핵심 광물별 공급위기 요소

탐지 모델 개발

(핵심광물별 가격 변동성, 시세, 생산·수입국 현황 등으로 핵심 광물별로
공급리스크를 줄일 수 있는 요소 탐지 분석 모델 개발)

작성자 김재환 , 한국외국어대학 경영학부 MBA
성지연 , Bloomberg City Initiatives
박기복 , 삼성전자로지텍

이메일 김재환 : kjh861213@gmail.com
성지연 : gigisung.53@gmail.com
박기복 : kibokpark9268@naver.com

목 차

I. 연복환 모델: 핵심광물의 가격 변동성 예측을 통한 공급리스크 관리

산업통상자원부의 공공데이터 활용 아이디어 공모전에 참여하게 된 한국외국어대학 경영학부 MBA 소속의 김재환, Bloomberg City Initiatives 소속의 성지연, 그리고 삼성전자로지텍 소속의 박기복은 각자의 전문성을 바탕으로 한 달 간의 치열한 논의와 협력을 통해 '연복환 모델'을 개발하게 되었습니다.

우리가 직면한 문제는 핵심광물의 가격 변동성, 시세, 생산 및 수입국 현황 등 다양한 요소로 인해 발생하는 공급리스크를 어떻게 효과적으로 줄일 수 있을지에 대한 것이었습니다. 이를 해결하기 위해 우리는 시계열 모델, Prophet 모형, 안정화 지수 예측 및 분류모형, 위험성 평가모델, 다변량 위험예측 모델, 이상치 탐지 모형, 앙상블 모형 등 여러 모델을 고안하고 분석하였습니다.

모델의 명칭은 우리 세 사람의 이름에서 따왔습니다. 성지연의 '연', 박기복의 '복', 김재환의 '환'을 합쳐 '연복환 모델'로 명명하였습니다. 이 명칭은 단순히 이름을 조합한 것 이상의 의미를 담고 있습니다. '연'은 연속성과 지속성을, '복'은 복잡한 문제를 해결하는 능력을, '환'은 새로운 혁신적인 변화를 의미합니다. 이는 우리가 개발한 모델이 지속적인 데이터 분석을 통해 복잡한 문제를 해결하고, 새로운 변화를 이끌어 낼 수 있음을 상징합니다.

연복환 모델은 다음과 같은 주요 요소들을 포함하고 있습니다:

1. **안정화 지수 예측 및 분류모형**: 각 광물의 안정화를 위한 지수를 계산하고, 이를 기반으로 안정화 전략을 제시합니다.
2. **위험성 평가모델**: 장 · 단기적 위험을 예측하고 그에 따른 위험 평가를 실시합니다.
3. **다변량 위험예측 모델**: 다양한 변수들을 동시에 고려하여 종합적인 위험 예측을 수행합니다.
4. **이상치 탐지 모형**: 비정상적인 데이터 변동을 신속하게 탐지하여 사전 대처할 수 있도록 합니다.
5. **앙상블 모델**: 다양한 모형들의 결합을 통하여 기존 모델의 성능향상과 혁신을 통한 문제 해결을 가능하게 합니다.

연복환 모델은 단순히 기술적 접근에 그치지 않고, 각 광물의 생산국과 수입국의 정치적, 경제적 상황까지 고려하여 종합적인 리스크 분석을 수행하는 것을 목표로 합니다. 이를 통해 핵심광물의 공급망에서 발생할 수 있는 다양한 리스크를 사전에 예측하고, 완화할 수 있는 방안을 제시하는 것을 목표로 합니다.

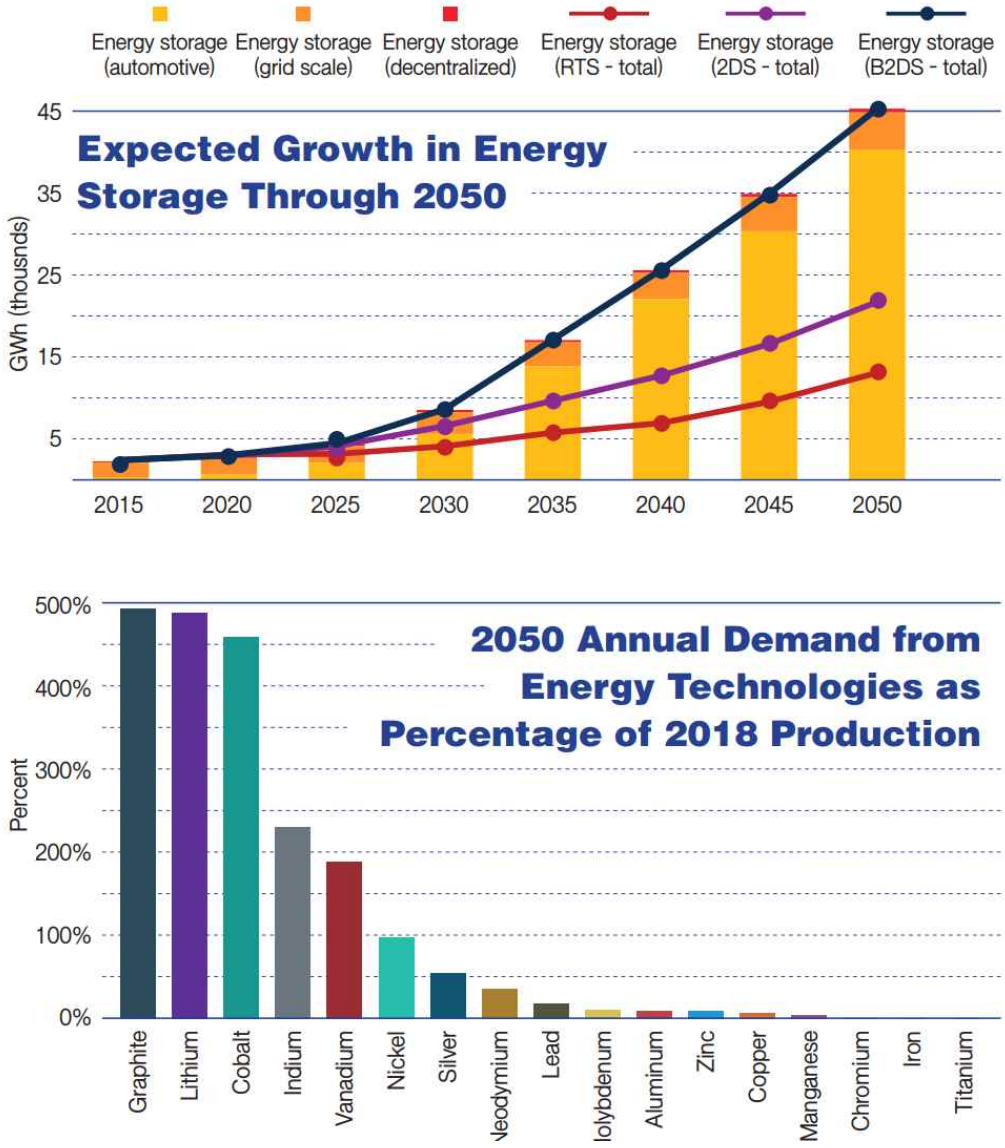
우리의 목표는 단순히 데이터를 분석하는 것에 그치지 않고, 이를 통해 실질적인 공급망 안정화 방안을 제공하는 것입니다. 연복환 모델은 데이터 기반의 과학적인 접근과 실무 경험을 결합하여, 핵심 광물의 공급 리스크를 효과적으로 줄일 수 있는 혁신적인 솔루션을 제공할 것입니다..

Ⅱ. 제안배경

1. 현황 및 제안목적

(1) 탄소중립 등 환경에 대한 관심과, 전기차 배터리 등 유망산업의 핵심광물 수요 증가

2050년까지 배터리(에너지) 시장은 크게 성장할 것으로 전망되며, 리튬, 니켈, 코발트, 망간 같은 핵심 금속들은 전기차 이차전지의 주요 원료로, 리튬은 배터리 양극재, 니켈과 코발트는 배터리의 에너지 밀도를 높이는데 사용됩니다. 세계은행의 예측치에 따르면, 리튬이나 코발트는 2018년 대비 2050년에는 450%이상으로 수요가 크게 증가할 것으로 예상됩니다.



세계 에너지(Energy Storage) 시장의 성장 예측치(상단)와 에너지 분야의 광물 수요증가 예측치(하단) [1)]

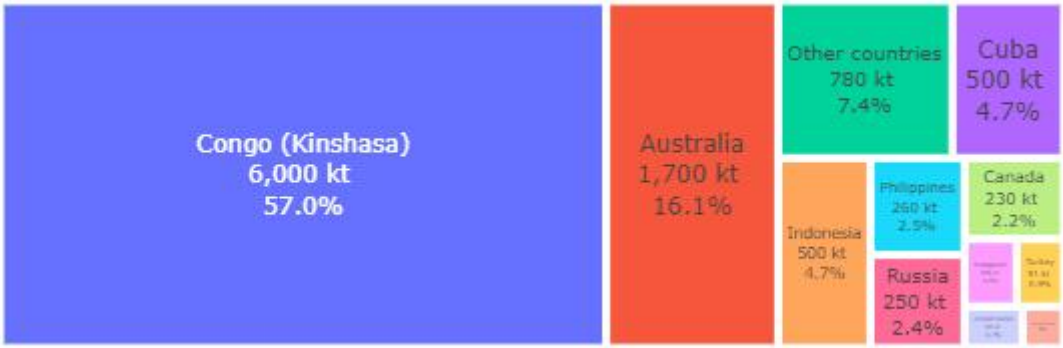
1) 출처 : The World Bank

(2) 핵심광물 매장이 특정 국가에 심하게 편중되어 있어 상당한 공급망 리스크 존재

USGS의 자료를 토대로 광물별 매장량을 산출했을 때, 분석대상 4종 광물 모두 매장량의 50%이상이 1~2개 국가에 편중되어있어, 공급망 리스크가 큰 것을 확인했습니다.

- 니켈 : 인도네시아(42%), 호주(18.3%), 브라질(12.2%) 등

- 코발트: 콩고(57%), 호주(16.1%) 등



- 리튬: 칠레(33.6%), 호주(22.4%), 아르헨티나(13%) 등

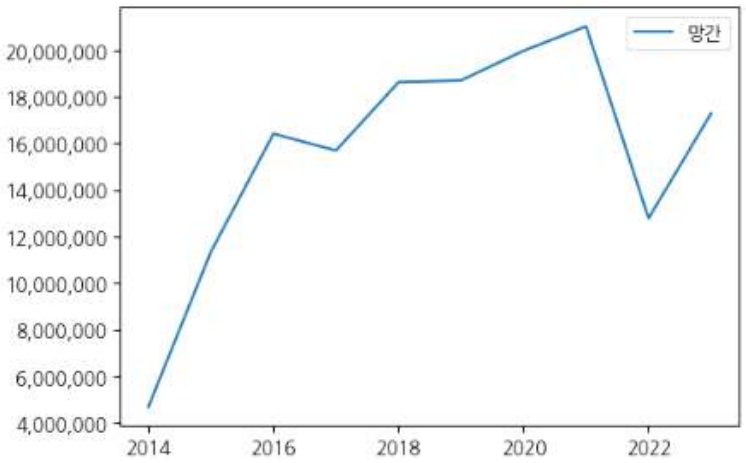
- 망간: 남아프리카(31.4%), 호주(26.2%), 중국(14.7%) 등[2)]

2) 원본 데이터(USGS Mineral Commodity Summaries 2024)에서 가공

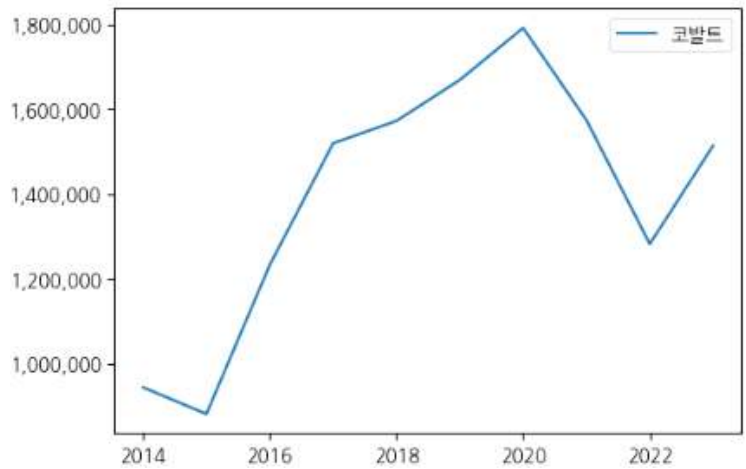
(3) 핵심광물 조달에 대한 수입경쟁 심화

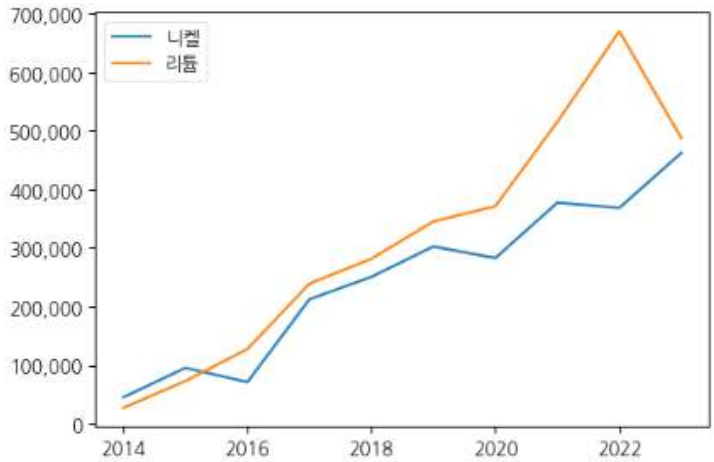
- 핵심광물의 전세계 이동(광물 수입량) 증가

UN Comtrade의 무역통계를 기준으로 각 광물에 대해 조사했을 때, 리튬을 제외한 각 광물의 국가별 수입량은 2014년 대비 증가세를 보였습니다. 2022년 이후 리튬수입 감소세는, 전기차 캐즘 현상(대중화 직전 수요 감소) 및 배터리시장 경쟁심화(한국업체 점유율 전년비 5.3% 감소)가 원인일 것으로 추정됩니다.



3)





3) 원본 데이터(UN Comtrade DB)에서 가공

- **핵심광물 확보를 위한 경쟁국가들의 전략**

다양한 전략을 통해 핵심 광물 확보를 목표로 하는 경쟁국들의 노력이 진행 중으로, 주로 자원의 확보와 비축 등에 대한 부분이 주를 이루고 있습니다. 국가별 확보전략에 대해 간략히 요약했습니다.

* 한국 : 희소금속 산업 발전대책 2.0 추진 (확보-비축-순환 수급에 대한 3중 안정망 등)
* 미국 : 중요 광물 공급을 위한 연방정부 전략 추진 (수입의존 저감 및 공급망 확보 등)
* 일본 : 신 국제자원 전략 추진 (비축제도 재검토 및 확보책 구축, 국제협력 도출 등)
* EU : EIP Raw materials 추진 (자원의 채광/재활용에 대한 시험적 대응 등)

- **광물별 상위 수입국 현황**[4)]

[니켈]

	2015	2016	2017	2018	2019	2020	2021	2022	2023
1위	Brazil 34,127	Brazil 19,720	Japan 112,033	Japan 135,167	Japan 191,793	Japan 200,127	Japan 206,021	Japan 173,627	China 256,142
2위	Rep. of Korea 27,394	China 15,197	Rep. of Korea 53,085	Rep. of Korea 52,446	Belgium 34,111	Belgium 31,329	China 98,853	China 119,071	Japan 124,779
3위	Canada 16,517	Germany 14,054	China 21,639	Australia 21,776	Rep. of Korea 31,817	Canada 21,353	Belgium 41,121	Belgium 43,286	Belgium 41,815
4위	Germany 12,654	Canada 13,500	Germany 13,939	Belgium 20,963	Canada 25,360	China 16,340	Canada 21,410	Canada 18,103	Malaysia 26,280
5위	Thailand 4,923	Rep. of Korea 8,951	India 11,978	China 20,569	Australia 19,640	Australia 13,834	Malaysia 10,150	Rep. of Korea 14,469	Canada 13,364

[코발트]

	2015	2016	2017	2018	2019	2020	2021	2022	2023
1위	Germany 366,375	USA 444,837	USA 496,572	USA 471,603	United Kingdom 546,036	Germany 571,008	Germany 660,773	Germany 488,970	Germany 529,288
2위	Sweden 220,380	Germany 347,106	United Kingdom 345,329	Germany 329,296	Germany 380,888	USA 472,193	USA 295,225	Italy 250,075	United Kingdom 395,371
3위	Brazil 124,744	Sweden 201,729	Germany 299,495	Finland 277,965	USA 348,198	United Kingdom 371,668	Italy 232,667	Brazil 191,783	USA 219,089
4위	Canada 93,406	Brazil 137,948	Italy 195,547	United Kingdom 264,720	Italy 217,586	Italy 217,804	United Kingdom 198,261	Austria 186,431	Japan 189,547
5위	Rep. of Korea 76,266	Netherlands 101,162	Sweden 182,634	Italy 228,860	France 176,344	Sweden 159,068	Sweden 187,343	Sweden 164,849	Sweden 180,823

4) 원본 데이터(UN Comtrade DB)에서 가공 (단위 : kt[킬로톤])

[리튬]

	2015	2016	2017	2018	2019	2020	2021	2022	2023
1위	Rep. of Korea 35,014	China 47,227	Japan 69,609	Rep. of Korea 91,029	Rep. of Korea 124,094	Rep. of Korea 130,051	Rep. of Korea 189,821	China 278,358	China 324,683
2위	Belgium 18,782	USA 33,806	China 61,411	Japan 86,273	Japan 118,216	China 101,259	China 169,226	Rep. of Korea 238,878	Japan 109,344
3위	Canada 7,502	Belgium 24,300	Rep. of Korea 55,620	China 46,158	China 59,437	Japan 97,427	Japan 108,999	Japan 120,350	USA 34,261
4위	Germany 6,237	Russian Federation 11,620	USA 34,820	USA 36,507	USA 26,740	USA 26,229	USA 27,766	Netherlands 17,332	Netherlands 12,724
5위	Spain 5,638	Canada 10,842	Belgium 17,840	Belgium 21,817	France 16,984	Belgium 16,554	Russian Federation 19,591	USA 15,192	United Kingdom 6,761

[망간]

	2015	2016	2017	2018	2019	2020	2021	2022	2023
1위	Hong Kong 6,596,207	Hong Kong 6,430,363	Hong Kong 4,445,408	USA 5,646,705	USA 5,337,518	USA 5,522,054	USA 6,726,526	Germany 4,296,848	USA 7,890,263
2위	Germany 2,047,818	USA 3,899,931	USA 4,356,834	Hong Kong 5,317,808	Hong Kong 5,014,159	Hong Kong 4,943,913	Germany 4,675,819	Japan 2,842,379	Germany 3,032,223
3위	Belgium 1,255,255	Germany 3,361,475	Germany 2,858,660	Germany 2,722,953	Germany 3,163,188	Germany 4,058,559	Hong Kong 4,648,238	Hong Kong 2,556,882	Japan 2,457,672
4위	Brazil 1,015,417	Russian Federation 1,509,007	Japan 2,302,316	China 2,504,820	China 2,789,753	China 2,812,012	Tunisia 2,484,856	Poland 1,565,781	Hong Kong 2,360,582
5위	Rep. of Korea 443,342	China 1,213,518	Russian Federation 1,732,776	Japan 2,433,654	Japan 2,398,831	Japan 2,636,763	Japan 2,484,354	China 1,529,201	Poland 1,547,787

(4) 핵심요약 및 제안목적

한국은 핵심 광물의 수요 증가와 공급망 위기, 수급 경쟁 심화 등의 상황 속에서도 핵심 광물의 비축 및 국산화 노력을 강화하고 있습니다. 그러나 생산지 편중과 경쟁국의 확보 전략은 통제하기 어려운 요소로 여전히 공급망 리스크를 초래하고 있습니다.

이러한 상황에서는 미래 핵심 산업의 경쟁력을 유지하기 위해 핵심 원재료(광물)의 안정적인 확보가 필수적입니다. 이를 위해 각 상황에 대한 예측과 위기 요소를 탐지할 수 있는 위기 요소 탐지 모델을 도입하고자 합니다.

2. 도입 방법

(1) 분석대상 정립

배터리 분야의 핵심광물인 니켈, 코발트, 리튬, 망간을 분석대상으로 지정했습니다. 핵심광물의 국제현황을 파악하고자 각 광물의 기준이 될 HSCODE를 정했습니다. 한국뿐 아니라 전세계의 현황을 파악할 목적이므로, 10자리의 HSK(HS of Korea)체계가 아닌, 국제 공통으로 사용되는 6자리의 HS체계를 기준으로 확정했습니다.

* 니켈 : 산화/수산화니켈(282540), 황산니켈(283324)
* 코발트 : 산화코발트(282200), 황산코발트(283329)
* 리튬 : 산화/수산화리튬(282520), 탄산리튬(283691)
* 망간 : 이산화망간(850610)

(2) 고려요소(Feature) 선정

- 가격추이 및 경쟁국 활동의 모니터링

UN Comtrade DB API 핵심광물의 수출입 현황 데이터를 산출할 수 있고, 이를 통해 가격추이(수입가격)와 경쟁자의 활동(수입량 증감)을 파악하고자 했습니다.

가격은 수입량 증감에 따른 영향을 배제하고자 중량당 가격(Price per Weight)로 계산하여 고려했습니다. UN Data의 Value(가격)와 Netweight(중량)의 연월별 데이터로 분석합니다.

전세계의 수입량을 고려대상으로 포함하여, 다양한 핵심광물 확보전략을 추진중인 경쟁국의 활동을 모니터링 하고자 하였습니다.

- 공급망 리스크 모니터링을 위한 물류지표 추가

핵심광물은 생산지가 해외에 편중되어 국제 운송이 필수로 수반됩니다. 단순히 가격만이 아닌 물류(해상운송)에 대해 고려하여 공급망에 대해 총체적으로 고려하고자, BDI(Baltic Dry Index)지수를 대상으로 포함했습니다.

광물 등 원자재는 일반 컨테이너선이 아닌 건화물선으로 운송되므로, 상하이컨테이너 운임지수가 아닌 건화물선 운임지수인 BDI를 포함했습니다. BDI는 건화물선 운송 자체에 대한 지표이기도 하지만, 구체적인 운송이 있을때만 예약/활용된다는 특성과 원자재 운송에 활용된다는 점에서 미래의 경제활동(생산)에 대해 미리 알 수 있는 경기선행지표로도 활용되는 지표입니다.

3. 예측 및 이상탐지

우리는 자원 수입 현황을 분석하기 위해 ARIMA와 홀트-윈터스 지수평활법을 처음으로 채택하여 예측 모델을 구축하고자 합니다. 이후에는 GARCH 모형을 도입하여 변동성을 상세히 분석하고, LSTM과 같은 AI 머신러닝 모델을 활용하여 더 정밀한 예측을 시도하고자 합니다.

이 예측 데이터를 기반으로, R이나 파이썬의 Prophet과 같은 Anomaly Detection 기법을 활용하여 이상 탐지를 수행할 계획입니다. 이를 통해 예기치 않은 자원 수입의 변동이나 국제 시장에서의 변화를 신속하게 감지하고 대응할 수 있는 시스템을 구축하고자 합니다

Ⅲ. 모형의 이론적 배경과 선정 이유

1. 주요 방법론

(1) GARCH 모형

1) 이론적 배경

GARCH(Generalized Autoregressive Conditional Heteroskedasticity) 모형은 시계열 데이터의 변동성을 모델링하는 데 널리 사용되는 방법입니다. 이 모형은 Engle(1982)이 개발한 ARCH(Autoregressive Conditional Heteroskedasticity) 모형을 확장한 것입니다. Bollerslev(1986)이 제안한 GARCH 모형은 ARCH 모형의 단점을 보완하여 더 일반적인 형태로 변동성을 설명합니다.

2) 핵심 개념

- 조건부 분산: GARCH 모형은 시간에 따라 변화하는 조건부 분산을 통해 데이터의 변동성을 모델링합니다.
- 이분산성(heteroskedasticity): 시계열 데이터에서 관측되는 분산의 변화(변동성)를 설명하는 데 중점을 둡니다.
- 변동성 클러스터링(volatility clustering): 자원 수입 데이터와 같은 금융 시계열에서 자주 나타나는 현상으로, 고변동성과 저변동성 구간이 군집을 이루는 패턴을 포착합니다.

3) GARCH(p, q) 모형의 수식

GARCH 모형은 다음과 같은 형태로 나타낼 수 있습니다:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

- σ_t^2는 시점 t에서의 조건부 분산(변동성)입니다.
- ϵ_t는 시점 t에서의 잔차입니다.
- $\alpha_0, \alpha_i, \beta_j$는 모형의 계수입니다.
- p는 ARCH 항의 개수, q는 GARCH 항의 개수입니다.

4) 선정이유

a. 변동성 클러스터링 포착

자원 수입 데이터는 종종 변동성 클러스터링을 보입니다. GARCH 모형은 이러한 패턴을 효과적으로 포착하여 변동성을 예측하는 데 강점을 보입니다.

b. 조건부 이분산 모델링

GARCH 모형은 시간에 따라 변화하는 분산을 모델링하는 데 적합합니다. 이는 자원 수입 데이터에서 발생하는 변동성의 동적 특성을 잘 반영합니다.

c. 리스크 관리 및 의사결정 지원

변동성 예측을 통해 자원 수입의 리스크를 관리하고, 전략적 의사결정에 중요한 정보를 제공합니다. 예를 들어, 높은 변동성이 예상되는 시기에 대비한 비축 전략 수립이 가능해집니다.

d. 금융 및 경제 데이터에서의 검증된 성과

GARCH 모형은 금융 및 경제 데이터 분석에서 그 성능이 입증되었습니다. 자원 수입 데이터 역시 경제적 요인에 크게 영향을 받으므로, GARCH 모형의 활용이 적절합니다.

e. 예측력 강화

GARCH 모형은 다른 시계열 모델에 비해 변동성 예측에서 높은 정확도를 보입니다. 이는 자원 수입의 불확실성을 줄이고, 안정적인 공급망 관리를 돕습니다.

