수집처	빈도	데이터 제공 기간	메모/비고
대륙별 수출입실적_2000~2023	수출입무역통계	관 세청	년	2000~2023	기간,대륙,수출건수,수출금액,수입건수,수입금액,무역수지
국제 주요국 수출	국가별 수출액	ECOS	월/분기	1998~2022	통관기준
국제 주요국 수입	국가별 수입액	ECOS	월/분기	1998~2022	통관기준
WITS-Import-Export-all	global import expoert	WITS	년	2000~2020	전세계 수출액, 수입액

무역 관련 데이터는 국내외 수출, 수입의 추세를 확인하기 위해 수집하여 사용하였다.

ㅇ 대한민국 무역
품목별 수출입실적(GW)와 품목별 국가별 수출입실적(GW) 데이터는 공공데이터포털의 Open API를 활용해서 1995년부터 2023년까지의 데이터를 수집했다. 나머지 데이터는 공공데이터포털, kotra 등에서 csv로 파일 형태로 수집하였다.

변수명	출처	수집처	빈도	데이터 제공 기간	메모/비고
표준품명	공공데이터포털	관세청			2022년도 표준품명 ,데이터로서 HS부호 단위별로 품명, 한글, 표준품명, 영문 표준품명, 규격번호, 한글 필수규격, 영문 필수규격, 규격값, 세부분류내용
품목번호별 관세율표	공공데이터포털	관세청			품목번호별로 관세율을 표시한 데이터로 품목번호, 관세율 구분, 단위당세액, 기준가격, 적용국가구분, 용도세율구분, 적용개시일, 적용만료일
품목별 수출입실적(GW)	공공데이터포털	관세청	월	1995~2023	AP를 활용해서 데이터 수집함 기간, 품목명, HS코드, 수입/수출 중량 및 금액, 무역수지
품목별 국가별 수출입실적(GW)	공공데이터포털	관세청	월	1995~2023	AP를 활용해서 데이터 수집함 기간, 국가명, 국가코드, 품목명, HS코드, 수출액, 수입액, 무역수지
수출지역순위	수출지역순위	Kotra			

해당 데이터는 종속 변수로 사용되었다.

ㅇ 뉴스 데이터
본 분석의 핵심이 되는 데이터 중 하나이며, 네이버 뉴스를 통해 수집하였다.
경제 섹션의 2018~2023.05월 기사를 수집하여 사용하였으며, 작업 환경의 한계로 2023년 4월과 2023년 5월은 전체 기사 중 절반을 수집하여 분석에 활용하였다.

석에는 기사 본문만을 사용하였으며, 기사 페이지에 명시된 업로드 시간을 기준으로 뉴스 기사를 월별로 그룹화하였다.

ㅇ 경제 변수
수출입 무역통계, 국가지표체계, ECOS, KOSIS 등에서 csv로 수집한 뒤 연도를 기준으로 병합하였다. 품목별 국가별 수출입 통계 데이터가 1995년부터 있었기 때문에 경제변수 데이터 역시 가능한 1995~2022년의 데이터를 수집하였다. 1995년 데이터가 없는 경우 수집 가능한 연도부터의 데이터를 수집하였다.

변수명	출처	수집처	빈도	데이터 제공 기간	메모/비고
수출금액	관세청	수출입무역통계	년	2000~2023	(단위: 1000달러)
수입금액	관세청	수출입무역통계	년	2000~2023	(단위: 1000달러)
무역수지	관세청	수출입무역통계	년	2000~2023	(단위: 1000달러)
경제성장률	한국은행	국가지표체계	년	1995~2022	실질, 전년대비, 기준년도 2015년
GDP 디플레이터	한국은행	ECOS	년	1995~2022	국내총생산 디플레이터 (단위: 2015=100)
상품수지	한국은행	KOSIS	년	1995~2022	(단위: 100만달러)
서비스수지	한국은행	KOSIS	년	1995~2022	(단위: 100만달러)
무역의존도 (수출)	관세청, 한국은행	KOSIS	년	1995~2021	수출입의 대 GDP 비율, 관세청 통관자료, 한국은행 (명목GDP)
무역의존도 (수입)	관세청, 한국은행	KOSIS	년	1995~2021	수출입의 대 GDP 비율, 관세청 통관자료, 한국은행 (명목GDP)
수출입 비율	한국은행	국가지표체계	년	1995~2022	GDP 대비 수출입 비율 ({(수출 총액 + 수입 총액 + 국외수취요소소득 + 국외지급요소소득) ÷ 명목GDP} × 100), 2015년 기준, 최근 연도는 잠정치
원/달러 환율	주가지수	ECOS	년	1995~2022	평균 자료
인플레이션 전망	OECD	KOSIS	년	1995~2023	유로 지역 국가, 유로 지역 합계, 영국의 소비자물가지수(CPI) 또는 조화소비자물가지수(HICP)로 측정 전망치는 개별 국가 및 세계 경제의 경제 환경 측정을 기반으로 하며, 모델 기반 분석 및 전문가 판단을 조합하여 산출
수출 물가지수	한국은행	ECOS	년	1995~2022	달러기준 (단위: 2015=100)
수입 물가지수	한국은행	ECOS	년	1995~2022	달러기준 (단위: 2015=100)
기준금리	한국은행	ECOS	년	1999~2022	말일 기준
주가지수	한 국 거 래 소	ECOS	년	1995~2022	말일 기준