(2) LSTM 모형

1) 이론적 배경

LSTM(Long Short-Term Memory) 모형은 Recurrent Neural Network(RNN)의 일종으로, 시계열 데이터 분석 및 예측에 널리 사용됩니다. LSTM은 Hochreiter와 Schmidhuber가 1997년에 처음 제안한 모델로, RNN의 장기 의존성 문제를 해결하기 위해 개발되었습니다.

2) 핵심 개념

- Recurrent Neural Network(RNN): 순환 신경망(RNN)은 이전 시간 단계의 출력을 현재 단계의 입력으로 사용하는 구조를 가지고 있습니다. 이는 시계열 데이터와 같이 시간의 흐름에 따라 변화하는 데이터를 처리하는 데 유리합니다.
- 장기 의존성 문제: RNN은 시간 단계가 길어질수록 이전 정보가 사라지거나 왜곡되는 '기울기 소실(Vanishing Gradient)' 문제를 겪습니다.
- LSTM 구조: LSTM은 이러한 문제를 해결하기 위해 고안된 특수한 구조를 가지고 있습니다. LSTM 셀은 입력 게이트, 출력 게이트, 망각 게이트라는 세 가지 게이트를 통해 정보의 흐름을 조절합니다.

3) LSTM 셀의 구성 요소

입력 게이트(Input Gate): 현재 입력과 이전 상태를 사용하여 새로운 정보를 얼마나 받아들일지 결정합니다.

망각 게이트(Forget Gate): 이전 상태의 정보를 얼마나 잊을지 결정합니다.

출력 게이트(Output Gate): 현재 상태를 출력으로 얼마나 보낼지 결정합니다.

이 게이트들은 아래와 같은 수식으로 정의됩니다:

$$f_t = \sigma(W_f \bullet [h_{t-x}, x_t] + b_f)$$

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i)$$

$$\widetilde{C_t} = \tanh(W_c \bullet [h_{t-1}, x_1] + b_c)$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t}$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times \tanh(C_t)$$

- f_t : 망각 게이트의 활성화 값
- i_t : 입력 게이트의 활성화 값
- $\widetilde{C_t}$: 새로운 기억 셀 상태
- C_t : 현재 기억 셀 상태
- o_t : 출력 게이트의 활성화 값
- h_t : 현재 숨겨진 상태
- tanh : 하이퍼볼릭 탄젠트 활성화 함수
- W와 b : 가중치와 바이어스 매개 변수

4) 선정이유

a. 장기 의존성 문제 해결

LSTM 모형은 RNN의 장기 의존성 문제를 해결하여, 긴 시간 단계의 시계열 데이터에서도 이전 정보의 중요성을 유지합니다. 이는 자원 수입 데이터의 장기적인 패턴을 예측하는 데 매우 유용합니다.

b. 복잡한 패턴 학습

LSTM은 복잡한 시계열 데이터의 패턴을 효과적으로 학습할 수 있습니다. 자원 수입

데이터는 다양한 외부 요인에 의해 영향을 받으며, 이러한 복잡한 상호작용을 학습하는 데 LSTM이 적합합니다.

c. 비선형 관계 모델링

자원 수입 데이터는 종종 비선형적인 특성을 가집니다. LSTM은 비선형 관계를 모델링하는 데 강점을 보이며, 이는 자원 수입 예측의 정확도를 높이는 데 기여합니다.

d. 예측력 강화

LSTM은 다른 전통적인 시계열 예측 모델에 비해 높은 예측력을 제공합니다. 이는 자원 수입의 변동성을 정확하게 예측하고, 안정적인 공급망 관리에 도움을 줍니다.

e. 광범위한 적용 사례

LSTM은 금융, 기후, 물류 등 다양한 분야에서 시계열 예측의 성능을 입증하였습니다. 자원 수입 데이터 역시 이러한 시계열 데이터의 특성을 가지므로, LSTM의 적용이 적절합니다.

f. 실시간 데이터 처리

LSTM은 실시간 데이터 스트리밍 환경에서도 효과적으로 작동합니다. 이는 자원 수입의 변동을 실시간으로 모니터링하고 신속하게 대응하는 데 유리합니다.

이와 같은 이유로 LSTM 모형을 자원 수입 현황 분석 및 예측의 주요 방법론 중 하나로 선정하였습니다. LSTM의 강력한 시계열 데이터 처리 능력을 활용하여 자원 수입의 변동성을 보다 정확하게 예측하고, 리스크를 효과적으로 관리할 수 있을 것으로 기대됩니다.

2. 보조적 방법론

(1) ARIMA 모형

1) 이론적 배경

ARIMA(Autoregressive Integrated Moving Average) 모형은 시계열 데이터를 분석하고 예측하는 데 널리 사용되는 통계적 방법입니다. ARIMA 모형은 Box와 Jenkins가 1970년에 제안한 모형으로, 자기회귀(AR)와 이동평균(MA) 성분을 통합한 형태입니다.

2) 핵심 개념

a. 자기회귀(AR) 모형

- 정의: 현재 시점의 값이 이전 시점들의 값에 의존하는 형태의 모형
- 수식: $X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t$
- 의미: Φ는 자기회귀 계수, ϵ_t는 백색 잡음.

b. 이동평균(MA) 모형

• 정의: 현재 시점의 값이 이전 시점들의 오차에 의존하는 형태의 모형

• 수식: $X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \Phi_t X_{t-p} + \epsilon_t$

• 의미: θ는 자기회귀 계수, ϵ_t는 백색 잡음.

c. 차분(differencing)

• 정의: 비정상성을 제거하기 위해 데이터에서 차분을 취하는 과정.

• 수식: $Y_t = X_t - X_{t-1}$

• 의미: 한 시점에서 이전 시점의 값을 빼서 새로운 시계열을 만듦

3) ARIMA(p, d, q) 모형의 수식

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots \theta_q\epsilon_{t-q}$$

• p: 자기회귀(AR) 항의 개수

• d: 차분 횟수

• q: 이동평균(MA)항의 개수

• ϕ: 이동평균 계수

• ϵ: 백색 잡음

4) 선정이유

a. 데이터의 자기상관성 분석

ARIMA 모형은 시계열 데이터의 자기상관성을 기반으로 예측을 수행합니다. 자원 수입 데이터는 과거 데이터에 의존하는 경향이 있기 때문에, ARIMA 모형을 통해 이러한 자기상관성을 효과적으로 분석할 수 있습니다.

b. 비정상성 제거

자원 수입 데이터는 종종 비정상적 특성을 가질 수 있습니다. ARIMA 모형은 차분 과정을 통해 비정상성을 제거하여 데이터를 정상화한 후 예측을 수행할 수 있습니다. 이는 데이터의 트렌드와 계절성을 더 명확하게 파악하는 데 도움이 됩니다.

c. 단기 예측의 정확성

ARIMA 모형은 단기 예측에서 높은 정확성을 보입니다. 자원 수입의 변동성을 단기적으로 예측하는 데 유용하며, 이를 통해 적절한 비축 전략을 수립할 수 있습니다.

d. 계절성 분석

ARIMA 모형의 변형인 SARIMA(Seasonal ARIMA)를 통해 계절적 변동성을 포함한 예측을 수행할 수 있습니다. 자원 수입 데이터는 계절적 요인에 의해 영향을 받을 수 있으므로, 이를 반영한 예측이 중요합니다.

e. 경제적 해석 용이

ARIMA 모형은 경제 데이터 분석에서 널리 사용되며, 예측 결과를 경제적으로 해석하는 데 용이합니다. 자원 수입 데이터의 변동성을 경제적 관점에서 분석하고 예측할 수 있습니다.

f. 구현과 사용의 용이성

ARIMA 모형은 다양한 통계 소프트웨어와 프로그래밍 언어에서 쉽게 구현할 수 있습니다. 이는 모델 구축과 결과 해석을 빠르고 효율적으로 수행할 수 있게 합니다.

(2) 홀트-윈터스 지수평활법

1) 이론적 배경

홀트-윈터스(Holt-Winters) 지수평활법은 시계열 데이터의 추세와 계절성을 고려한 예측 방법입니다. 이 방법은 단순 지수평활법에서 발전하여, 데이터의 트렌드와 계절적 변동을 동시에 반영할 수 있도록 고안되었습니다. Holt(1957)와 Winters(1960)가 각각 개발한 방법을 결합하여 사용되며, 가법적(추가형) 및 승법적(곱셈형) 두 가지 형태로 나뉩니다.

2) 핵심 개념

a. 수준(Level)

- 데이터의 현재 상태를 나타내는 값입니다.
- 수식: $L_t = \alpha(X_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1})$

b. 추세(Trend)

- 데이터의 장기적인 증가 또는 감소를 나타냅니다.
- 수식: $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

c. 계절성(Seasonality)

- 일정한 주기로 반복되는 패턴을 나타냅니다.
- 수식(가법적): $S_t = \gamma(X_t - L_{t-1} - T_{t-1}) + (1-\gamma)S_{t-m}$

• 수식(승법적): $S_t = \gamma(\frac{x_t}{L_{t-1} + T_{t-1}}) + (1-\gamma)S_{t-m}$

d. 예측(Forecast)

• 미래 시점의 값을 예측합니다.

• 수식(가법적): $\hat{X}_{t+h} = L_t + hT_1 + S_{t+h-m(k+1)}$

• 수식(승법적): $\hat{X}_{t+h} = (L_t + hT_1) \cdot S_{t+h-m(k+1)}$

• L_t: 시점 t에서의 수준(Level)

• T_1: 시점 t에서의 추세(Trend)

• S_t: 시점 t에서의 계절성(Seasonality)

• α, β, γ: 각각 수준, 추세, 계절성의 평활 계수

• X_t: 시점 t에서의 실제 값

• $\hat{X}_{t+h}$: 시점 t+h에서의 예측 값

• m: 계절 주기(예: 월별 데이터에서 12개월)

• h: 예측하고자 하는 미래 시점의 길이

• k: 계절 주기 횟수

3) 선정 이유

a. 계절적 변동성 반영

자원 수입 데이터는 계절적 요인에 의해 영향을 받습니다. 홀트-윈터스 지수평활법은 이러한 계절적 패턴을 효과적으로 반영하여 예측할 수 있습니다. 이는 자원 수입의 계절적 변동성을 예측하는 데 매우 유용합니다.

b. 단기 및 중기 예측의 적합성

홀트-윈터스 지수평활법은 단기 및 중기 예측에 적합합니다. 자원 수입의 단기적인 변화와 중기적인 패턴을 예측하여, 적절한 비축 및 공급 전략을 수립할 수 있습니다.

c. 추세 및 수준 변화 반영

이 방법은 데이터의 추세와 수준 변화를 모두 반영할 수 있습니다. 이는 자원 수입 데이터의 장기적인 증가 또는 감소 추세를 예측하고, 현재 수준을 정확하게 평가하는 데 도움을 줍니다.

d. 실시간 적용 가능성

홀트-윈터스 지수평활법은 계산이 비교적 간단하고, 실시간 데이터에 쉽게 적용할 수 있습니다. 이는 자원 수입의 변동을 실시간으로 모니터링하고 신속하게 대응하는 데 유리합니다.

e. 다양한 산업에서 검증된 성능

홀트-윈터스 지수평활법은 다양한 산업에서 그 성능이 검증되었습니다. 이는 자원 수입 데이터와 같은 시계열 데이터를 분석하는 데 적합하며, 예측의 신뢰성을 높이는 데 기여합니다.

f. 모델의 유연성

이 방법은 가법적(추가형) 및 승법적(곱셈형) 형태로 사용할 수 있어, 데이터의 특성에 맞게 모델을 유연하게 적용할 수 있습니다. 자원 수입 데이터의 특성에 따라 적절한 형태를 선택하여 예측의 정확도를 높일 수 있습니다.

(3) Prophet 모형

1) 이론적 배경

Prophet 모형은 Facebook의 데이터 과학팀이 개발한 시계열 예측 도구입니다. Prophet은 비정상적이고 불규칙한 데이터, 특히 시즌 패턴과 휴일 효과를 반영하는 데 뛰어난 성능을 보입니다. 이 모형은 데이터의 패턴을 자동으로 감지하고, 사용자에게 직관적이고 해석 가능한 예측을 제공하는 것을 목표로 합니다.

2) 핵심 개념

a. 가법 모델(Additive Model)

Prophet 모형은 가법 모델을 기반으로 합니다. 이는 시간의 함수로 표현되는 트렌드, 계절성, 휴일 효과를 더하여 시계열 데이터를 설명합니다.

수식 : $y(t) = g(t) + s(t) + h(t) + \epsilon_t$

- $g(t)$: 트렌드 함수로 데이터의 장기적인 변화를 나타냅니다.
- $s(t)$: 계정성 함수로 연간 주기의 반복되는 패턴을 나타냅니다.
- $h(t)$: 휴일 효과를 나타내는 함수로, 특정 이벤트나 공휴일의 영향을 반영합니다.
- ϵ_t 오차 항으로, 예측할 수 없는 잡음을 나타냅니다.

b. 트렌드(Trend)

트렌드는 데이터의 장기적인 증가 또는 감소를 나타내며, Prophet에서는 두 가지 형태로 모델링할 수 있습니다:

- 선형 트렌드: $g(t) = k + mt$

• 로그 트렌드: $g(t) = k + m\log(t)$

Prophet은 데이터에 가장 적합한 트렌드를 자동으로 선택합니다.

c. 계절성(Seasonality)

계절성은 일정한 주기로 반복되는 패턴을 설명합니다. Prophet은 연간 계절성 외에도 주간 및 일간 계절성을 모델링할 수 있습니다. Fourier 시리즈를 사용하여 계절성을 모델링 할 수 있습니다.

• 수식: $s(t) = \sum_{n=1}^{N}(a_n\cos(\frac{2\pi nt}{P}) + b_n\sin(\frac{2\pi nt}{P}))$

d. 휴일 효과(Holiday Effects)

특정 이벤트나 공휴일이 데이터에 미치는 영향을 반영합니다. 사용자가 휴일 목록을 제공하면 Prophet이 이들 날짜를 특별하게 처리합니다.

e. 변화점(Changepoints)

트렌드 변화가 발생하는 지점을 자동으로 감지하고 반영합니다. Prophet은 변화점을 통해 트렌드가 변화하는 시점을 효과적으로 모델링합니다.

3) 선정 이유

a. 비정상적이고 불규칙한 데이터 처리 능력

Prophet 모형은 비정상적이고 불규칙한 데이터에서 뛰어난 성능을 보입니다. 자원 수입 데이터는 종종 비정상적이고 불규칙한 변동을 보이기 때문에, Prophet 모형을 통해 이러한 데이터를 효과적으로 처리할 수 있습니다.

b. 직관적이고 해석 가능한 예측

Prophet 모형은 사용자가 쉽게 이해할 수 있는 예측 결과를 제공합니다. 이는 자원 수입 데이터의 변동성을 직관적으로 해석하고, 이를 기반으로 전략적 의사결정을 내리는 데 유용합니다.

c. 계절성 및 휴일 효과 반영

자원 수입 데이터는 계절적 요인과 특정 이벤트의 영향을 받을 수 있습니다. Prophet 모형은 이러한 계절성 및 휴일 효과를 정확하게 반영하여 예측의 정확도를 높입니다.

d. 자동 변화점 감지

Prophet 모형은 데이터의 트렌드 변화점을 자동으로 감지하여 반영합니다. 이는 자원 수입 데이터에서 발생하는 중요한 변화를 놓치지 않고 예측에 반영할 수 있도록 합니다.

e. 빠른 계산 속도와 확장성

Prophet 모형은 대규모 데이터셋에서도 빠르게 계산할 수 있으며, 다양한 환경에 쉽게 적용할 수 있습니다. 이는 실시간 데이터 처리 및 예측에 유리합니다.

f. 사용자 친화적 인터페이스

Prophet은 Python과 R에서 쉽게 사용할 수 있으며, 직관적인 파라미터 설정과 시각화 도구를 제공합니다. 이는 분석가가 빠르게 모델을 구축하고 결과를 해석하는 데 도움을 줍니다.

3. 변동성 예측의 보조 도구

(1) Isolation Forest

1) 이론적 배경

Isolation Forest 모델은 데이터에서 이상치(outliers)를 탐지하는 데 사용되는 비지도 학습 알고리즘입니다. 이 모델은 Liu, Ting, and Zhou가 2008년에 제안한 방법으로, 데이터 포인트를 분리하는 데 필요한 횟수를 기반으로 이상치를 식별합니다. Isolation Forest는 기존의 거리 기반 및 밀도 기반 이상치 탐지 기법과 달리, 데이터의 분포나 밀도를 추정하지 않습니다.

2) 핵심 개념

a. 분리(Isolation)

Isolation Forest는 데이터 포인트를 분리하는 과정을 통해 이상치를 탐지합니다. 이상치는 일반적으로 정상치에 비해 쉽게 분리되기 때문에, 분리하는 데 필요한 횟수가 적습니다.
데이터 분리를 위해 랜덤한 서브셋(subset)을 선택하고, 각 서브셋에서 임의의 분할(potential split)을 수행합니다.

b. 분할(Splitting)

각 분할은 임의의 특징(feature)와 해당 특징의 임의의 값을 기준으로 이루어집니다. 이러한 과정을 반복하여 트리를 생성합니다.
생성된 트리는 데이터 포인트를 분리하는 데 필요한 단계 수(트리의 깊이)를 기록합니다.

c. 트리(Tree)

Isolation Forest는 여러 개의 랜덤한 분할 트리를 생성하여 포레스트(forest)를 만듭니다.
각 트리는 데이터 포인트를 분리하는 데 필요한 단계 수를 기반으로 이상치를 판단합니다.

d. 이상치 점수(Anomaly Score)

각 데이터 포인트에 대해 분리하는 데 필요한 평균 단계를 계산하여 이상치 점수를 부여합니다. 이 점수는 트리의 깊이와 반비례합니다.
이상치 점수가 높을수록 해당 데이터 포인트가 이상치일 가능성이 큽니다.

Isolation Forest 알고리즘의 핵심 아이디어는 이상치가 정상 데이터 포인트보다 적은 단계에서 분리될 것이라는 가정에 기반합니다.

3) 선정 이유

a. 효율적인 계산 성능

Isolation Forest는 데이터 포인트의 하위 샘플링과 트리 구조를 활용하여 이상치를 탐지하기 때문에 계산이 매우 효율적입니다. 이는 대규모 자원 수입 데이터셋에서도 빠르게 이상치를 탐지할 수 있게 합니다.

b. 비지도 학습 방식

이 알고리즘은 비지도 학습 방법을 사용하기 때문에, 이상치 탐지를 위해 레이블이 필요하지 않습니다. 이는 레이블링이 어려운 자원 수입 데이터에서도 효과적으로 이상치를 탐지할 수 있습니다.

c. 고차원 데이터 처리 능력

Isolation Forest는 고차원 데이터에서도 이상치를 잘 탐지할 수 있습니다. 자원 수입 데이터는 종종 여러 변수로 구성된 고차원 데이터이기 때문에, 이러한 특성을 가진 모델이 필요합니다.

d. 불규칙한 데이터 분포 처리

이 알고리즘은 데이터의 분포나 밀도에 의존하지 않기 때문에, 불규칙한 데이터 분포에서도 강력한 성능을 발휘합니다. 자원 수입 데이터는 비정상적이고 불규칙한 패턴을 보일 수 있어, Isolation Forest의 적용이 적절합니다.

e. 직관적인 이상치 점수

Isolation Forest는 각 데이터 포인트에 대해 직관적인 이상치 점수를 제공하여, 이상치의 정도를 쉽게 해석할 수 있습니다. 이는 자원 수입 데이터의 변동성을 모니터링하고, 이상치를 신속하게 감지하는 데 유용합니다.

f. 실시간 이상치 탐지

이 모델은 실시간 데이터 스트리밍 환경에서도 이상치를 효율적으로 탐지할 수 있습니다. 이는 자원 수입의 변동을 실시간으로 모니터링하고 대응하는 시스템 구축에 유리합니다.

(2) Prophet 모형

위와 상동.

Ⅳ. 분석 내용 및 분석결과

1. 핵심 광물의 리스크 요인분석

(1) 공급망 리스크

주요 생산국 의존성: 리튬, 니켈, 코발트 및 망간과 같은 자원은 특정 국가(예: 칠레, 콩고 민주 공화국, 인도네시아)에 집중적으로 매장되어 있습니다.

사례: 콩고 민주 공화국은 전 세계 코발트 생산의 70%를 차지합니다. 정치적 불안정성이나 무역 제한이 발생하면 글로벌 공급망에 큰 영향을 미칠 수 있습니다.

지정학적 리스크: 자원 생산국의 정치적, 경제적 상황에 따라 공급이 불안정할 수 있습니다.

사례: 2021년 칠레의 리튬 생산 관련 환경 규제 강화는 글로벌 리튬 공급에 영향을 미쳤습니다.

(2) 환경 및 건강 영향

환경 파괴: 자원 추출 과정에서 생태계 파괴와 오염이 발생할 수 있습니다.

(3) 가격 변동성

수요 증가: 전기차, 재생 에너지 저장장치 등 배터리 산업의 급성장으로 원자재 수요가 급증하고 있습니다.

사례: 2021년 전기차 수요 증가로 인해 리튬 가격이 급등한 바 있습니다.

시장 변동: 시장의 수요와 공급 변화, 투기적 거래 등이 원자재 가격의 급격한 변동을 초래할 수 있습니다.

사례: 코발트 가격은 2017년과 2018년 사이에 2배 이상 급등했으나, 이후 수요 감소와 공급 증가로 인해 가격이 급락했습니다.

여기에서 저희가 주목한 것은 가격 변동성과 가격과 수요량, 공급량 혹은 선행지표들이고, 공급에 대한 위험을 직접적 혹은 간접적으로 예측할 수 있다면 리스크에 대비할 수 있기 때문에 예측 모델을 만들기 위해 노력했습니다.

2. 세부 분석 결과

(1) 시계열 모델

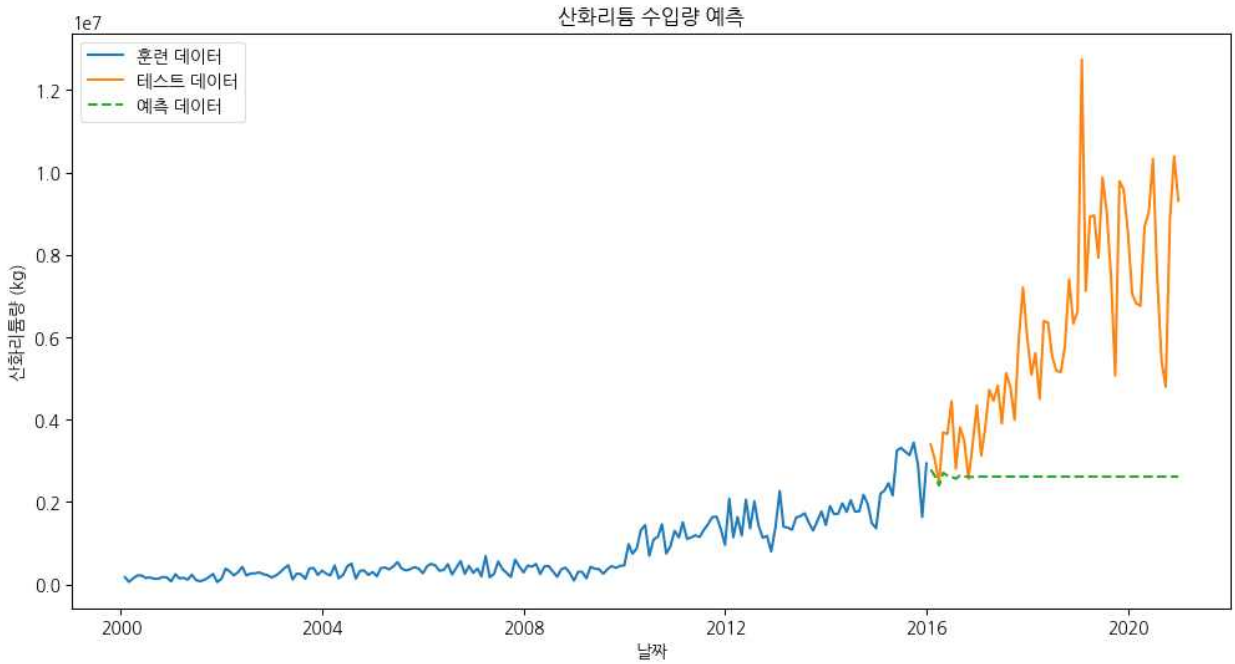
UNcomtrade[5] 의 공개된 수입, 수출자료를 가지고 1. ARIMA 모델, 2. 홀트-윈터스법, 3. 기본머신러닝, 4. LSTM , 5. 트랜스포머 방식의 시계열 예측 모델을 수행하였습니다.

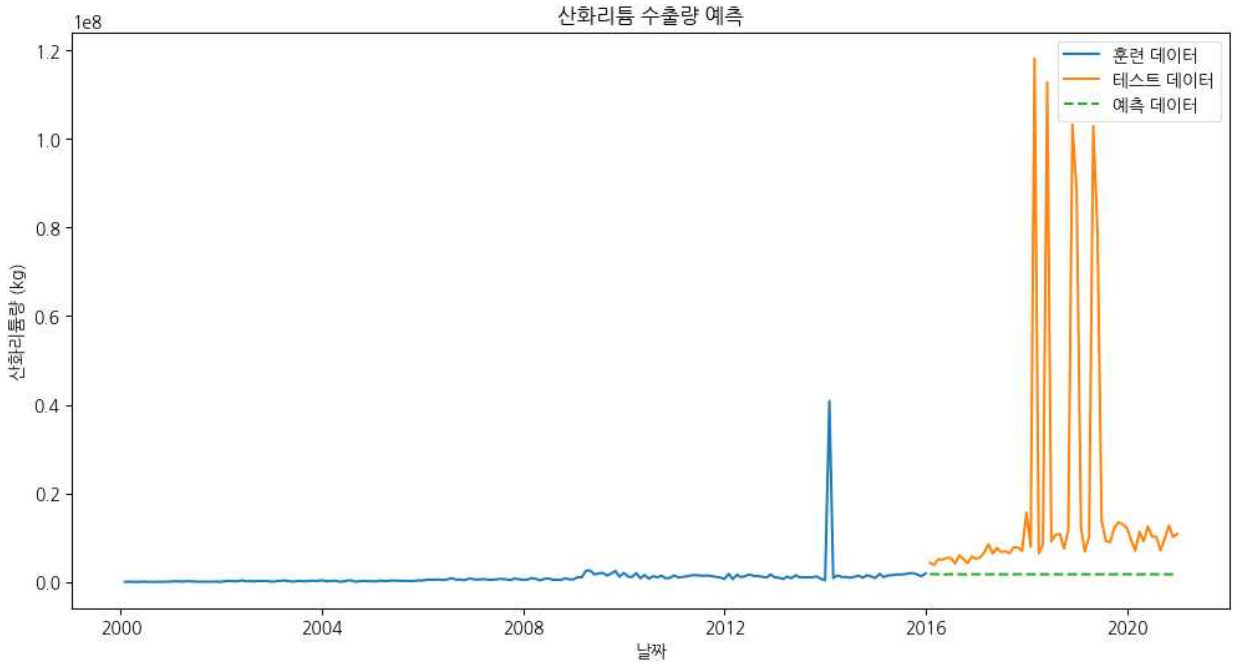
5) UNcomtrade : https://comtradeplus.un.org/

1) ARIMA 모형

산화리튬의 수입량과 수출량의 데이터를 가지고 가장 고전적인 시계열 방식의 모델로 예측을 해보았습니다. UNcomtrade의 사이트에서 다운로드 받은 데이터를 사용하였으며 2000년 1월부터 2015년 데이터를 가지고 학습하고, 2016년부터 2020년 까지의 산화리튬 전세계 수입량과 수출량을 예측하는 모델을 실행하였습니다. 아래는 결과입니다.

산화리튬 수출량 예측의 MSE: 1045397438694954.5
수입 데이터의 MSE: 17985598629690.15
수출 데이터의 MSE: 1045397438694954.5





산화리튬 수출량 예측의 MSE: 1045397438694954.5

SARIMAX Results

Dep. Variable:	NetWgt	No. Observations:	192
Model:	ARIMA(5, 1, 0)	Log Likelihood	-3127.623
Date:	Tue, 25 Jun 2024	AIC	6267.246
Time:	14:11:26	BIC	6286.760
Sample:	01-31-2000 - 12-31-2015	HQIC	6275.150
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8520	0.038	-22.481	0.000	-0.926	-0.778
ar.L2	-0.6750	0.046	-14.581	0.000	-0.766	-0.584
ar.L3	-0.4971	0.048	-10.297	0.000	-0.592	-0.402
ar.L4	-0.3269	0.045	-7.225	0.000	-0.416	-0.238
ar.L5	-0.1631	0.036	-4.577	0.000	-0.233	-0.093
sigma2	1.002e+13	3.81e-15	2.63e+27	0.000	1e+13	1e+13

Ljung-Box (L1) (Q):	0.13	Jarque-Bera (JB):	152891.62
Prob(Q):	0.72	Prob(JB):	0.00
Heteroskedasticity (H):	4031.92	Skew:	10.63
Prob(H) (two-sided):	0.00	Kurtosis:	139.97

산화리튬 수입량 예측의 MSE: 17985598629690.15

SARIMAX Results

Dep. Variable:	NetWgt	No. Observations:	192
Model:	ARIMA(5, 1, 0)	Log Likelihood	-2656.764
Date:	Tue, 25 Jun 2024	AIC	5325.528
Time:	14:11:26	BIC	5345.041
Sample:	01-31-2000 - 12-31-2015	HQIC	5333.431
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6144	0.050	-12.215	0.000	-0.713	-0.516
ar.L2	-0.3960	0.070	-5.680	0.000	-0.533	-0.259
ar.L3	-0.2419	0.083	-2.916	0.004	-0.404	-0.079
ar.L4	0.0562	0.058	0.977	0.328	-0.057	0.169
ar.L5	0.0024	0.054	0.044	0.965	-0.104	0.109
sigma2	7.236e+10	7.21e-13	1e+23	0.000	7.24e+10	7.24e+10

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	402.73
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	15.47	Skew:	-0.31
Prob(H) (two-sided):	0.00	Kurtosis:	10.09

수입량 예측의 MSE는 17,985,598,629,690.15로, 수출량 예측의 MSE(1,045,397,438,694,954.5)보다 훨씬 작습니다.

이는 수입량 예측이 수출량 예측보다 상대적으로 더 정확하다는 것을 의미할 수 있습니다.

수입량 예측 모델의 AIC는 5325.528, BIC는 5345.041입니다.

수출량 예측 모델의 AIC는 6267.246, BIC는 6286.760입니다.

AIC와 BIC 값이 작을수록 모델이 데이터를 더 잘 설명한다는 의미입니다. 수입량 예측 모델이 상대적으로 더 낮은 값을 가지므로 더 적합하다고 할 수 있습니다.

두 모델 모두 AR(1), AR(2), AR(3) 계수가 통계적으로 유의미합니다 (p-값 < 0.05).

AR(4), AR(5) 계수는 수입량 예측에서는 유의미하지 않으나, 수출량 예측에서는 유의미합니다.

사실상 ARIMA는 AR(자기회귀), I(차분), MA(이동평균) 으로 모두 과거 데이터로부터 데이터를 예측합니다. 따라서 과거와 비슷한 추세를 따라서 예측하는 것을 볼수 있으며 과거와 다른 새로운 추세 혹은 변동성이 생긴다면 예측이 어려워지는 것을 볼 수 있습니다.

2) 지수평활법(홀트 윈터스법)

홀트-윈터스법 (Holt-Winters Method)은 추세와 계절성이 있는 데이터를 예측할 때 사용하며, 일반적으로 단순 지수 평활법보다 주기성이 있는 데이터를 잘 분석 하는 것으로 알려져 있습니다.

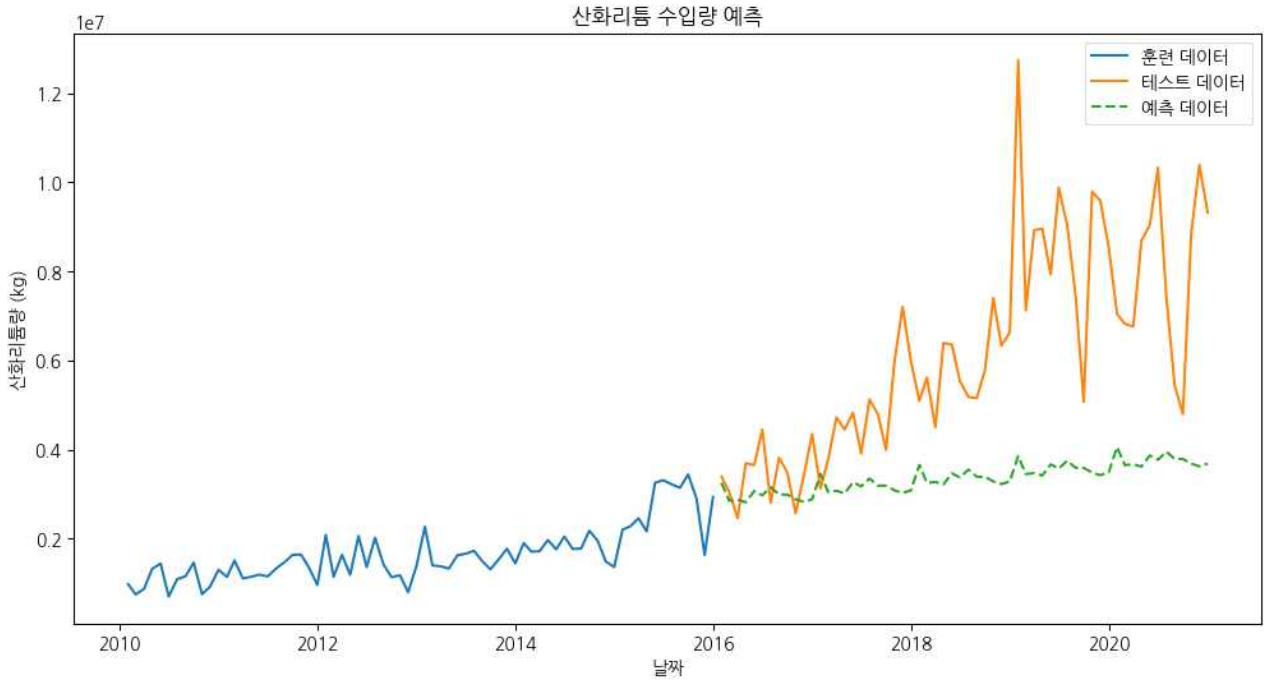
산화리튬 수입량 예측의 MSE: 12367788934872.885
산화리튬 수입량 예측의 MAE: 2834412.1284889993
산화리튬 수출량 예측의 MSE: 1042825483593919.8
산화리튬 수출량 예측의 MAE: 15763756.769611005
수입 데이터의 MSE: 12367788934872.885
수입 데이터의 MAE: 2834412.1284889993
수출 데이터의 MSE: 1042825483593919.8
수출 데이터의 MAE: 15763756.769611005

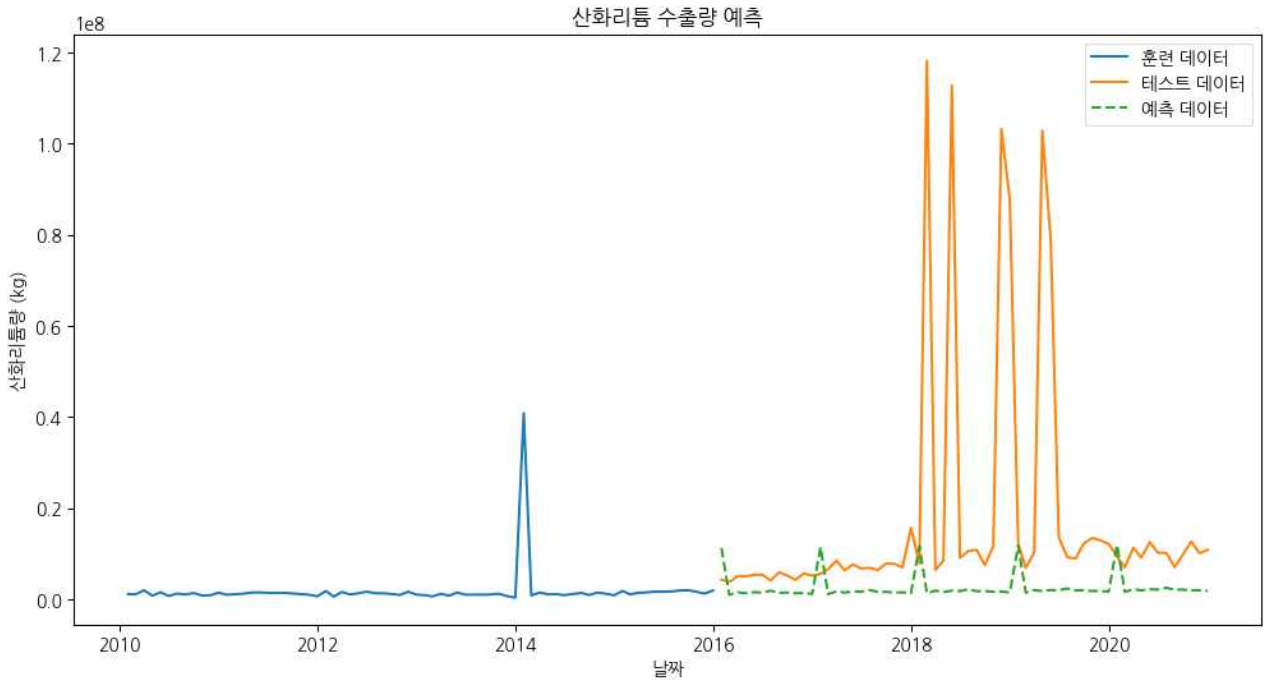
수출 데이터의 MSE가 수입 데이터의 MSE보다 훨씬 큽니다.

이는 수출 데이터 예측에서 예측값과 실제값 간의 차이가 더 크다는 것을 나타냅니다.

수출 데이터의 MAE가 수입 데이터의 MAE보다 훨씬 큽니다.

이는 수출 데이터 예측에서 절대적 오차가 더 크다는 것을 의미합니다.





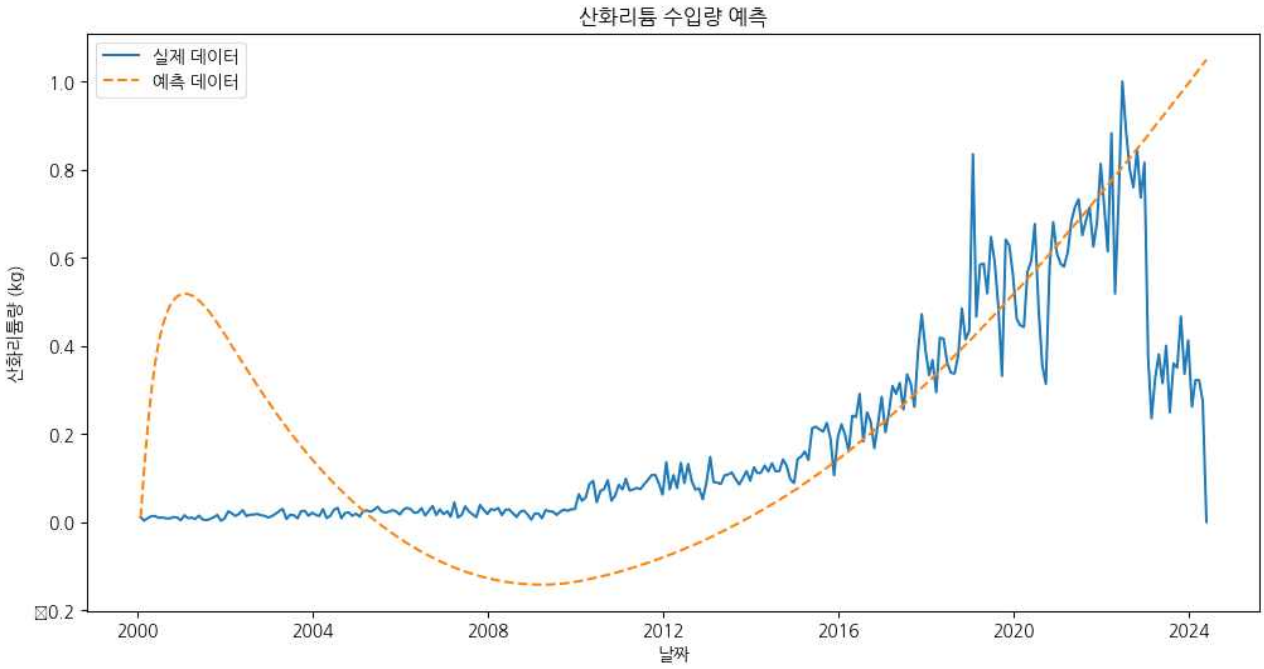
두 예측 결과 모두 MSE와 MAE 값이 매우 크기 때문에 예측 정확도가 낮음을 나타냅니다.

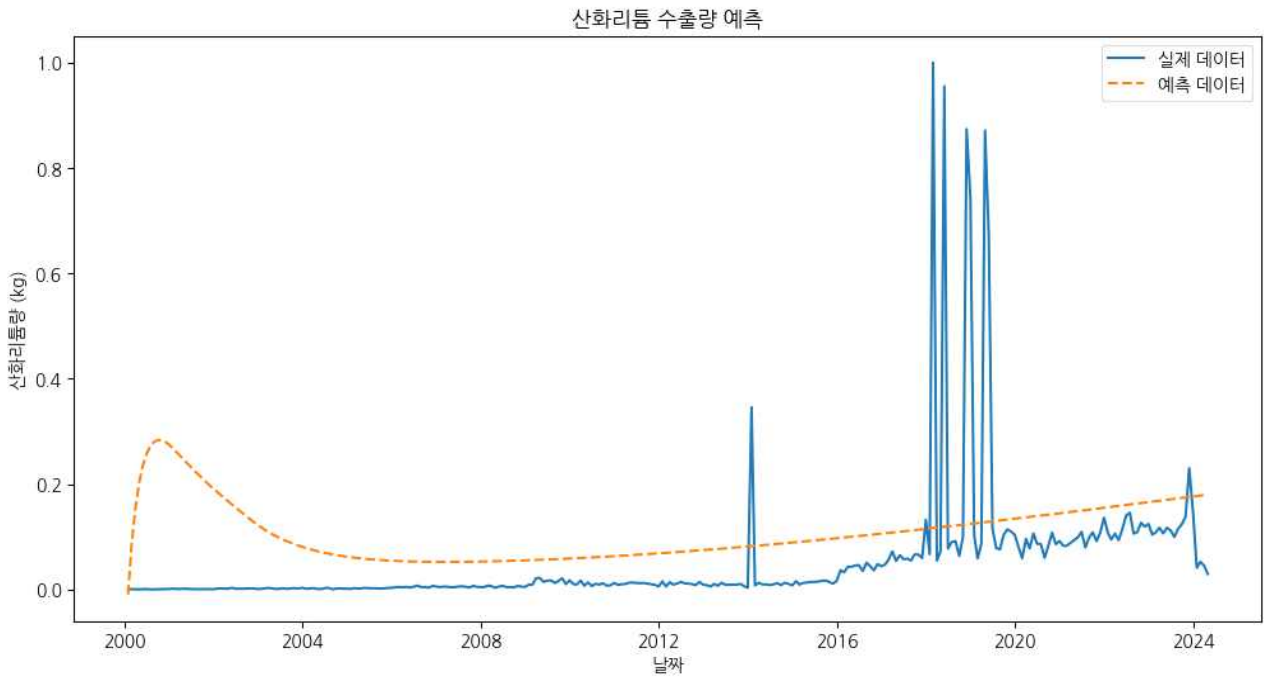
이는 Holt-Winters 모델이 산화리튬의 수입량과 수출량을 예측하는 데 적합하지 않을 수 있음을 의미합니다.

ARIMA 모델보다는 주기성과 추세를 반영하지만 역시 과거의 데이터와 다른 결과를 예측하는 것에는 상대적으로 낮은 성능을 보여주고 있습니다.

3) 기본머신러닝(합성곱)

다음은 Keras를 사용한 신경망모델을 사용하여 회귀분석을 진행하였습니다. 두 개의 dense layer(첫번째 64개, 두번째 32개)로 나누어서 DNN(딥러닝) 방식의 알고리즘을 실행하였습니다.





OLS Regression Results

Dep. Variable:	y	R-squared:	0.710
Model:	OLS	Adj. R-squared:	0.709
Method:	Least Squares	F-statistic:	498.1
Date:	Tue, 25 Jun 2024	Prob (F-statistic):	1.54e-56
Time:	13:30:03	Log-Likelihood:	397.19
No. Observations:	205	AIC:	-790.4
Df Residuals:	203	BIC:	-783.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0313	0.005	-6.421	0.000	-0.041	-0.022
x1	3.031e-05	1.36e-06	22.318	0.000	2.76e-05	3.3e-05

Omnibus:	32.169	Durbin-Watson:	0.517
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.108
Skew:	0.909	Prob(JB):	5.90e-11
Kurtosis:	4.487	Cond. No.	7.15e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.15e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
다중 상관계수 (R^2): 0.7104

OLS Regression Results

Dep. Variable:	y	R-squared:	0.123
Model:	OLS	Adj. R-squared:	0.119
Method:	Least Squares	F-statistic:	28.39
Date:	Tue, 25 Jun 2024	Prob (F-statistic):	2.64e-07
Time:	13:30:03	Log-Likelihood:	472.04
No. Observations:	204	AIC:	-940.1
Df Residuals:	202	BIC:	-933.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0054	0.003	-1.602	0.111	-0.012	0.001
x1	5.004e-06	9.39e-07	5.328	0.000	3.15e-06	6.86e-06

Omnibus:	424.534	Durbin-Watson:	2.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	237328.092
Skew:	12.369	Prob(JB):	0.00
Kurtosis:	168.254	Cond. No.	7.11e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.11e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
다중 상관계수 (R^2): 0.1232

첫 번째 OLS 회귀 분석 결과에서 결정 계수 (R-squared)는 0.710으로, 모델이 종속 변수의 변동성 중 71%를 설명한다는 의미입니다. 높은 설명력입니다.

t-통계량과 P-값은 x1의 t-통계량은 22.318로 매우 크며, p-값은 0.000으로 통계적으로 매우 유의미합니다. 상수항(const)의 t-통계량은 -6.421로, p-값은 0.000으로 유의미합니다.

두 번째 OLS 회귀 분석 결과에서 결정 계수 (R-squared)는 0.123으로, 모델이 종속 변수의 변동성 중 12.3%만을 설명한다는 의미입니다. 설명력이 낮습니다.

t-통계량과 P-값은 x1의 t-통계량은 5.328로 크며, p-값은 0.000으로 통계적으로 유의미합니다.

상수항(const)의 t-통계량은 -1.602로, p-값은 0.111로 유의미하지 않습니다.

첫 번째 모델은 높은 설명력을 가지며, 독립 변수 x1이 종속 변수 y에 미치는 영향이 통계적으로 유의미합니다.

두 번째 모델은 낮은 설명력을 가지며, 상수항이 통계적으로 유의미하지 않습니다.

두 모델 모두 잔차가 정규 분포를 따르지 않을 가능성이 있으며, 더빈-왓슨 통계량을 통해 첫 번째 모델은 잔차의 자기 상관 가능성을 확인할 수 있습니다.

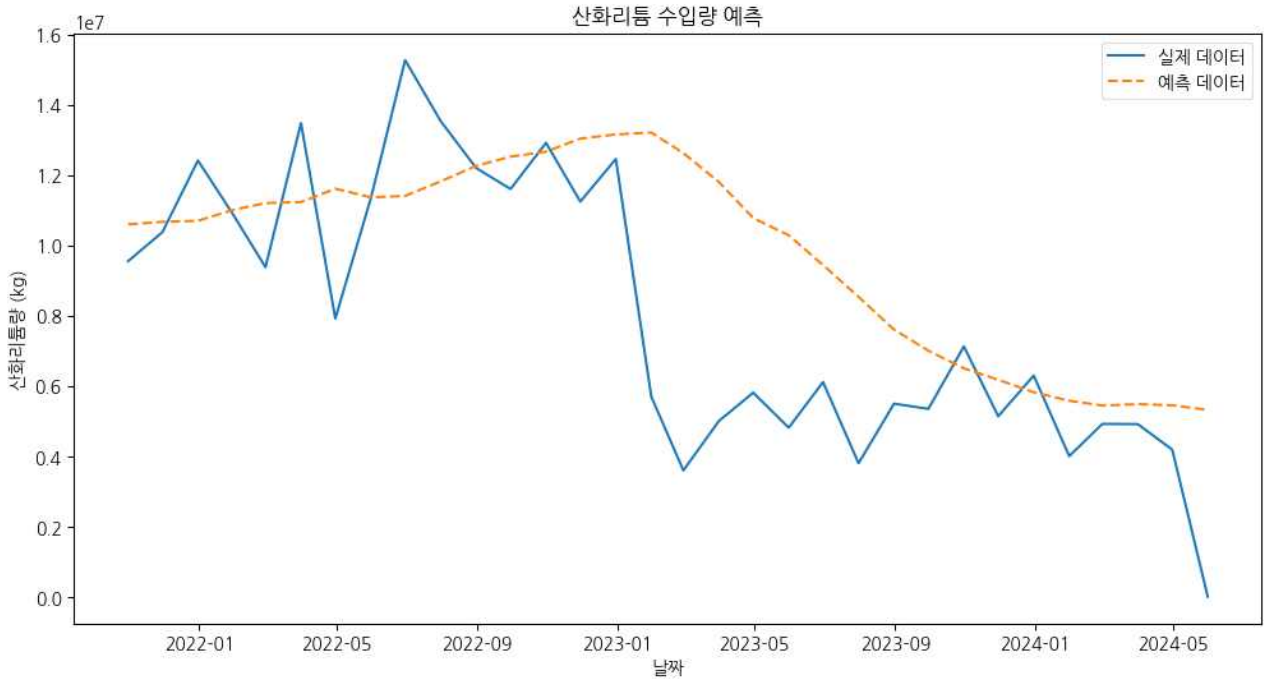
전제적으로 보면 처음에는 모델이 전혀 예측을 하지 못하다가 점차 학습을 진행 할수록 장기적인 변동을 예측하고 있습니다. 따라서 앞선 두 모델보다 어느정도 예측 성능이 나아졌습니다.