해당 데이터는 아래의 파생변수 목록과 같이 가공하여 경제변수와 수출과의 상관관계를 파악하는 데 활용하였다.

o 파생변수

앞서 수집한 경제 변수를 아래와 같이 가공하여 분석 시 활용하였다.

파생변수명	계산 방식
GDP디플레이터 변화율	(금년도 GDP디플레이터 - 전년도 GDP디플레이터) / 전년도 GDP디플레이터
원/달러환율 변화율	(금년도 원/달러환율 - 전년도 원/달러환율) / 전년도 원/달러환율
상품수지 변화율	(금년도 상품수지 - 전년도 상품수지) / 전년도 상품수지
서비스수지 변화율	(금년도 서비스수지 - 전년도 서비스수지) / 전년도 서비스수지
수출물가지수 변화율	(금년도 수출물가지수 - 전년도 수출물가지수) / 전년도 수출물가지수
주가지수 변화율	(금년도 주가지수 - 전년도 주가지수) / 전년도 주가지수
실업률 변화 (%p)	금년도 실업률 - 전년도 실업률
기준금리 변화 (%p)	금년도 기준금리 - 전년도 기준금리
인플레이션전망 변화 (%p)	금년도 인플레이션전망 - 전년도 인플레이션전망
수입무역의존도 변화 (%p)	금년도 수입무역의존도 - 전년도 수입무역의존도

5 사업화방안 및 기대효과

본 분석에서 그 효과를 파악하고자 한 추천 시스템은 기존의 여러 분석들과 달리 세분화된 국가-품목 그룹에 대해서 유의미한 수출 예측을 만들어내는 방법으로 뉴스 데이터와 자기회귀 모형을 활용하였다.

분석 결과 설정한 가설을 온전히 채택하는 데에는 무리가 있었으나, 그룹별로 뉴스 데이터를 기반으로 추론하는 데 있어 특징이 다르게 나타난다는 것을 예상해볼 수 있었다.

이를 기반으로 향후 뉴스 데이터를 다양화하고, 딥러닝 기반 자연어 모델을 도입하고, 기사 필터링 방식을 개선하는 등의 모델 고도화 작업을 수행할 경우 뉴스 기반의 세부 그룹별 수출 전망 파악의 정확도를 높일 수 있을 것으로 기대된다.

뉴스 텍스트를 기반으로 수출에 대한 전망을 사전에 추론할 수 있게 된다면, 한국의 경제에 가장 중요한 요소라고도 할 수 있는 무역에 대한 예측력이 높아진다. 이는 수출 측면에서의 충격에 대해 사전에 대비하거나 미리 파악할 수 있게 된다는 것을 의미하므로, 경제 정책을 수립하고 시행하는 주체나 실제 해당 파트에서 경제활동을 수행하는 개인들에게 있어 큰 의사결정에 큰 정보적 이득이 될 수 있을 것으로 기대된다.

다만 이와 같이 실질적인 효과를 얻어내기 위해서는 현 분석 수준에서 제안한 모델을 고도화하여 더욱 정확도를 높이는 과정을 반드시 수반해야 하므로, 보다 면밀한 분석과 검증 과정이 필요할 것이다.

※ 분량제한은 없으며, 공모요강에 적시된 평가항목을 참고하여 작성하여 주시기 바람 (상세 설명을 위해 도표, 스케치 등 별도파일 추가 가능)