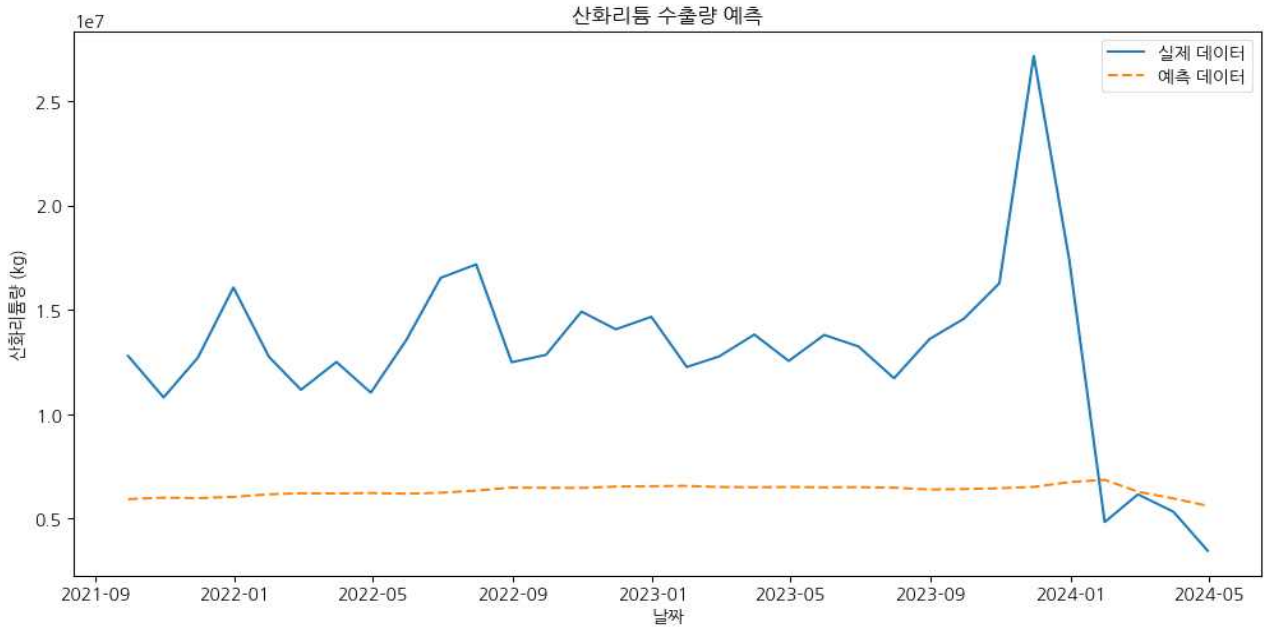
4) LSTM

LSTM은 장, 단기메모리 알고리즘으로 복잡한 흐름을 예측하는 모델입니다. 결과는 아래와 같습니다.

수입 모델의 MSE는 0.0680, MAE는 0.1921로 나타났습니다.

수출 모델의 MSE는 0.0046, MAE는 0.0605로 나타났습니다.





모델의 요약 및 성능 평가를 통해 수입 및 수출 예측 모델이 적절하게 구성되고, 성능이 확인되었습니다.

LSTM 레이어는 시계열 데이터를 효과적으로 처리하며, Dense 레이어는 최종 출력 값을 생성합니다.

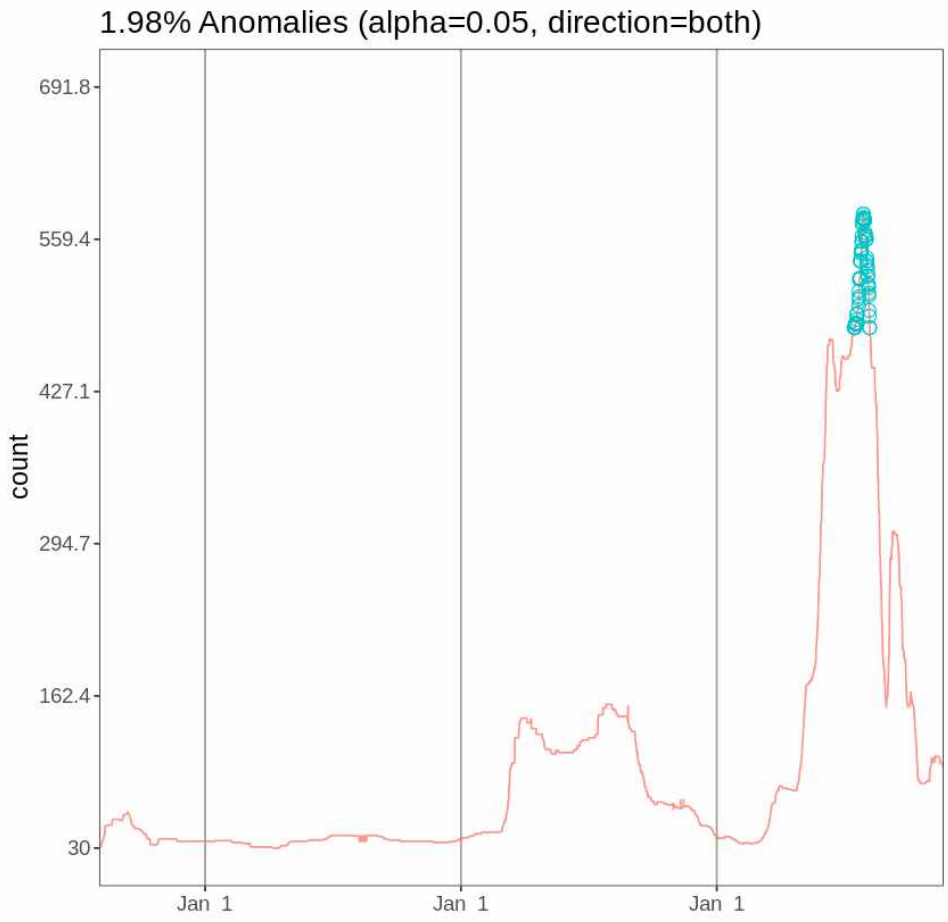
모델 성능 평가를 통해 예측 정확도를 확인할 수 있으며, 가중치 분석을 통해 모델의 학습 상태를 확인할 수 있습니다.

수출량에서도 추세는 예측하지만 변동성 부분에서 약간의 부적합이 일어난 것을 확인할 수 있었지만 대체적으로 수입량, 수출량 모두에서는 변동에 대해서 앞선 모형들 보다 잘 예측하고 있다고 볼 수 있습니다.

결론 : 이제까지 다양한 시계열 모델을 살펴보았지만 완벽하게 예측하는 모델은 없었으며 장, 단기적인 추세를 파악할때 딥러닝 혹은 LSTM 모형이 ARIMA 혹은 지수평활법 보다 성능이 상대적으로 좋았으며, 모든 시계열 모형이 단기적으로 예측하는 것에는 어느 정도의 한계가 있다는 것을 알 수 있었습니다.

(2) Prophet 모형

위험을 찾는다는 것은 어떠한 흐름에서 이상 흐름을 발견한다는 말과 동의어 이기 때문에 이 모델을 먼저 구현하였습니다.

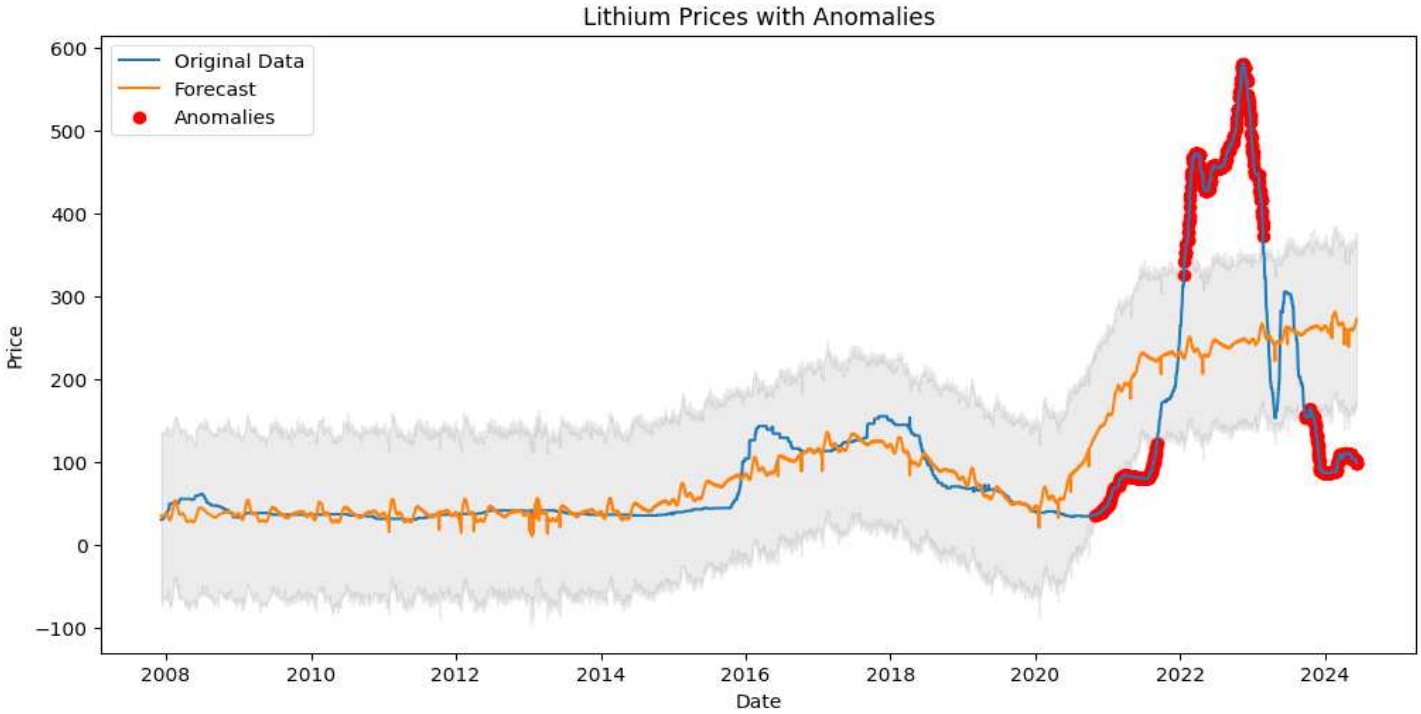


이 모델은 AnomalyDetection 패키지를 사용하여 R로 만들었고, 차후에 동일한 모델을 prophet 모듈을 사용하여 파이썬으로 만들었습니다.

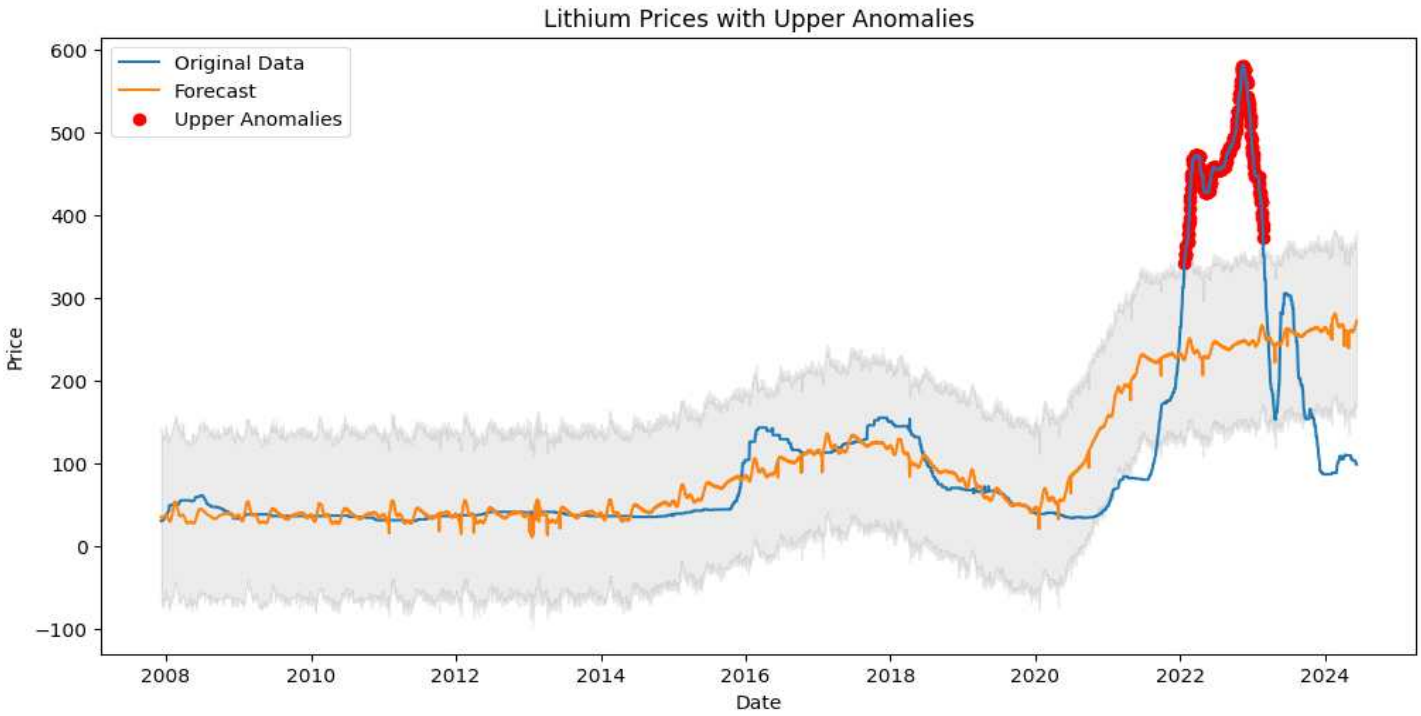
해당 데이터는 코미스[6]의 리튬 가격 데이터를 바탕으로 작성하였으며, 녹색부분은 전제데이터의 상위 2% 초과한 이상치들을 시각화 하여 나타낸 모델입니다.

6) 코미스 : https://www.komis.or.kr

다음은 동일한 내용의 모델로 파이썬으로 구현하였습니다.



주황색 선은 forecast 함수로 예측을 하였으며, 회색구간은 신뢰구간이고, 파란선은 실제 데이터입니다. 실제 데이터가 상위, 하위 20%를 초과하면 이상치로 표시하도록 하였습니다. 가격 데이터라서 아래의 이상치는 필요없을것 같지만 자세히 보면 가격이 비정상적으로 낮은 시기 이후에는 가격이 급등하는 것을 알 수 있습니다. 따라서 이상치를 측정한다면 하위와 상위 모두 확인해야 된다고 생각합니다. 아래는 상위 20%만 이상치를 표시하는 모델입니다.



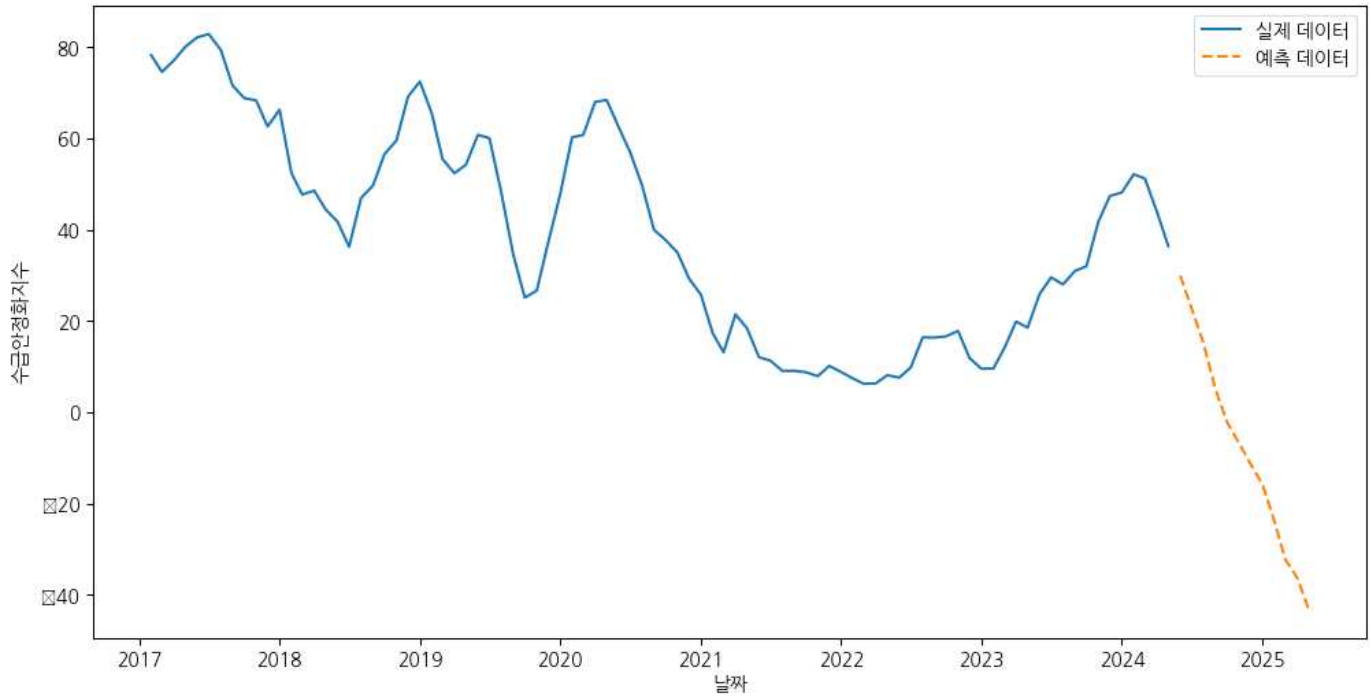
(3) 안정화 지수예측 및 분류모델

1) 홀트-윈터스법에 의한 예측 분류모형

해당 데이터는 코미스(https://www.komis.or.kr)의 니켈의 수급안정화지수 데이터를 활용하였습니다. 앞서 설명했던 시계열 방식을 적용하여 12개원(1년)의 안정화지수를 예측하고 예측한 데이터의 값이 지수들의 어느 등급에 맞는지 알려주는 모형입니다.

예측결과 :

2024-06-01	수급주의
2024-07-01	수급주의
2024-08-01	수급불안
2024-09-01	수급위기
2024-10-01	수급위기
2024-11-01	수급위기
2024-12-01	수급위기
2025-01-01	수급위기
2025-02-01	수급위기
2025-03-01	수급위기
2025-04-01	수급위기
2025-05-01	수급위기



2) DNN(LSTM)과 GARCH 모형을 이용한 예측 분류모형

Baltic Exchange[7]의 BDI[8] 데이터를 대상으로 LSTM모델을 적용하여 예측하는 모델과 GARCH 예측 모델을 동시에 적용하였습니다.

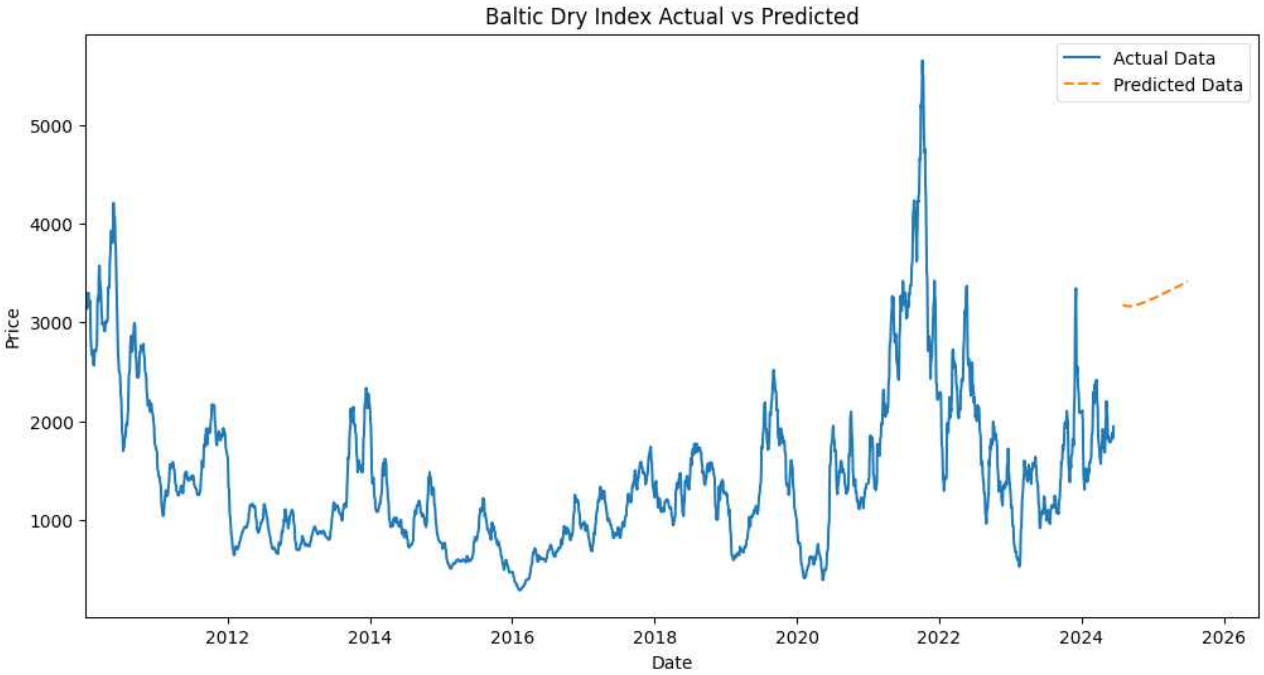
BMI 지수는 글로벌 무역상태를 반영하는 선행지표로 사용이되며 원자재의 가격의 변동성을

7) Baltic Exchange : www.balticexchange.com
8) 세계 주요 해운 노선의 운임을 기반으로 하는 지수

예측하는데 유용합니다. 또한 2008년 금융위기 당시는 BDI 지수가 급락하여 경기 침체의 선행지표로 활용되었습니다.

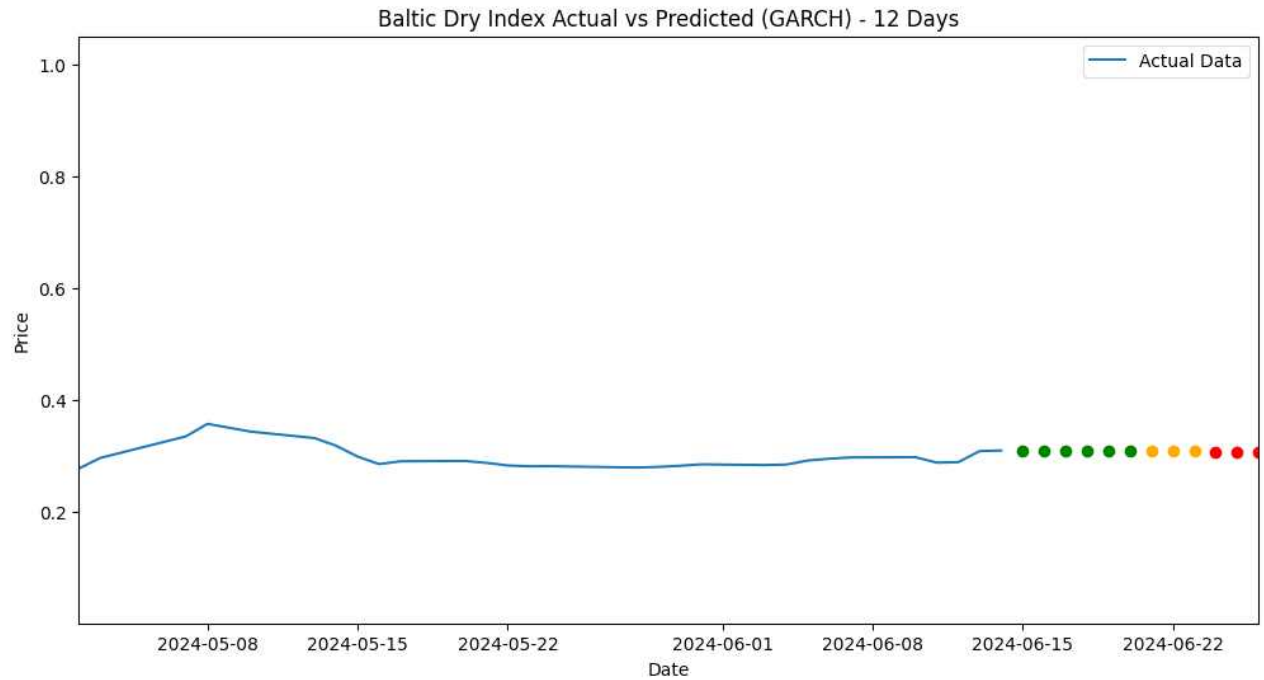
최근 몇 년간 중국의 경제성장 둔화와 환경규제 강화로 BDI 변동성이 커졌습니다. 2020년 코로나 19 팬데믹으로 인해 BDI 지수는 큰폭으로 하락했다가 이후 반등했습니다.

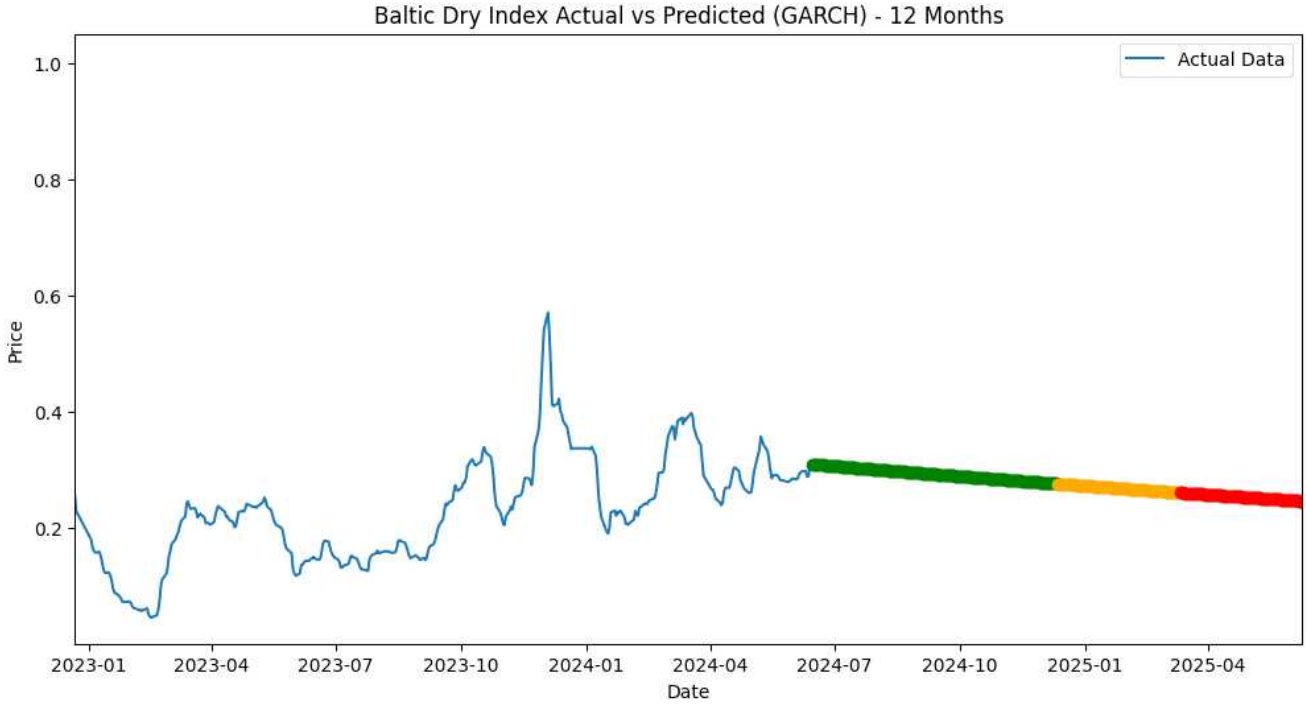
아래는 LSTM으로 12개월의 BDI지수를 예측한 결과입니다.



동일한 데이터로 GARCH 예측모형을 적용하여 변동성을 예측하고 위험지수를 분류하였습니다.

다음은 BDI 지수를 학습하고 12일 간의 변동성과 12개월간의 변동성을 예측하여 위험을 구분하여 표시한 그래프입니다.





각각의 위험도를 변동성 기준으로 4분위로 구분하여 위험도를 구분하였으며 녹색은 Low, 노랑색은 Medium, 빨강은 High로 각각 위험도를 구분하였습니다. 다음은 예측 데이터에 대한 구제적인 수치입니다.

12일 예측 데이터:

	Date	Predicted Price	Volatility	Risk
0	2024-06-15	0.309131	3.175029	Low
1	2024-06-16	0.308933	3.566663	Low
2	2024-06-17	0.308736	3.892836	Low
3	2024-06-18	0.308539	4.170891	Low
4	2024-06-19	0.308341	4.411636	Low
5	2024-06-20	0.308144	4.622401	Low
6	2024-06-21	0.307947	4.808455	Medium
7	2024-06-22	0.307751	4.973750	Medium
8	2024-06-23	0.307554	5.121353	Medium
9	2024-06-24	0.307358	5.253703	High
10	2024-06-25	0.307161	5.372780	High
11	2024-06-26	0.306965	5.480224	High

12개월 예측 데이터:

	Date	Predicted Price	Volatility	Risk
0	2024-06-15	0.309131	3.175029	Low
1	2024-06-16	0.308933	3.566663	Low
2	2024-06-17	0.308736	3.892836	Low
3	2024-06-18	0.308539	4.170891	Low
4	2024-06-19	0.308341	4.411636	Low
..	...	...	...	...
355	2025-06-05	0.246377	6.612591	High
356	2025-06-06	0.246219	6.612591	High
357	2025-06-07	0.246062	6.612591	High
358	2025-06-08	0.245905	6.612591	High
359	2025-06-09	0.245747	6.612591	High

[360 rows x 4 columns]

```
          h.1        h.2        h.3        h.4        h.5
436  159.223157  159.17418  159.125448  159.076958  159.02871

h.1: 일일 변동성 수준 = 12.62%, 변동성 수준 = 매우 높은 변동성 (매우 위험한 상태)
h.2: 일일 변동성 수준 = 12.62%, 변동성 수준 = 매우 높은 변동성 (매우 위험한 상태)
h.3: 일일 변동성 수준 = 12.61%, 변동성 수준 = 매우 높은 변동성 (매우 위험한 상태)
h.4: 일일 변동성 수준 = 12.61%, 변동성 수준 = 매우 높은 변동성 (매우 위험한 상태)
h.5: 일일 변동성 수준 = 12.61%, 변동성 수준 = 매우 높은 변동성 (매우 위험한 상태)

월간 변동성 수준:
Month 1: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 2: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 3: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 4: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 5: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 6: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 7: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 8: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 9: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 10: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 11: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
Month 12: 월간 변동성 수준 = 0.38%, 변동성 수준 = 낮은 변동성 (안정적인 상태)
```

(4) 위험성 평가 모델

1) GARCH 모형

GARCH 모형은 시계열 데이터의 변동성을 모델링하기 위해 사용되는 통계적 모델입니다. GARCH 모형은 시계열 데이터의 변동성이 시간에 따라 변화하는 특성을 반영합니다. 특히 특정 광물에 지금처럼 변동성이 커지는 경우 적합한 모형이라고 생각되어 집니다.

GARCH 모형을 적용하기 위해서 한국무역협회(https://stat.kita.net/newMain.screen)의 산화리튬의 수입금액 데이터에 대한 5일간의 변동성을 예측하고 그 위험성의 구간에 따라 위험도를 구분하였습니다. 또한 12개월의 월간 변동성을 예측하고 월간 변동성 위험도의 구간에 따라 위험도를 구분하였습니다. 결과는 아래와 같습니다.

또한 코미스(https://www.komis.or.kr)의 니켈 가격 데이터를 이용하여 동일한 분석을 진행하였습니다.

결과는 아래와 같습니다.

```
          h.1       h.2     h.3       h.4       h.5
3684  1.025256  1.042271  1.0591  1.075744  1.092207

h.1: 일일 변동성 수준 = 1.01%, 변동성 수준 = 중간 변동성 (일반적인 상태)
h.2: 일일 변동성 수준 = 1.02%, 변동성 수준 = 중간 변동성 (일반적인 상태)
h.3: 일일 변동성 수준 = 1.03%, 변동성 수준 = 중간 변동성 (일반적인 상태)
h.4: 일일 변동성 수준 = 1.04%, 변동성 수준 = 중간 변동성 (일반적인 상태)
h.5: 일일 변동성 수준 = 1.05%, 변동성 수준 = 중간 변동성 (일반적인 상태)
```

```
12개월치 월간 변동성 예측:
h.01     5.925951
h.02     6.620848
h.03     7.249439
h.04     7.827713
h.05     8.366112
h.06     8.871897
h.07     9.350363
h.08     9.805510
h.09    10.240447
h.10    10.657649
h.11    11.059124
h.12    11.446525
Name: 2024-06-30 00:00:00, dtype: float64

12개월치 월간 변동성 예측 및 수준 구분:
Month 1: 변동성 = 5.93%, 수준 = 높은 변동성 (불안정한 상태)
Month 2: 변동성 = 6.62%, 수준 = 높은 변동성 (불안정한 상태)
Month 3: 변동성 = 7.25%, 수준 = 높은 변동성 (불안정한 상태)
Month 4: 변동성 = 7.83%, 수준 = 높은 변동성 (불안정한 상태)
Month 5: 변동성 = 8.37%, 수준 = 높은 변동성 (불안정한 상태)
Month 6: 변동성 = 8.87%, 수준 = 높은 변동성 (불안정한 상태)
Month 7: 변동성 = 9.35%, 수준 = 높은 변동성 (불안정한 상태)
Month 8: 변동성 = 9.81%, 수준 = 높은 변동성 (불안정한 상태)
Month 9: 변동성 = 10.24%, 수준 = 높은 변동성 (불안정한 상태)
Month 10: 변동성 = 10.66%, 수준 = 높은 변동성 (불안정한 상태)
Month 11: 변동성 = 11.06%, 수준 = 높은 변동성 (불안정한 상태)
Month 12: 변동성 = 11.45%, 수준 = 높은 변동성 (불안정한 상태)
```

무역의 수입금액은 일일 변동수준이 커 일일변동성은 매우 위험한 상태로 나왔지만 월별 변동성은 안정화되어 있으며 반대로 니켈의 가격은 일일변동은 안정화되어 있지만 월별 변동성은 매우 불안정한 상태로 볼 수 있습니다.

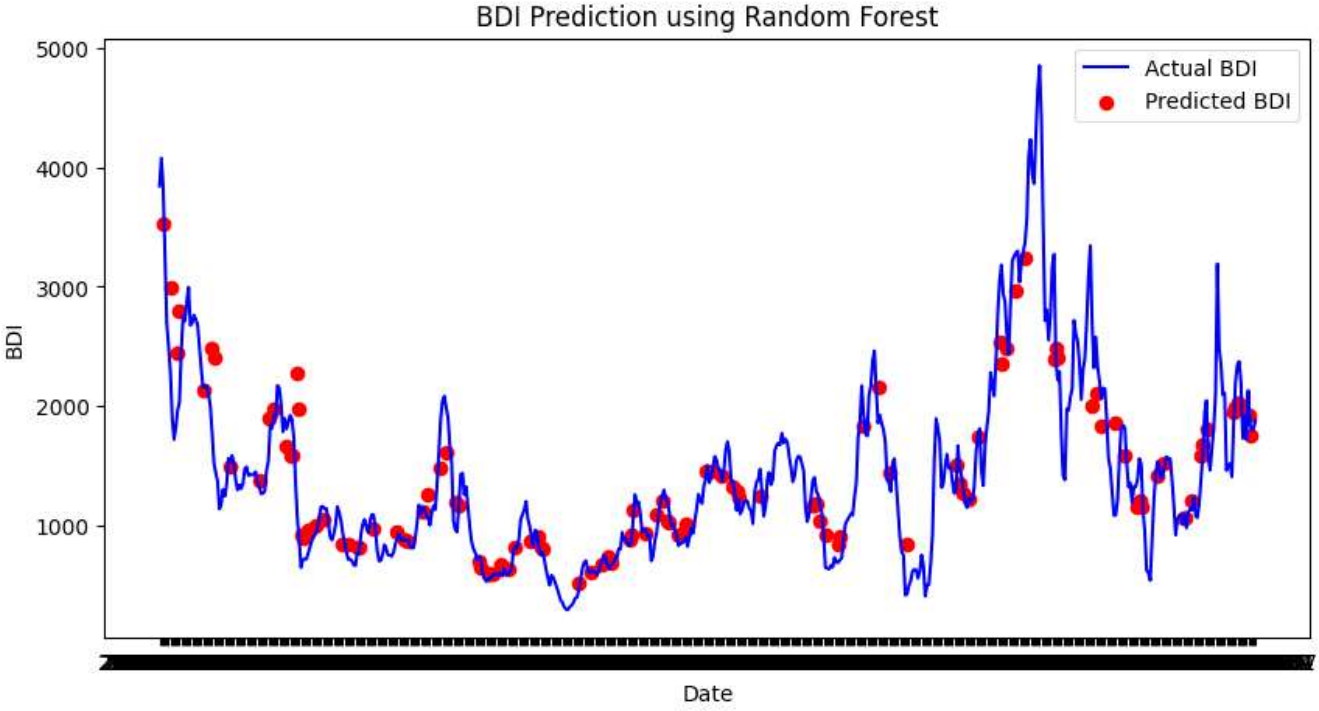
GARCH 모형은 앞서 언급했던 시계열 모형의 단점인 이상치를 예측하는 것에 성능이 떨어지는 부분을 간접적으로 보완할 수 있습니다.

만약 GARCH 모형에 의하여 어떠한 수치가 변동성이 커지는 방향으로 결과값이 나왔다면 위에서 예측했던 예측모델 예측치를 넘어서는 이상치가 나올확률이 커지는 것이기 때문에 예측모델과 병행하여 사용한다면 효과가 높을 것으로 생각되어 집니다.

실제 활용에서는 시계열 예측에서 데이터의 장, 단기방향성을 예측하고, GARCH 모형으로 데이터의 변동성(리스크)를 측정해서 보완한다면, 예측의 정확도가 높아질 것이라 생각합니다.

2) 변동성 예측 모델(랜덤 포레스트 모델)

원래는 금가격의 변동성을 예측하는 모델로 시간에 따라 변화하는 가격 혹은 변동하는 데이터를 예측하는 모형으로 많이 사용됩니다. 같은 알고리즘을 활용하여 니켈과 리튬, 망간, 텅스텐, 코발트 가격을 이용하여 BDI 지수를 예측하는 모델을 구현해보았습니다.



Mean Squared Error: 94039.93743471075
R^2 Score: 0.8052221161382161

위의 결과에서 실제의 결과에 비하여 예측치가 BDI의 급등(이상치)을 정확하게 예측하지 못했기 때문에 MSE 가 높게 나왔습니다. 하지만 대략적인 변동성을 모두 예측하고 있으며, 정확성의 측면(R^2)에서는 상당히 높은 수치가 나오고 있습니다. 따라서 다른 지수를 참고하면 활용한다면 매우 중요한 의사결정 모델로 활용할 수 있다고 생각합니다.

이 모델의 장점은 독립변수의 제한이 없으며, 상관관계만 있다면 어떠한 변수든지 투입값으로 활용할 수 있다는 점입니다. 또한 종속변수도 위의 5개의 변수중 어느 것이든 선택할 수 있습니다.(예를 들어 BDI지수와 망간, 텅스텐, 코발트, 니켈의 가격으로 리튬의 가격을 예측할수 있습니다.) 하지만 예측을 위해서는 사전에 나머지 변수들이 확정이 되어 있어야 한다는 전제가 필요합니다.

따라서 이 모델로 미래값을 예측한다면 다른 독립변수들의 예측치를 활용하여, 그 예측치를 랜덤포레스트 모형의 투입값으로 활용한다면 또 하나의 예측 모델로써 활용할 수 있다고 생각합니다.

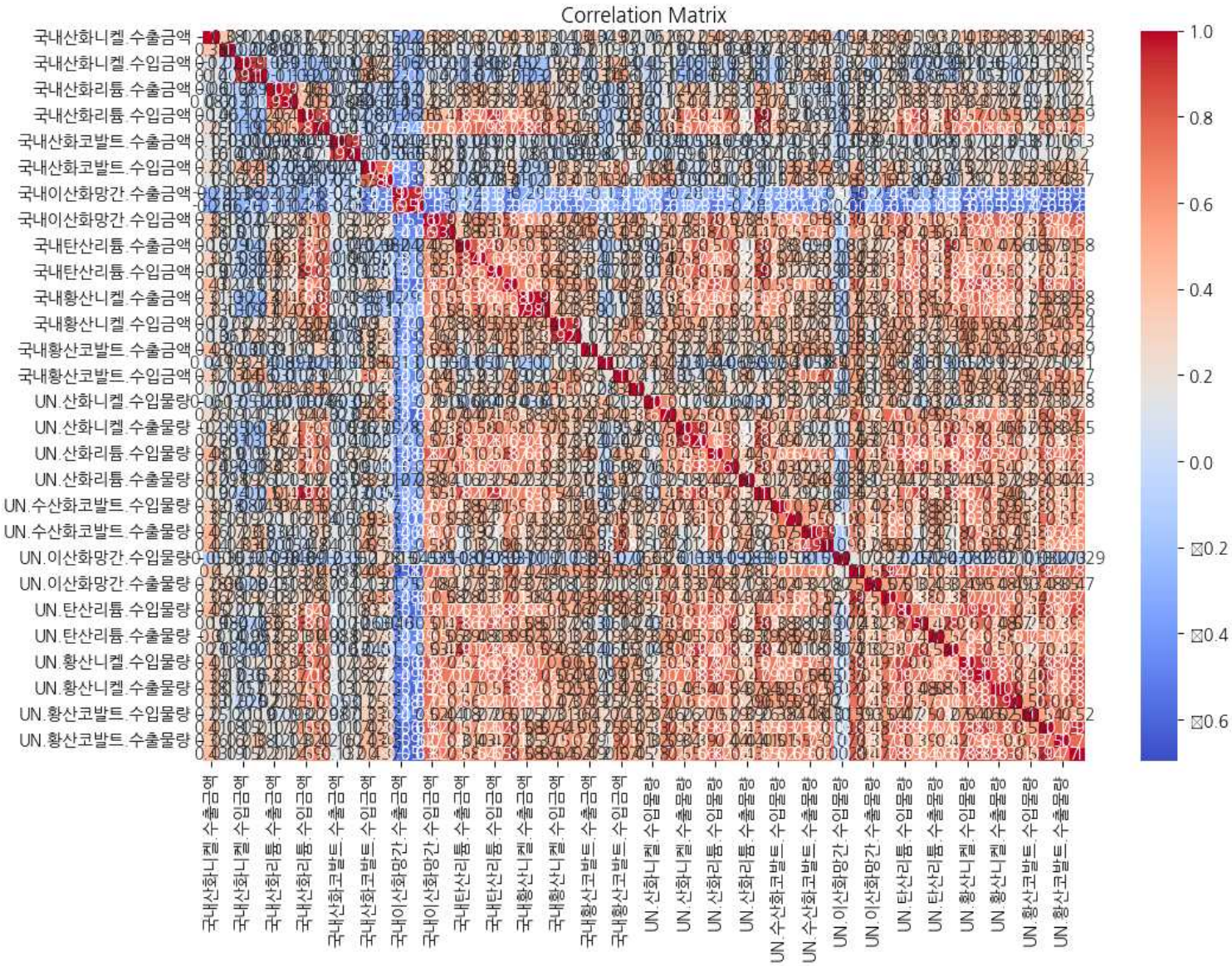
(5) 다변량 위험 예측모델

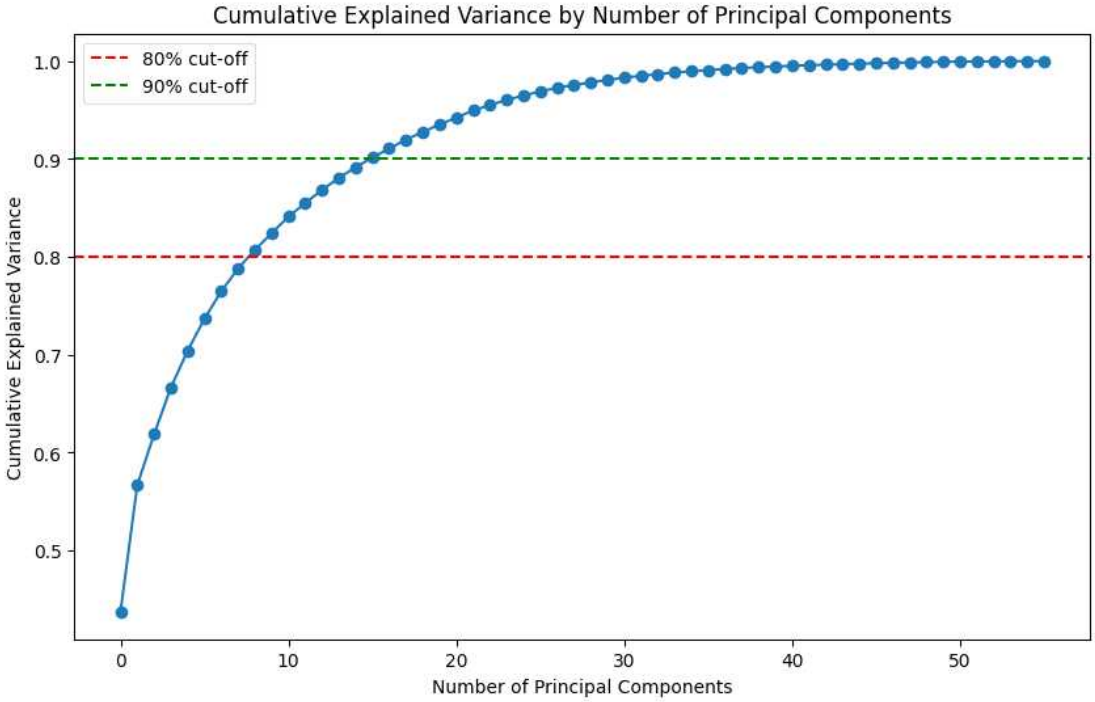
우선적으로 다변량 모델의 분석을 위해서 UNcomtrade(https://comtradeplus.un.org/)의 전세계 주요광물(니켈과 리튬, 망간, 텅스텐, 코발트)의 수입중량과, 수입금액, 수출중량과, 수출금액과 한국무역협회(https://stat.kita.net/newMain.screen)의 주요광물의 국내 수입금액과 수입중량, 국내 수출금액과 수출중량의 데이터를 편집하여 하나의 데이터로 편집하였습니다.

그로인하여 총 56개의 변수의 데이터가 만들어졌으며 변수들간의 상관성이 너무 높기 때문에

PCA분석(주성분분석)방법을 활용하여 변수들의 차원을 낮추고 상관성을 낮추었습니다.
우선적으로 모든 다변량 분석에선 국내 광물의 수입금액(7개의 변수)들을 종속변수로 선정하였고 그에 따라 PCA를 진행하였을 때 그것들을 제외하고 선정하였습니다.
나머지 49개의 변수들을 모두 종속 변수로 선정하였으며 적합한 변수의 개수는 모델 판정 결과 15개가 나왔기 때문에 49개를 15개로 줄이고 분석을 진행 하였습니다.

분석을 진행하기 전에 사전에 변수들간의 사전 상관관계를 히트맵으로 표시하였습니다.
상관관계 매트릭스는 다음과같습니다.



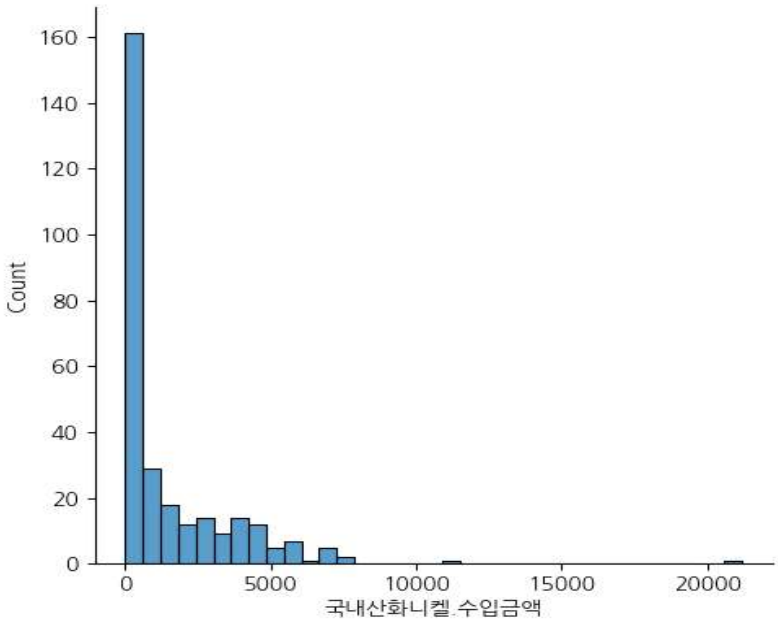


위의 그림은 PCA를 적용하기위한 설명 분산비율을 그래프로 시각화 시킨 내용입니다. 90%의 설명 비율이 가장 모델 효율성이 좋다고 판단하였고 그에 해당하는 주성분의 개수를 15개로 선정하였습니다.

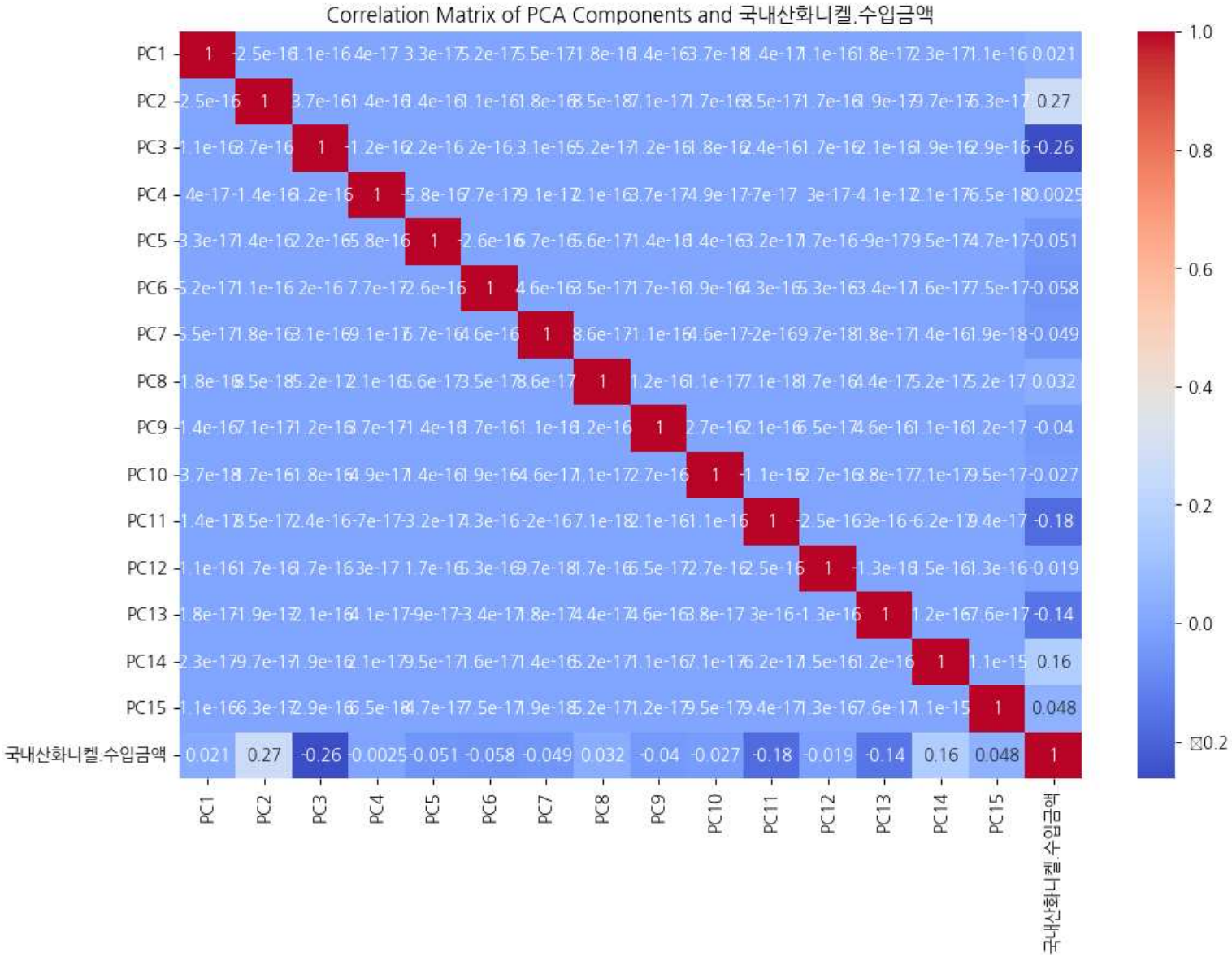
1) 다변량 회귀 분석(기본모델)

모든 종속 변수에 대한 분석에 대한 결과를 분석하기 어렵기 때문에 국내 산화니켈의 수입금액에 대한 다변량 회귀분석을 진행하였습니다. 나머지 변수들도 같은 방법으로 분석이 가능합니다.

다음은 국내산 니켈의 수입금액에 대한 히스토그램입니다.

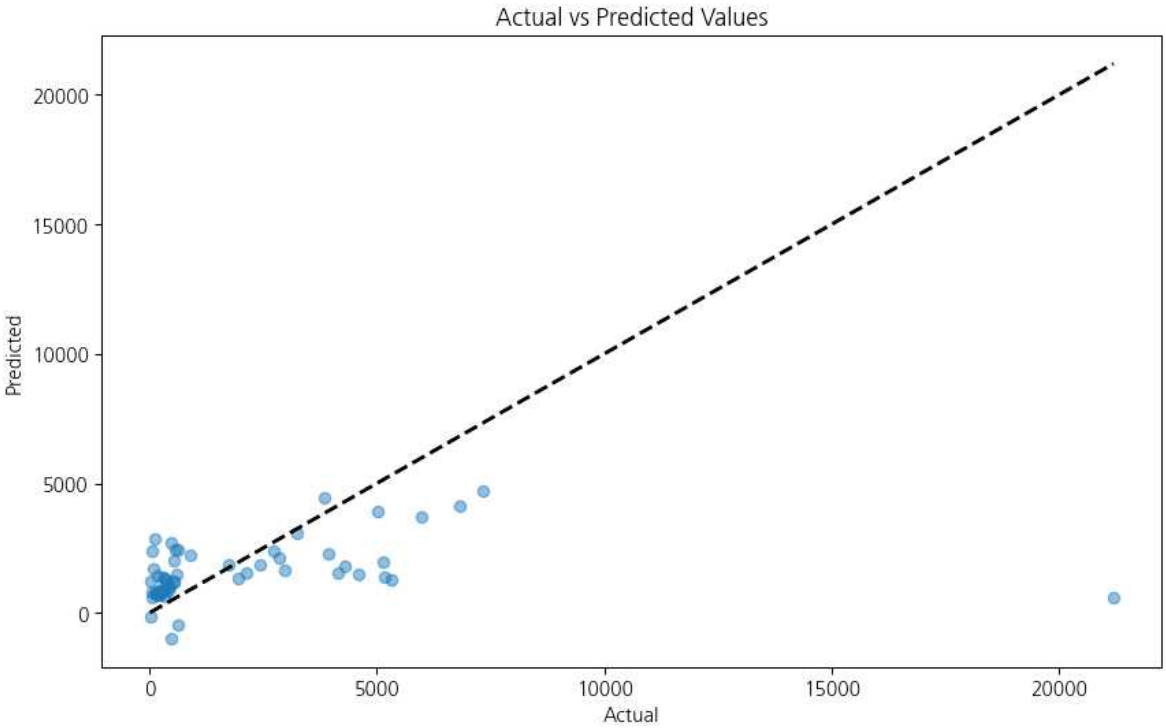


우선 PCA적용 후 상관 분석에 대한 시각화 매트릭스는 다음과 같습니다.



15개의 변수로 차원을 낮추어 분석하기 용이함과 동시에 변수들간의 상관관계가 확연히 줄어든 것을 확인할 수 있습니다.

적용한 모델은 LinearRegression으로 일반적인 다변량 분석을 진행하였으며, 일변량의 종속변수가 나오고 다변량의 독립변수가 투입되는 다변량 분석을 실시하였습니다.
아래는 모델 적용의 결과입니다.



OLS Regression Results

Dep. Variable:	국내산화니켈.수입금액	R-squared:	0.325
Model:	OLS	Adj. R-squared:	0.278
Method:	Least Squares	F-statistic:	6.918
Date:	Tue, 25 Jun 2024	Prob (F-statistic):	3.31e-12
Time:	14:34:35	Log-Likelihood:	-2030.9
No. Observations:	232	AIC:	4094.
Df Residuals:	216	BIC:	4149.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1480.3363	104.765	14.130	0.000	1273.843	1686.829
PC1	1.482e-07	1.2e-07	1.236	0.218	-8.82e-08	3.85e-07
PC2	1.03e-06	1.73e-07	5.944	0.000	6.88e-07	1.37e-06
PC3	-1.468e-06	3.14e-07	-4.678	0.000	-2.09e-06	-8.49e-07
PC4	-5.096e-08	4.57e-07	-0.112	0.911	-9.51e-07	8.49e-07
PC5	-8.834e-07	7.93e-07	-1.114	0.266	-2.45e-06	6.79e-07
PC6	-4.386e-07	9.79e-07	-0.448	0.655	-2.37e-06	1.49e-06
PC7	-2.281e-06	1.76e-06	-1.295	0.197	-5.75e-06	1.19e-06
PC8	9.15e-07	1.89e-06	0.485	0.628	-2.8e-06	4.63e-06
PC9	-1.459e-06	2.42e-06	-0.602	0.548	-6.23e-06	3.31e-06
PC10	-2.793e-06	2.66e-06	-1.051	0.295	-8.03e-06	2.45e-06
PC11	-9.497e-06	2.73e-06	-3.473	0.001	-1.49e-05	-4.11e-06
PC12	-2.142e-06	3.27e-06	-0.656	0.513	-8.58e-06	4.3e-06
PC13	-1.101e-05	3.41e-06	-3.228	0.001	-1.77e-05	-4.29e-06
PC14	2.105e-05	6.12e-06	3.441	0.001	8.99e-06	3.31e-05
PC15	3.499e-06	5.82e-06	0.601	0.549	-7.98e-06	1.5e-05

Omnibus:	128.507	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	912.912
Skew:	2.092	Prob(JB):	5.80e-199
Kurtosis:	11.772	Cond. No.	8.81e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
[2] The condition number is large, 8.81e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Mean Squared Error: 9541897.973048335
R-squared: 0.08131860712998551

다음은 회귀 분석 요약입니다.

Dep. Variable: 종속 변수는 국내산화니켈.수입금액입니다.
R-squared는 결정 계수는 0.325로, 이 모델이 종속 변수의 변동성을 약 32.5% 설명함을 나타냅니다.
Adj. R-squared는 수정된 결정 계수는 0.278로, 변수의 수를 고려한 결과입니다.
F-statistic는 6.918로, 모델의 전체 유의성을 검정합니다.
Prob (F-statistic)는 F 통계량의 유의 확률은 3.31e-12로, 모델이 통계적으로 유의미함을 나타냅니다.

계수 (Coefficients)
절편(1480.3363)은 모든 PCA 주성분이 0일 때의 국내산화니켈.수입금액을 나타냅니다.
PC1 ~ PC15: 각 주성분의 회귀 계수입니다.
PC2: 계수는 1.03e-06으로, p값은 0.000으로 유의미합니다.
PC3: 계수는 -1.468e-06으로, p값은 0.000으로 유의미합니다.
PC11: 계수는 -9.497e-06으로, p값은 0.001로 유의미합니다.
PC13: 계수는 -1.101e-05으로, p값은 0.001로 유의미합니다.
PC14: 계수는 2.105e-05으로, p값은 0.001로 유의미합니다.
나머지 주성분은 p값이 0.05 이상으로 유의미하지 않습니다

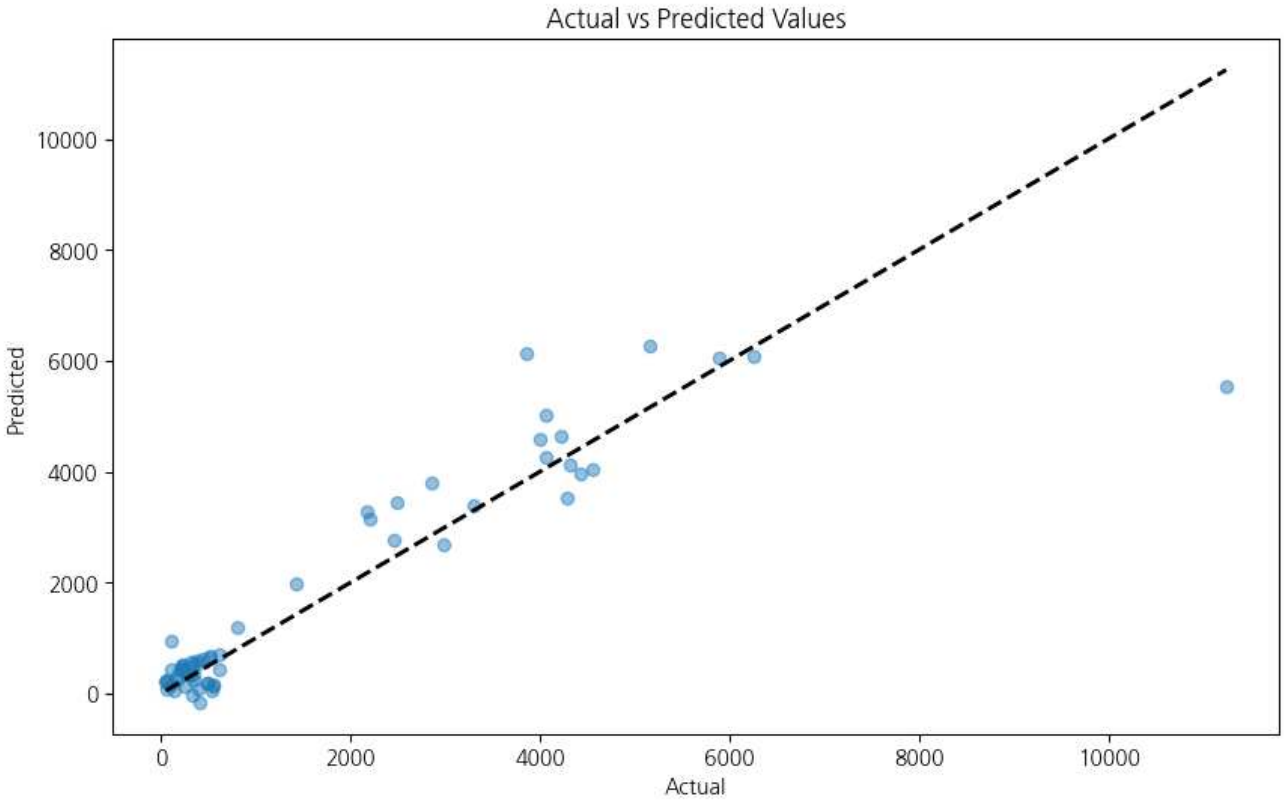
통계적 유의성 (P>|t|)
P>|t|: 각 변수의 p값으로, 0.05 이하인 경우 해당 변수가 종속 변수에 유의미한 영향을 미친다고 볼 수 있습니다.
유의미한 변수: PC2, PC3, PC11, PC13, PC14
유의미하지 않은 변수: PC1, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC12, PC15

Mean Squared Error (MSE)은 9541897.97로, 예측값과 실제값 간의 평균 제곱 오차를 나타냅니다.
R-squared는 0.0813로, 테스트 데이터에 대한 결정 계수입니다. 이는 모델이 테스트 데이터의 변동성을 약 8.13% 설명함을 나타냅니다.

모델의 결정 계수는 낮아, 독립 변수들이 국내산화니켈.수입금액을 설명하는데 충분하지 않습니다.
일부 PCA 주성분은 통계적으로 유의미하지만, 독립변수가 종속변수를 설명하는 비율이 32.5% 이므로 모델의 신뢰성이 매우 낮습니다.

따라서 같은 모델을 진행하되 변수들의 설명력을 높이기 위하여 기존의 모든 데이터를 투입하는 대신에 국외 변수들를 모두 제외하고 국내 변수들만을 대상으로 다시 모델 학습을 진행 하였습니다.

그에 대한 결과는 다음과 같습니다.



OLS Regression Results

Dep. Variable:	국내산화니켈.수입금액	R-squared:	0.865
Model:	OLS	Adj. R-squared:	0.855
Method:	Least Squares	F-statistic:	92.43
Date:	Tue, 25 Jun 2024	Prob (F-statistic):	2.54e-85
Time:	10:10:17	Log-Likelihood:	-1893.5
No. Observations:	233	AIC:	3819.
Df Residuals:	217	BIC:	3874.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1566.5557	57.057	27.456	0.000	1454.098	1679.013
PC1	-0.0001	2.19e-05	-4.603	0.000	-0.000	-5.76e-05
PC2	0.0004	3.08e-05	13.648	0.000	0.000	0.000
PC3	0.0002	5.39e-05	4.164	0.000	0.000	0.000
PC4	0.0005	7.75e-05	7.059	0.000	0.000	0.001
PC5	-4.456e-05	8.34e-05	-0.534	0.594	-0.000	0.000
PC6	0.0006	0.000	4.587	0.000	0.000	0.001
PC7	0.0005	0.000	2.506	0.013	0.000	0.001
PC8	-0.0028	0.000	-10.375	0.000	-0.003	-0.002
PC9	-0.0004	0.000	-1.514	0.131	-0.001	0.000
PC10	0.0033	0.001	6.418	0.000	0.002	0.004
PC11	-0.0049	0.001	-8.512	0.000	-0.006	-0.004
PC12	0.0181	0.001	26.091	0.000	0.017	0.020
PC13	-0.0031	0.002	-1.644	0.102	-0.007	0.001
PC14	0.0007	0.001	0.500	0.617	-0.002	0.003
PC15	0.0118	0.014	0.853	0.394	-0.015	0.039

Omnibus:	249.356	Durbin-Watson:	1.928
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19790.620
Skew:	3.946	Prob(JB):	0.00
Kurtosis:	47.455	Cond. No.	3.19e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
[2] The condition number is large, 3.19e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
다중 상관계수 (R^2): 0.8647

모델의 요약은 다음과 같습니다.

결정 계수 (R^2)는 0.865이므로 모델이 국내산화니켈.수입금액의 변동성을 86.5% 설명할 수 있음을 나타냅니다.
수정된 결정 계수 (Adj. R^2)는 0.855이고 이는 변수의 수를 고려한 R^2 값으로, 85.5%의 설명력을 가집니다. 모델이 과적합(overfitting)되지 않았음을 시사합니다.
F-통계량은 92.43이며, 이는 모델 전체의 유의성을 검정하는 지표로, 매우 높은 값입니다.
모델이 유의미함을 나타내며, 독립 변수들이 종속 변수에 유의한 영향을 미친다는 것을 의미합니다.
F-통계량의 p-값은 2.54e-85인 매우 낮은 값으로, 모델이 유의함을 강하게 지지합니다.

상수항은 1566.5557 모든 독립 변수가 0일 때 국내산화니켈.수입금액의 예상 값입니다.

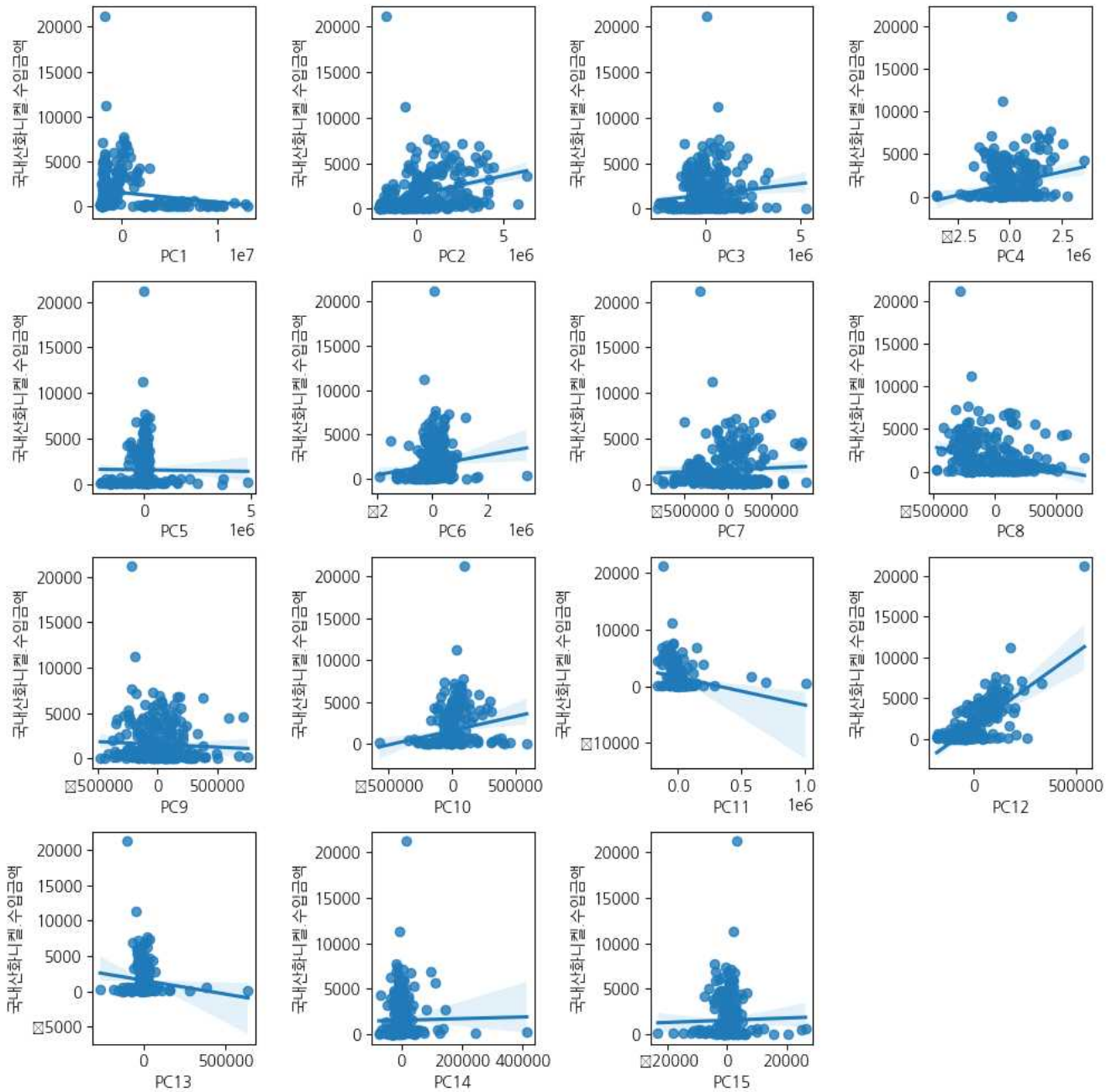
유의한 주성분은 다음과 같습니다.
PC1: -0.0001 (매우 유의)
PC2: 0.0004 (매우 유의)
PC3: 0.0002 (매우 유의)
PC4: 0.0005 (매우 유의)
PC6: 0.0006 (매우 유의)
PC7: 0.0005 (유의)
PC8: -0.0028 (매우 유의)
PC10: 0.0033 (매우 유의)
PC11: -0.0049 (매우 유의)
PC12: 0.0181 (매우 유의)

유의하지 않은 주성분 은 다음과 같습니다.
PC5: -4.456e-05 (유의하지 않음)
PC9: -0.0004 (유의하지 않음)
PC13: -0.0031 (유의하지 않음)
PC14: 0.0007 (유의하지 않음)
PC15: 0.0118 (유의하지 않음)

유의미한 변수: PC1, PC2, PC3, PC4, PC6, PC7, PC8, PC10, PC12

유의미하지 않은 변수: PC5, PC9, PC11, PC13, PC14, PC15

주성분의 유의성에 대한 확인은 다음의 회귀분석 plot으로도 확인할 수 있습니다.
아래의 그래프에서 유의미한 변수는 y축인 국내 산화니켈 수입금액과 서로 강한 상관성을 보아고, 유의미하지 않은 변수는 상대적으로 서로 상관성이 서로 약하게 표시되어 있습니다.

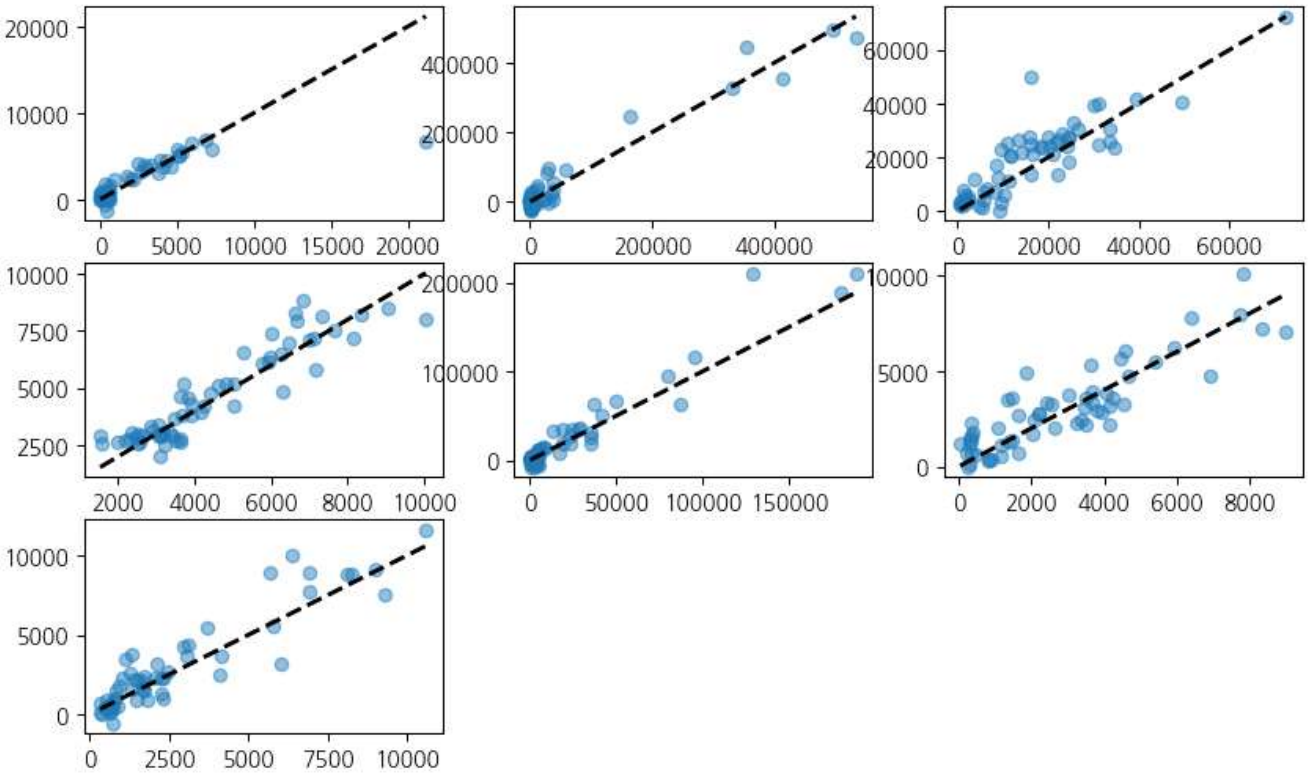


결론적으로 R^2 값이 0.865로, 국내외 주요광물의 수입과 수출의 중량과 금액을 반영하는 모델보다 국내의 주요광물의 수입과 수출의 중량과 금액만을 반영한 모델이 국내 산화니켈의 수입금액의 변동성을 매우 잘 설명하고 있는 것을 알 수 있습니다.

2) 다변량 회귀분석(MultiOutput) 모형

위에서 분석했었던 데이터와 동일한 데이터로 분석하되 이번에는 종속변수를 개별적으로 하나씩 투입하여 분석하는 것이 아닌 국내 주요광물의 수입금액 데이터(7개)를 가지고 모두 종속변수로 두고 나머지 49개의 요인을 투입변수로 하여 다변량 회귀 MultiOutput 모델을 구현하였습니다.

다음은 각 종속변수인 국내 주요광물의 수입금액 데이터에 대한 모델의 분석결과에 대한 산점도 및 추세선을 시각화한 그래프입니다.



위에서부터 왼쪽에서 오른쪽까지 차례대로 국내의 산화니켈, 산화리튬, 산화코발트, 이산화망간, 탄산리튬, 황산니켈, 황산코발트 이며 대체적으로 모델의 실제치(산포도)와 예상치(점선)의 차이가 크지않아 예측치가 매우 높은 것을 확인할 수 있습니다.
모델의 성능 평가 결과는 다음과 같습니다.

MSE: [1.63162084e+06 1.85806627e+08 6.39224091e+07 2.66861101e+05
4.99503315e+08 6.78700179e+05 4.17693852e+06], R2: 0.9535266553086184
MSE (PCA): [3.96846083e+06 7.47982796e+08 5.88207020e+07 6.29722159e+05
1.97349511e+08 1.25236745e+06 1.29380201e+06], R2 (PCA): 0.9378317303278392

다변량 회귀 모델과 PCA 적용 모델의 결과는 다음과 같습니다.

각 모델의 MSE 값을 비교해보면 다음과 같습니다:

국내산화니켈.수입금액:
일반 모델: 1,631,620.84
PCA 모델: 3,968,460.83
일반 모델이 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

국내산화리튬.수입금액:
일반 모델: 185,806,627.00
PCA 모델: 747,982,796.00
일반 모델이 훨씬 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

국내산화코발트.수입금액:
일반 모델: 63,922,409.10
PCA 모델: 58,820,702.00
PCA 모델이 약간 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

국내이산화망간.수입금액:
일반 모델: 266,861.10
PCA 모델: 629,722.16
일반 모델이 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

국내탄산리튬.수입금액:
일반 모델: 499,503,315.00
PCA 모델: 197,349,511.00
PCA 모델이 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

국내황산니켈.수입금액:
일반 모델: 678,700.18
PCA 모델: 1,252,367.45
일반 모델이 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

국내황산코발트.수입금액:

일반 모델: 4,176,938.52

PCA 모델: 1,293,802.01

PCA 모델이 더 낮은 MSE 값을 가지며, 예측 정확도가 높습니다.

R^2 (R-squared)는 0.954로, 모델이 약 95.4%의 변동성을 설명할 수 있습니다.

R^2 (PCA)는 0.938로, 모델이 약 93.8%의 변동성을 설명할 수 있습니다.

일반 모델에서는 대부분의 종속 변수에서 MSE 값이 낮아 예측 정확도가 높습니다.

R^2 값이 0.954로, 매우 높은 설명력을 가집니다.

PCA 모델에서는 일부 종속 변수(국내산화코발트.수입금액, 국내탄산리튬.수입금액, 국내황산코발트.수입금액)에서 MSE 값이 더 낮아 예측 정확도가 개선되었습니다.

R^2 값이 0.938로, 여전히 높은 설명력을 가지지만 일반 모델보다는 낮습니다.

결론적으로 일반 모델은 전체적으로 더 높은 예측 정확도를 가지고 있으며, 대부분의 종속 변수에서 낮은 MSE 값을 가집니다.

PCA 모델은 일부 종속 변수에서 더 낮은 MSE 값을 보여주며, 차원 축소를 통해 데이터의 복잡성을 줄였지만, 전체적인 예측 정확도는 다소 낮아졌습니다.

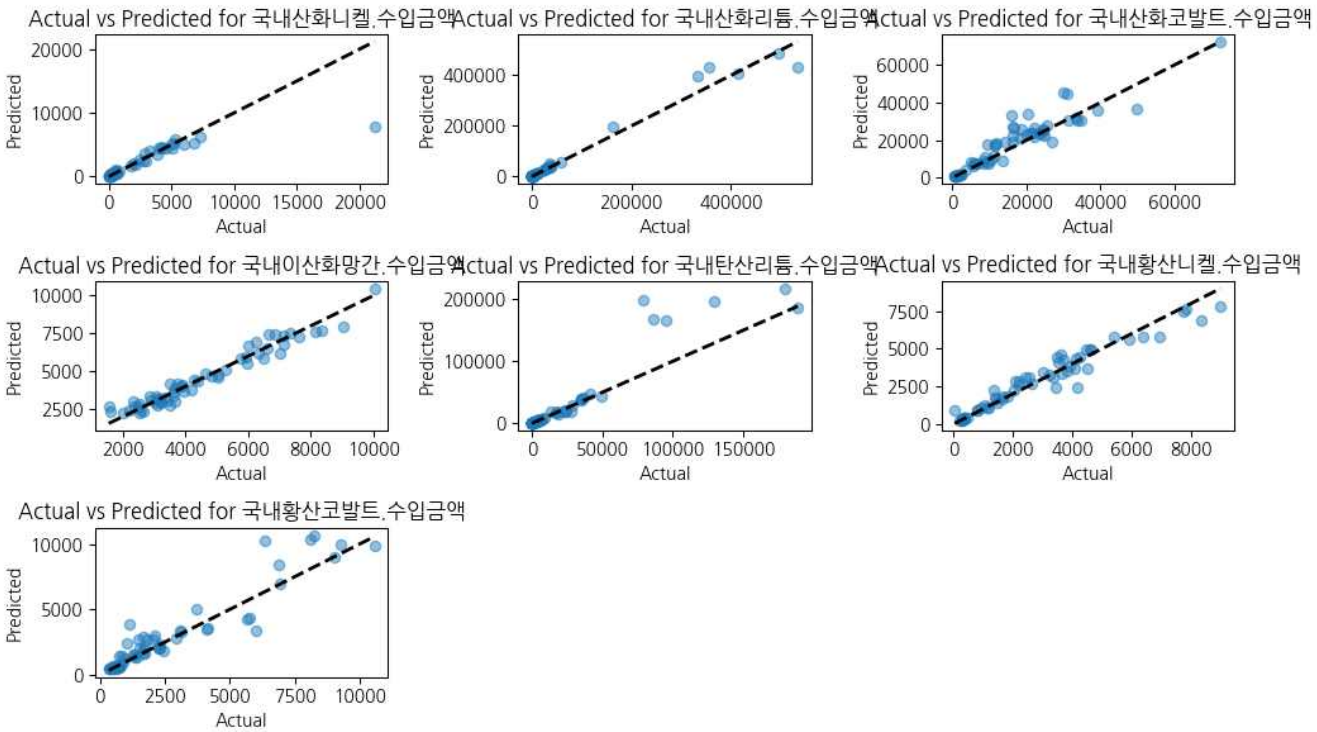
두 모델 모두 높은 설명력을 가지고 있으며, 특정 상황에서는 차원 축소를 통해 모델의 성능을 개선할 수 있는 가능성이 있습니다.

하지만 MSE값으로만 모든 모델의 검정력을 판단할 수 없고 변수가 많아지면 R^2의 값이 증가하는 기술통계 평가의 성질을 고려할 때 개별적인 종속변수들에 대한 다변량 회귀 모델 분석과 MultiOutput 모델과는 특정한 의사결정 수립에 있어서 두 모델의 분석이 같이 병행 되어야 한다고 생각되어집니다.

3) 다변량 회귀 MultiOutput_비선형 모형

비선형 모델에는 각각 Random Forest Regression, SVR, Gradient Boosting Regressor, Neural Network 모형들이 존재하며 위에서 진행 하였던 동일한 데이터로 분석을 실시하였습니다.

(1) Random Forest Regressor 모형



MSE: [3.20140677e+06 3.77331763e+08 3.00321020e+07 1.97632153e+05
5.40289540e+08 3.00132452e+05 1.01881080e+06], R2: 0.9414541325459412

MSE (Mean Squared Error)

국내산화니켈.수입금액: 3,201,406.77

국내산화리튬.수입금액: 377,331,763.00

국내산화코발트.수입금액: 30,032,102.00

국내이산화망간.수입금액: 197,632.15

국내탄산리튬.수입금액: 540,289,540.00

국내황산니켈.수입금액: 300,132.45

국내황산코발트.수입금액: 1,018,810.80

국내이산화망간.수입금액의 MSE가 197,632.15로 가장 낮고, 국내탄산리튬.수입금액의 MSE가 540,289,540.00로 가장 높습니다.

이는 모델이 국내이산화망간.수입금액을 비교적 정확히 예측했지만, 국내탄산리튬.수입금액은 예측 정확도가 떨어짐을 나타냅니다.

R2 (R-squared)

전체 모델의 R2 값은 0.9414541325459412입니다.

이는 모델이 전체 데이터 변동의 약 94.15%를 설명할 수 있음을 의미합니다.

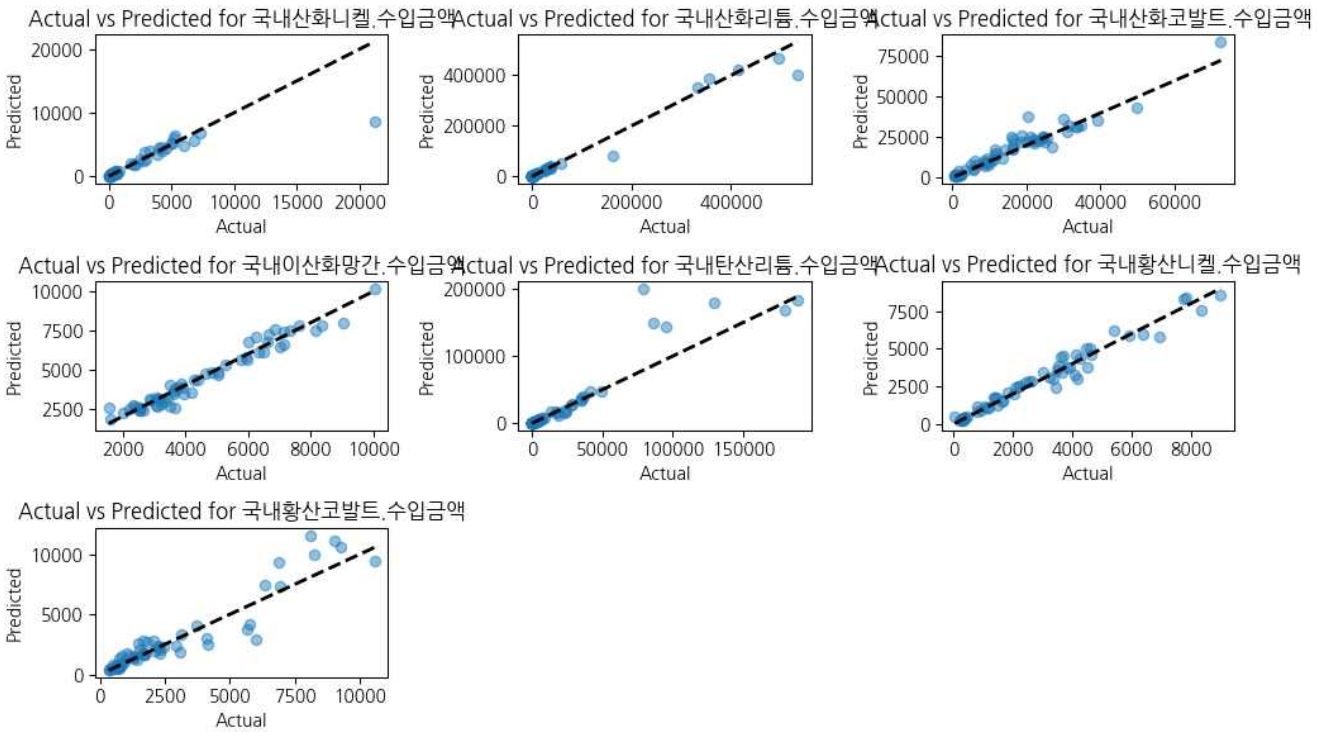
즉, 모델의 성능이 매우 높다고 할 수 있습니다.

모델은 전반적으로 높은 R2 값을 보이며, 다수의 종속 변수에 대해 낮은 MSE 값을 가지고 있어 높은 예측 정확도를 가집니다.

하지만 일부 종속 변수, 예를 들어 국내탄산리튬.수입금액의 경우 MSE가 높아 개선의 여지가 있습니다.

모델 성능을 더 높이기 위해선 데이터 전처리나 모델 튜닝 등의 추가 작업이 필요할 수 있습니다.

(2) Gradient Boosting Regressor 모형



MSE: [2.78994801e+06 4.55180489e+08 1.71052485e+07 1.86467023e+05
4.06935624e+08 1.93927183e+05 1.00078494e+06], R2: 0.9456945260453631

MSE (Mean Squared Error)

국내산화니켈 수입금액: 2,789,948.01

국내산화리튬 수입금액: 455,180,489.00

국내산화코발트 수입금액: 17,105,248.50

국내이산화망간 수입금액: 186,467.02

국내탄산리튬 수입금액: 406,935,624.00

국내황산니켈 수입금액: 193,927.18

국내황산코발트 수입금액: 1,000,784.94

국내산화리튬 수입금액과 국내탄산리튬 수입금액의 MSE 값이 매우 큽니다. 이는 해당 예측에서 큰 오차가 발생했음을 의미합니다.

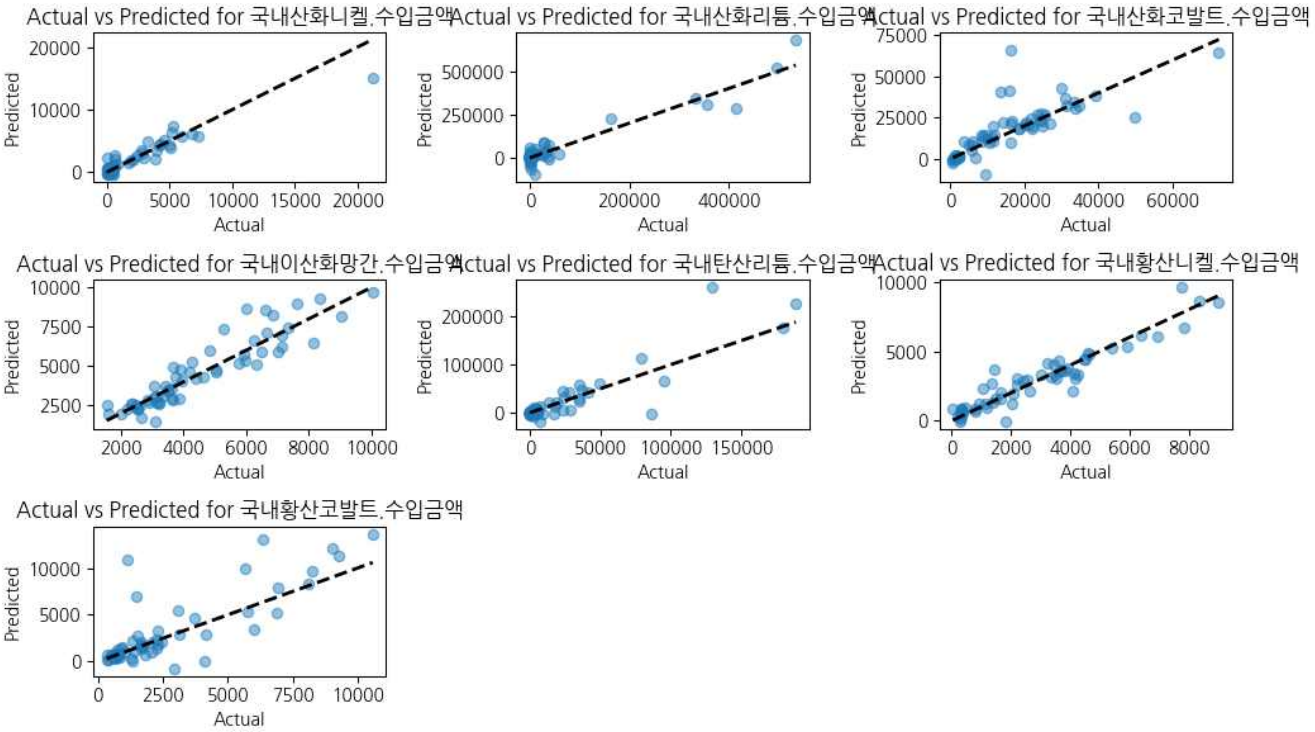
R^2 (R-squared)

0.9456945260453631 값은 매우 높은 값으로, 모델이 대부분의 변동성을 잘 설명하고 있음을 나타냅니다.

즉, 모델의 예측이 실제 데이터와 상당히 일치함을 의미합니다.

모델의 전반적인 성능은 R^2 값이 높아 매우 좋습니다. 하지만 국내산화리튬 수입금액과 국내탄산리튬 수입금액의 예측 성능을 높이기 위한 추가적인 모델 개선이 필요할 수 있습니다.

(3) Neural Network 모형



MSE: [1.23519047e+06 1.48411656e+09 9.62115784e+07 7.30371059e+05
5.66494687e+08 5.40920638e+05 4.66583944e+06], R2: 0.8675857798617864

MSE (Mean Squared Error)
국내산화니켈.수입금액: 1,235,190.47
국내산화리튬.수입금액: 1,484,116,560.00
국내산화코발트.수입금액: 96,211,578.40
국내이산화망간.수입금액: 730,371.06
국내탄산리튬.수입금액: 566,494,687.00
국내황산니켈.수입금액: 540,920.64
국내황산코발트.수입금액: 4,665,839.44
국내산화리튬.수입금액과 국내탄산리튬.수입금액의 MSE 값이 상대적으로 높습니다.
이는 모델이 이 두 변수의 예측에서 큰 오차를 발생시키고 있음을 나타냅니다.

R2 (R-squared)
R2 값이 0.8676인 것은 모델이 전체 데이터의 86.76%를 설명하고 있음을 나타냅니다.

리튬 관련 변수에서 높은 MSE가 관찰되어 개선이 필요하지만, 전체적으로 높은 R2값을 보이며 비교적 성능이 좋다고 할 수 있습니다.

결론적으로 다변량 회귀 MultiOutput_비선형 모형들의 전체적인 성능은 대체적으로 매우 높았으며, 공통적으로 탄산리튬 등에서 MSE값이 높은 것이 확인되었습니다. 그만큼 리튬의 가격이 비정상적으로 변동했다는 것을 의미하고 대부분의 모형에서 예측이 어려웠음을 의미합니다. 하지만 나머지 광물에 대해서는 높은 예측성을 보였습니다. 따라서 리튬의 가격에 대해서는 SingleOutput 모형방식의 분석과 더불어서 보조적인 방법들인 위험성 평가모델(GARCH모형, Random forest 모형 등)을 병행하여 활용, 예측하는 것이 필요하다고 생각되어 집니다.

(6) 이상치 탐지 모형(Abnormaly Detection Model)

이상치 탐지 모형은 주어진 데이터에서 머신러닝과 트리분할, 벡터분류, 유클리드 거리 계산 등의 여러 가지 방법을 활용하여 이상치를 찾아내는 모형을 의미합니다. 모델로는 Isolation Forest, One-Class SVM, Local Outlier Factor, Elliptic Envelope 등이 있습니다.
여기에서는 Isolation Forest와 시계열에서 이상치를 탐지하는 모델인 Prophet 모형을 적용하였습니다.

1) Isolation Forest 모델

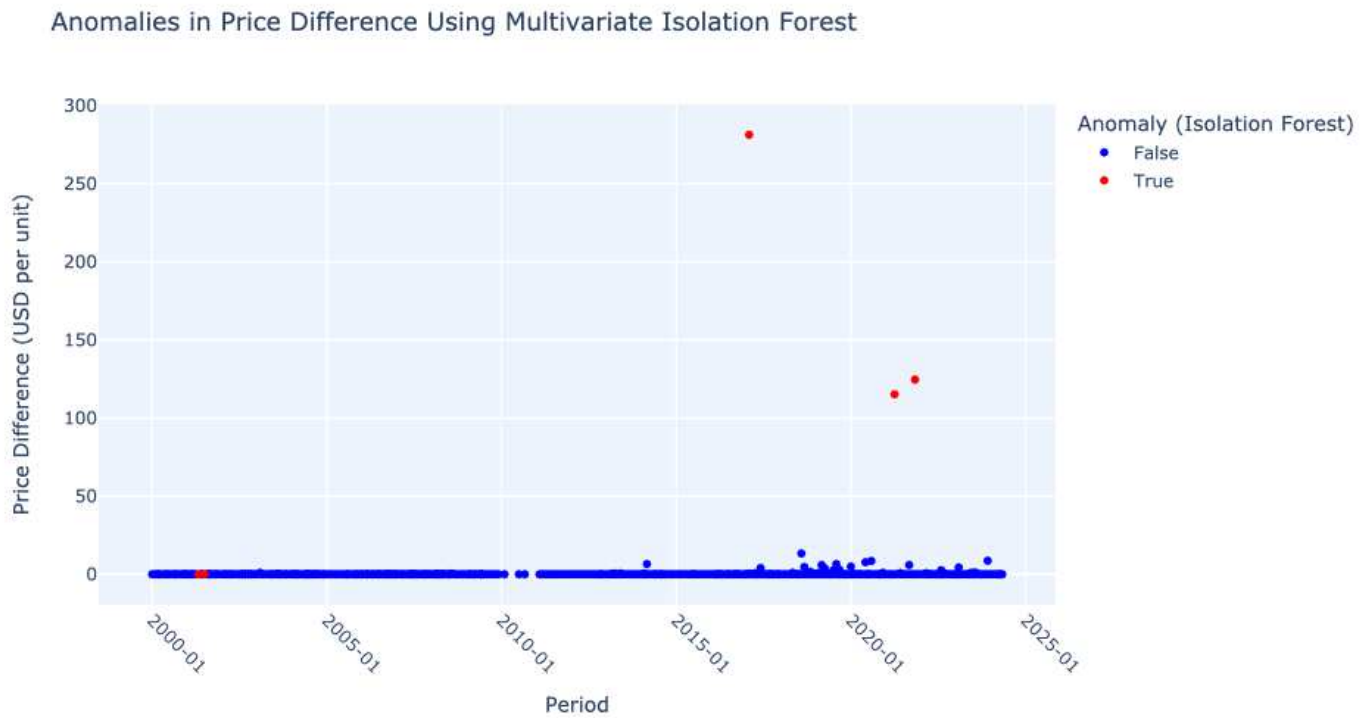
a. 사전 예측 방식

데이터에 isolation forest 모델을 적용하기에 앞서, UNcomtrade 의 이산화망간 데이터로 파일럿 분석을 진행하였습니다.

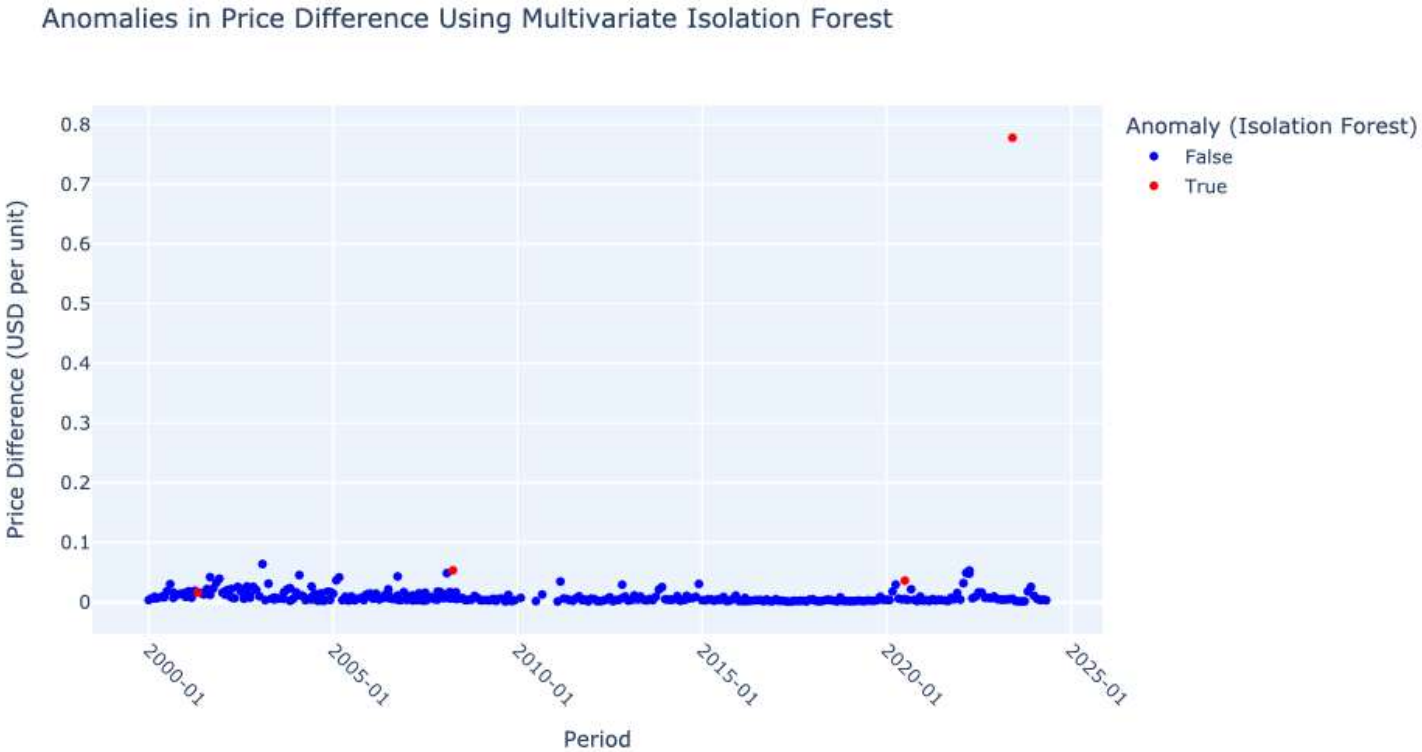
Isolation Forest 모델에서 사용된 특징(feature)은 다음과 같습니다:
1. qty (수량)
2. cifvalue (운임 및 보험료 포함 가격)
3. fobvalue (운임 및 보험료 제외 가격)
4. price_difference (가격 차이)
추가적으로, reporterDesc 열은 더미 변수로 변환되어 모델에 포함되었습니다.

Isolation Forest 모델은 데이터 포인트가 나머지 데이터와 얼마나 다른지를 측정하여 이상치를 감지합니다. 이를 통해 데이터 내의 비정상적이거나 예외적인 패턴을 식별할 수 있습니다.
모델의 동작 원리는 데이터를 여러 번 무작위로 분할하여 각 데이터 포인트를 격리시키는 방식입니다. "격리"는 데이터 포인트를 다른 데이터와 분리하는 과정을 의미합니다.
즉, 특정 데이터 포인트가 다른 데이터 포인트들과 얼마나 쉽게 분리될 수 있는지를 측정하는 것입니다. 예를 들어 “일반적인” 데이터 포인트는 데이터셋 내의 다른 데이터 포인트들과 비슷한 특징을 가지므로, 여러 번의 분할이 필요합니다.
예를 들어, 일반적인 데이터 포인트는 여러 번의 조건을 통해서만 고유하게 식별될 수 있습니다. 하지만 이상치는 다른 데이터 포인트들과 크게 다르기 때문에, 몇 번의 분할만으로 쉽게 분리됩니다. 이는 이상치가 데이터셋 내에서 고립되어 있다는 것을 의미합니다. 예를 들어, 극단적인 값이나 드문 패턴을 가진 데이터 포인트는 빠르게 분리될 수 있습니다

따라서 몇 번의 격리 과정을 거치면서 분할 횟수가 적게 필요한 데이터 포인트일수록 이상치일 가능성이 높다고 모델은 판단하게됩니다.
이산화 망간 수입데이터를 모델로 분석했을 때 우리는 다음과 같은 결과를 얻었습니다.

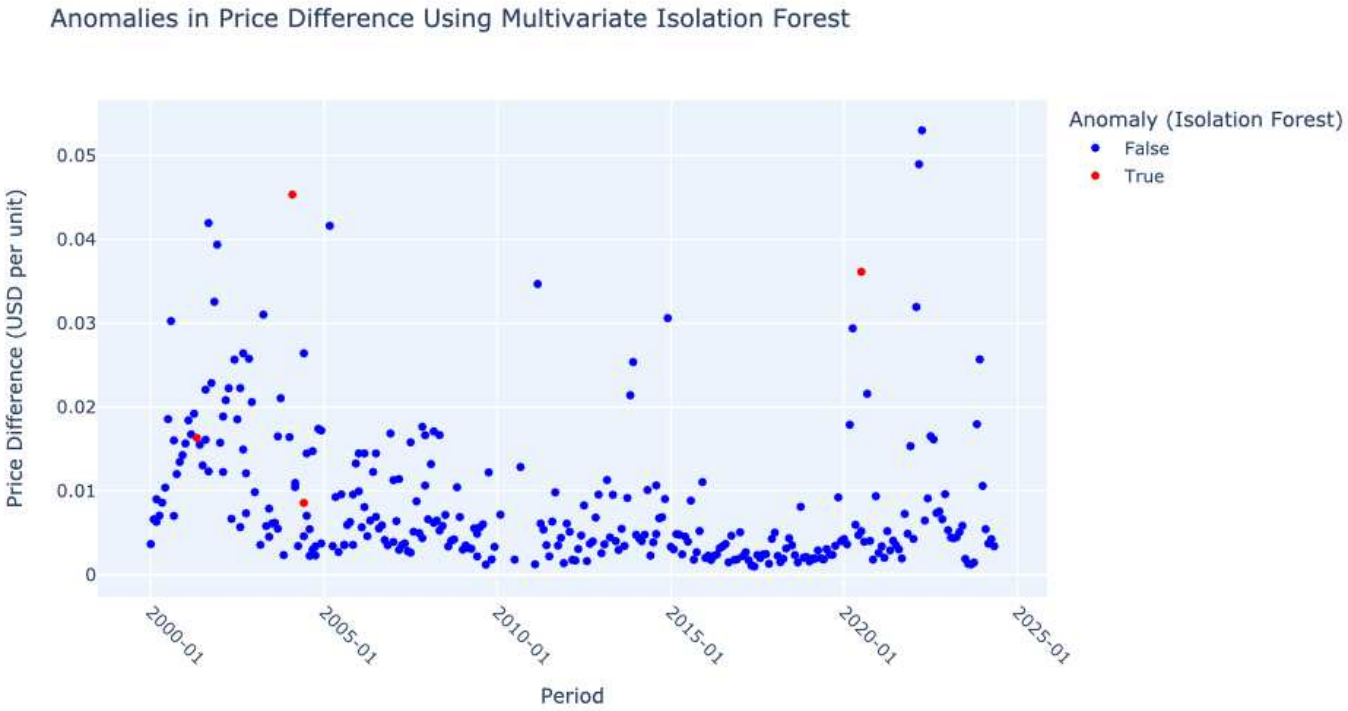


먼저 수량 기준을 0으로 설정한 데이터셋에서의 이상치 감지 결과입니다. 대부분의 데이터가 낮은 가격 차이 구간에 위치하고 있습니다. 일부 데이터 포인트가 높은 가격 차이를 나타내며 이상치로 감지되었습니다. 하지만 너무 작은 수량들이 이상치로 관측됩니다.



수량 기준을 1,000 으로 올려보았습니다. 수량 기준0 의 그래프와 유사한 패턴을 보이나, 이상치의 개수가 감소한 것을 알 수 있습니다. 이는 더 높은 수량 기준이 설정됨에 따라

이상치로 감지되는 데이터 포인트의 수가 줄어든 것을 의미합니다. 하지만 이상치가 감지된 시점은 유사합니다.



수량 기준을 10000으로 설정한 데이터셋에서의 이상치 감지 결과입니다. 수량 기준이 더 높아짐에 따라 이상치로 감지된 데이터 포인트의 수가 더욱 감소했습니다. 그러나 여전히 일부 데이터 포인트는 높은 가격 차이를 나타내며 이상치로 감지되었습니다.

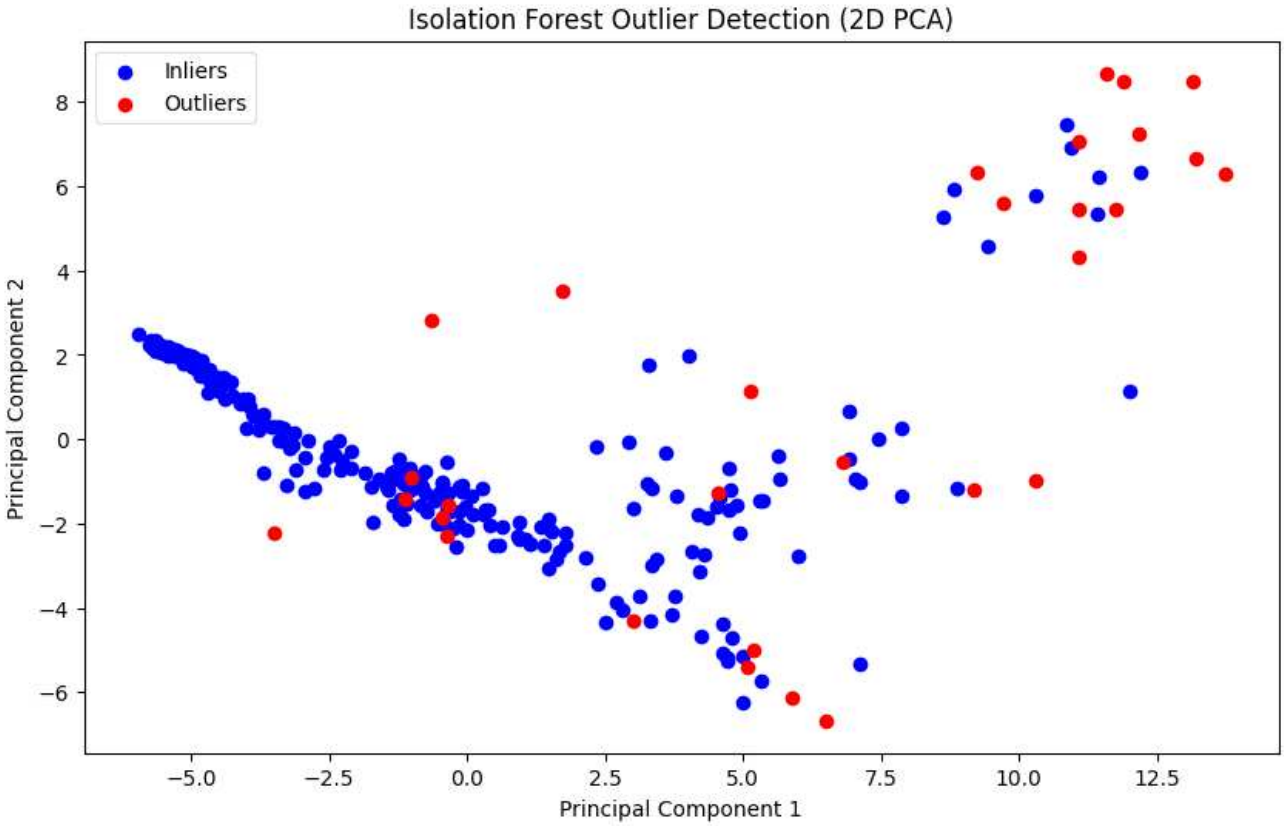
종합해보았을 때, 수량 기준이 낮을수록 더 많은 데이터 포인트가 이상치로 감지됩니다. 이는 작은 거래량에서도 큰 가격 차이가 발생할 수 있기 때문입니다. 반대로, 수량 기준이 높아질수록 이상치의 수가 줄어들며, 이는 큰 거래량에서 가격 차이가 상대적으로 적기 때문입니다.

주의할 점은 모델은 이상치의 정확한 발생 원인을 제공하지 않는다는 것입니다. 예를 들어, 특정 시점의 이상치가 시장 조건 변화 때문인지, 정책 변화 때문인지 등을 알 수 없습니다.

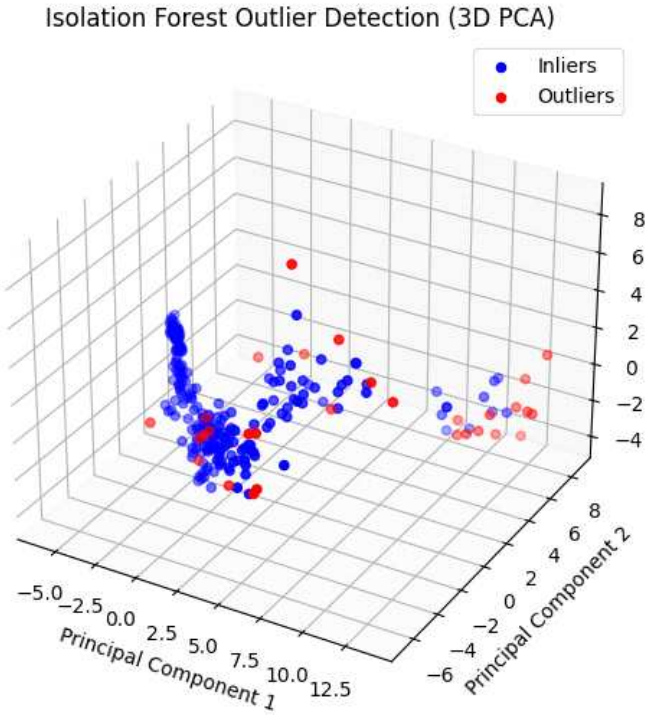
하지만 단점에도 불구하고, CIF와 FOB 가격의 차이인 해상 보험료, 운임가격 등의 선행지표로 이상치를 사전에 찾을 수 있어, 다른 이상치 모델처럼 특별한 가공이 필요하지 않습니다. 따라서 주요광물의 공급망 리스크를 판단하는 주요 선행지표로서 활용가능하다고 생각합니다.

b. 사후 검증 방식

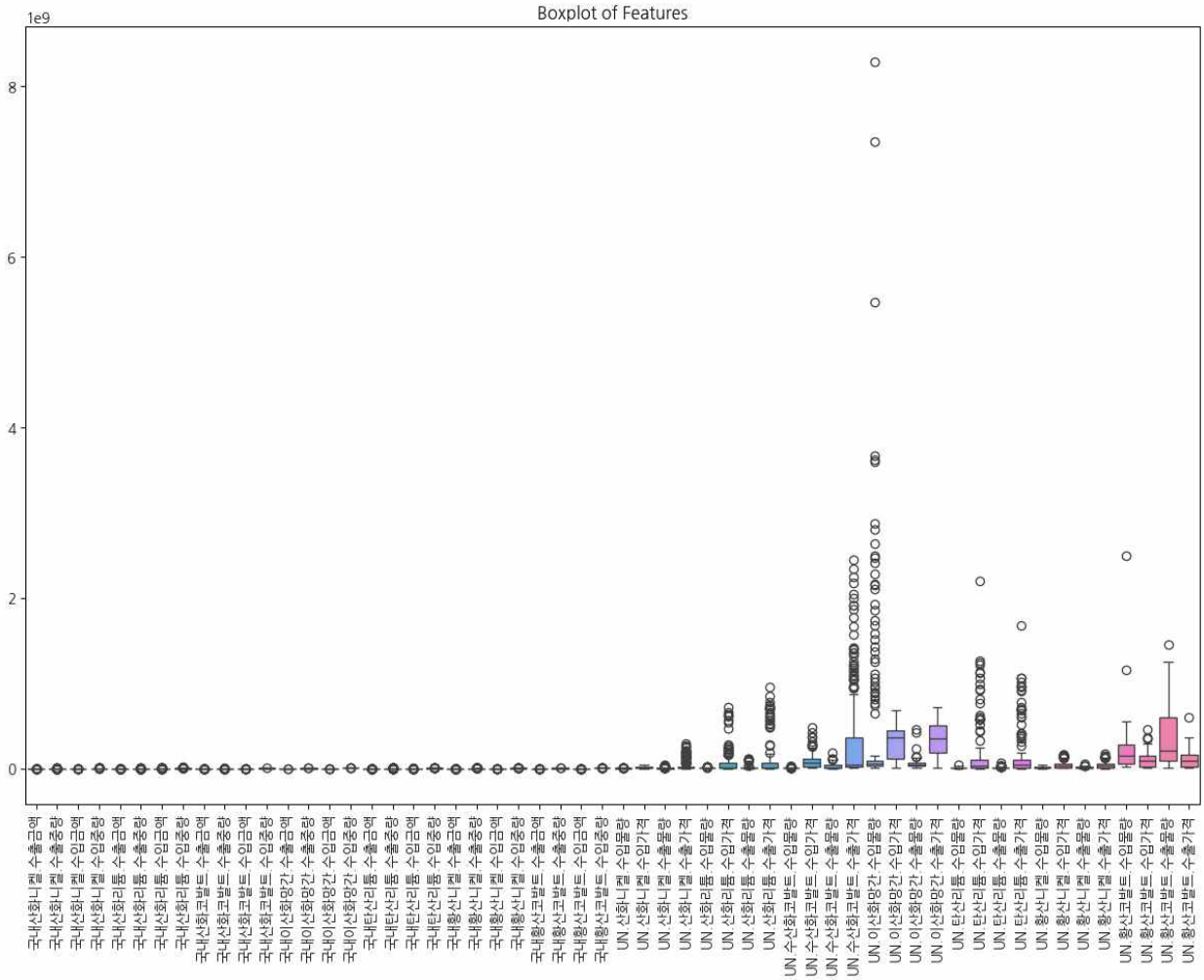
데이터는 앞에서 활용하였던, UNcomtrade(https://comtradeplus.un.org/)의 전세계 주요광물(니켈과 리튬, 망간, 텅스텐, 코발트)의 수입중량과, 수입금액, 수출중량과, 수출금액과 한국무역협회(https://stat.kita.net/newMain.screen)의 주요광물의 국내 수입금액과 수입중량, 국내 수출금액과 수출중량의 데이터를 활용하였습니다. Isolation 모델을 적용하기 앞서서 데이터를 PCA 처리하였습니다. (주성분 15개), 그리고 전체 데이터에서 이상치를 찾아내어 'isolated_outliers.csv' 파일로 그 목록을 저장 하였습니다.
그리고 이상치를 확인하기 위해 PCA 성분1과 성분2 에 대하여 시각화를 하였습니다.



다음은 성분1과 성분2와 성분3에 대하여 시각화를 하였습니다.



원래 전체 요소에 대한 이상치를 확인하기 위하여 boxplot을 활용하였습니다.

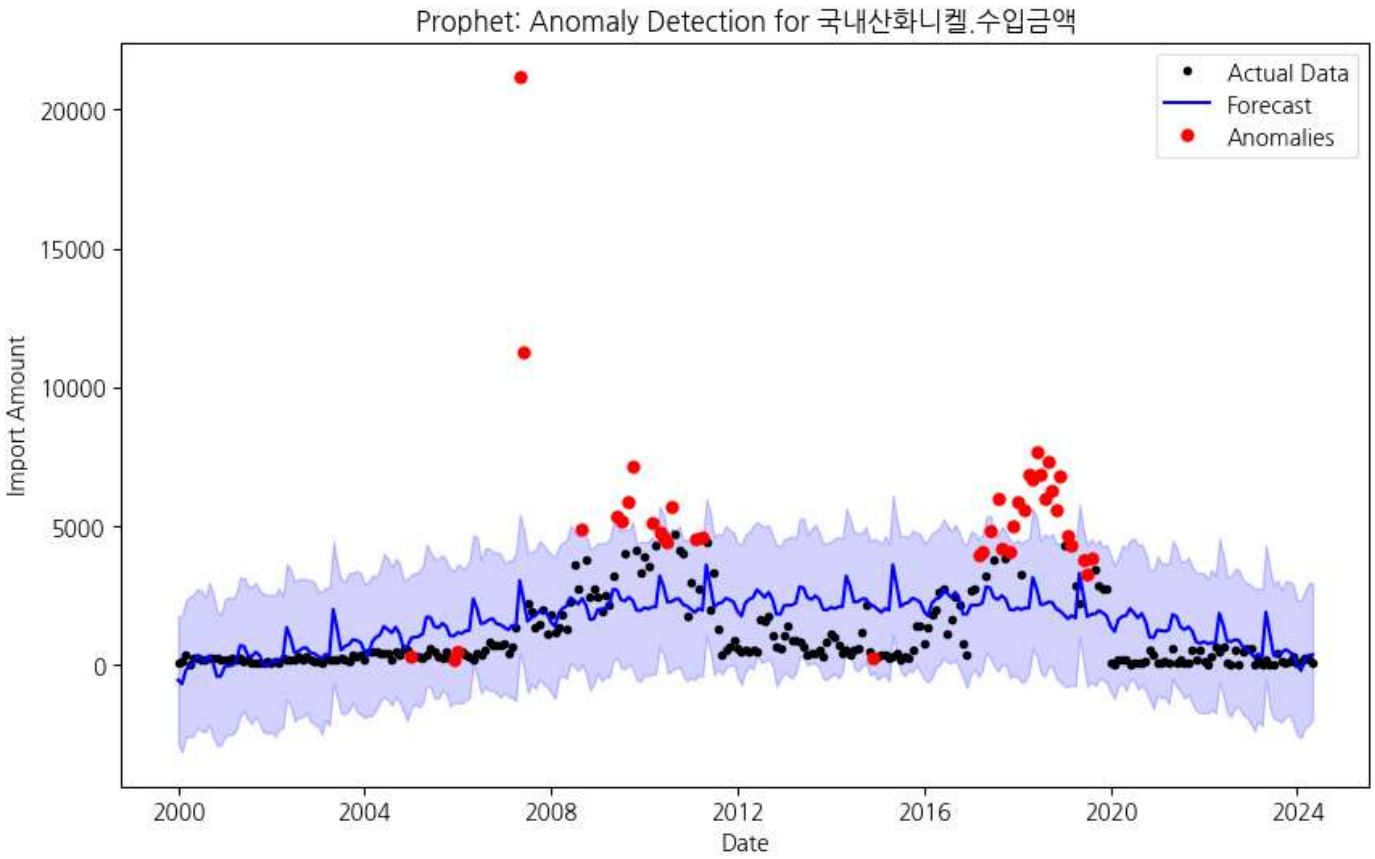


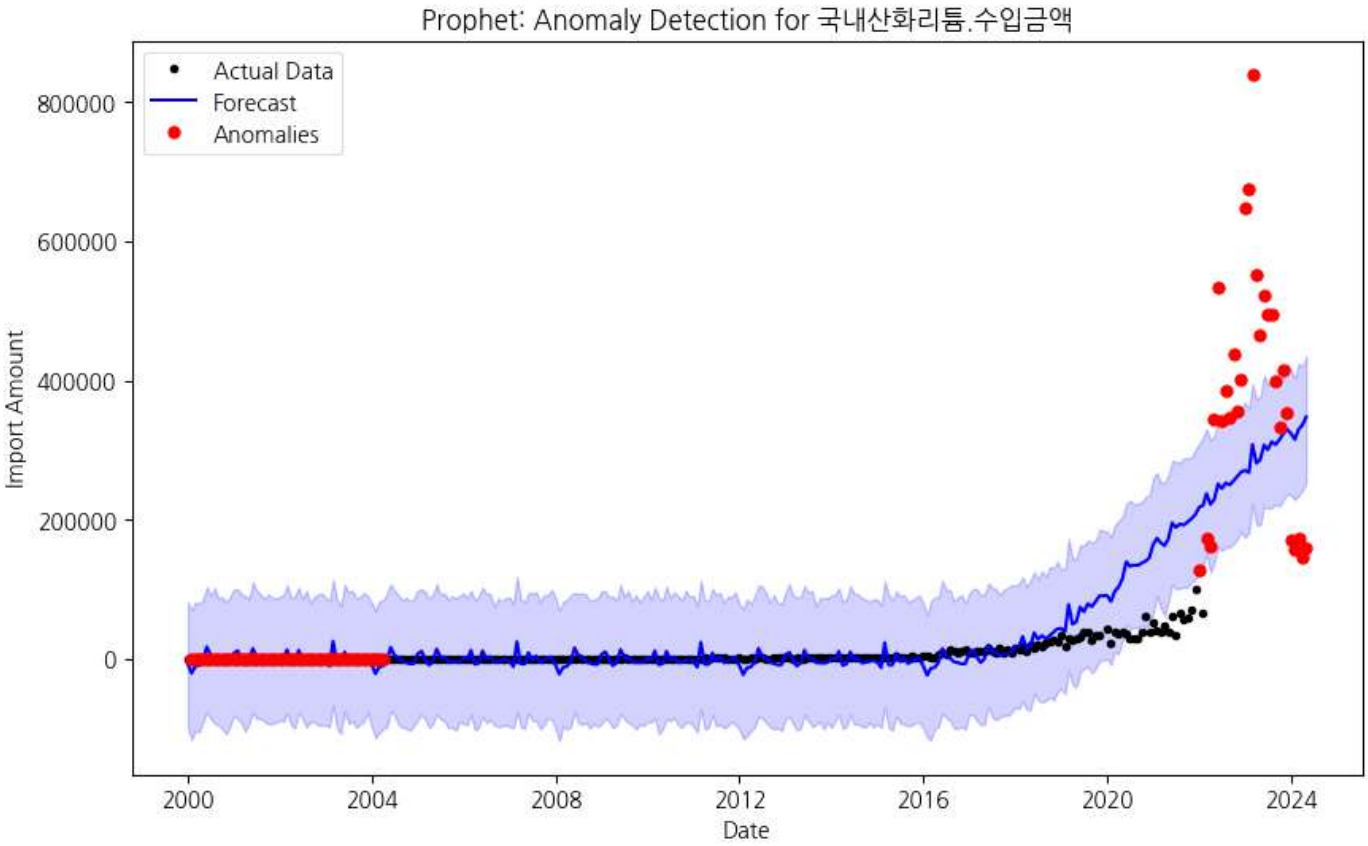
위의 데이터를 보면 UN 이산화망간의 수입물량과, UN 탄산리튬, UN 황산코발트 수입가격 등에서 이상치가 나온 것을 확인할 수 있습니다.

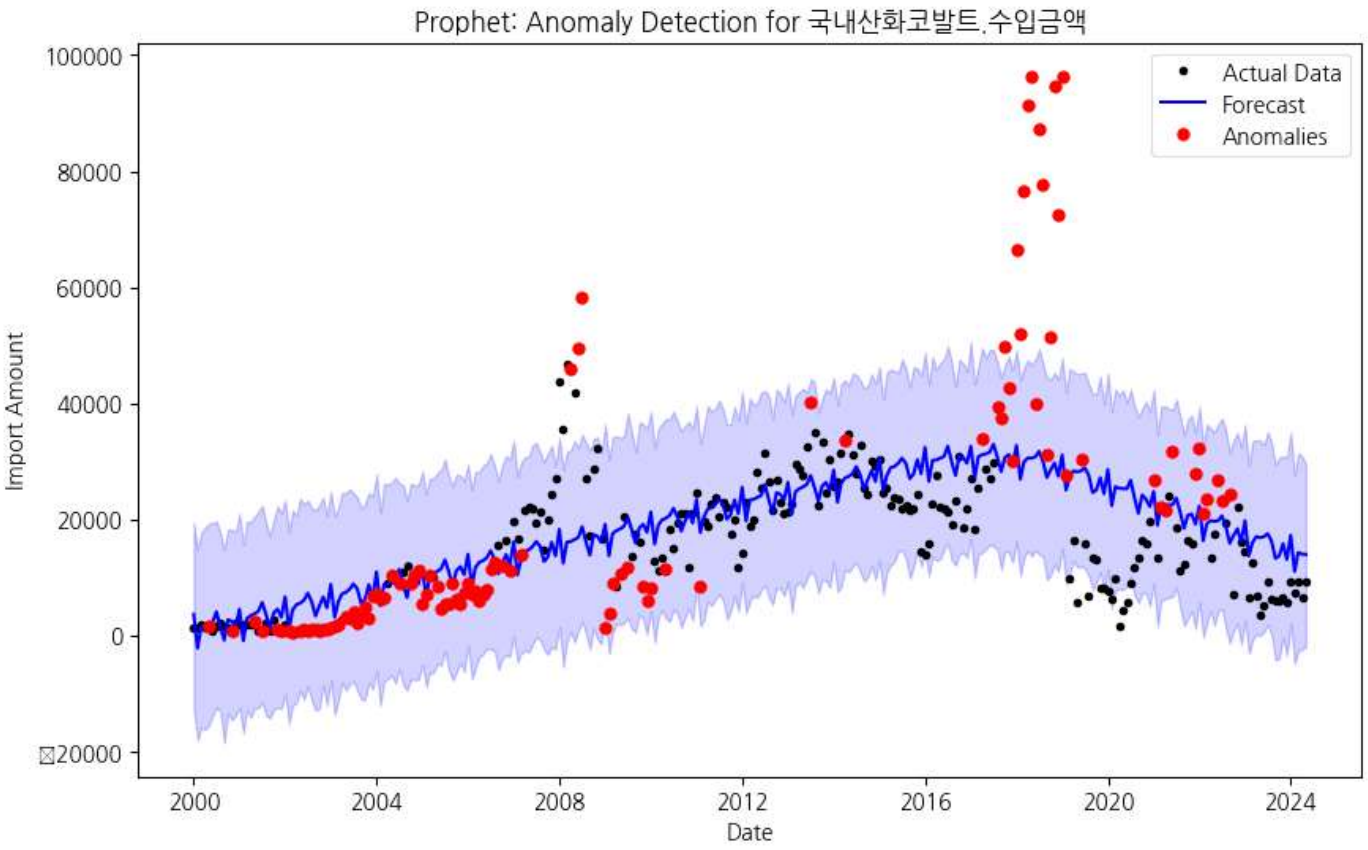
PCA처리된 데이터에서 이상치를 확인하기 위해서 모델을 돌린후 원래의 데이터로 변환하여 목록을 다운받았으며 약 30개의 데이터가 이상치로 발견되었습니다.
시간 순서대로의 총 293행의 데이터들에서 30개의 이상치를 발견하였으며 모두 전체 시계열 중에서 특정 시점(월)에서 거래들의 이상치가 존재한다는 의미입니다. 따라서 사후적으로 데이터를 활용할 수 있으며, 혹은 위에서 설명한 금가격 변동성 예측모형(Random Forest)처럼 예측에 대한 데이터가 충분하다면 이상 거래를 발견하는 방식으로 활용할 가능성이 있습니다.

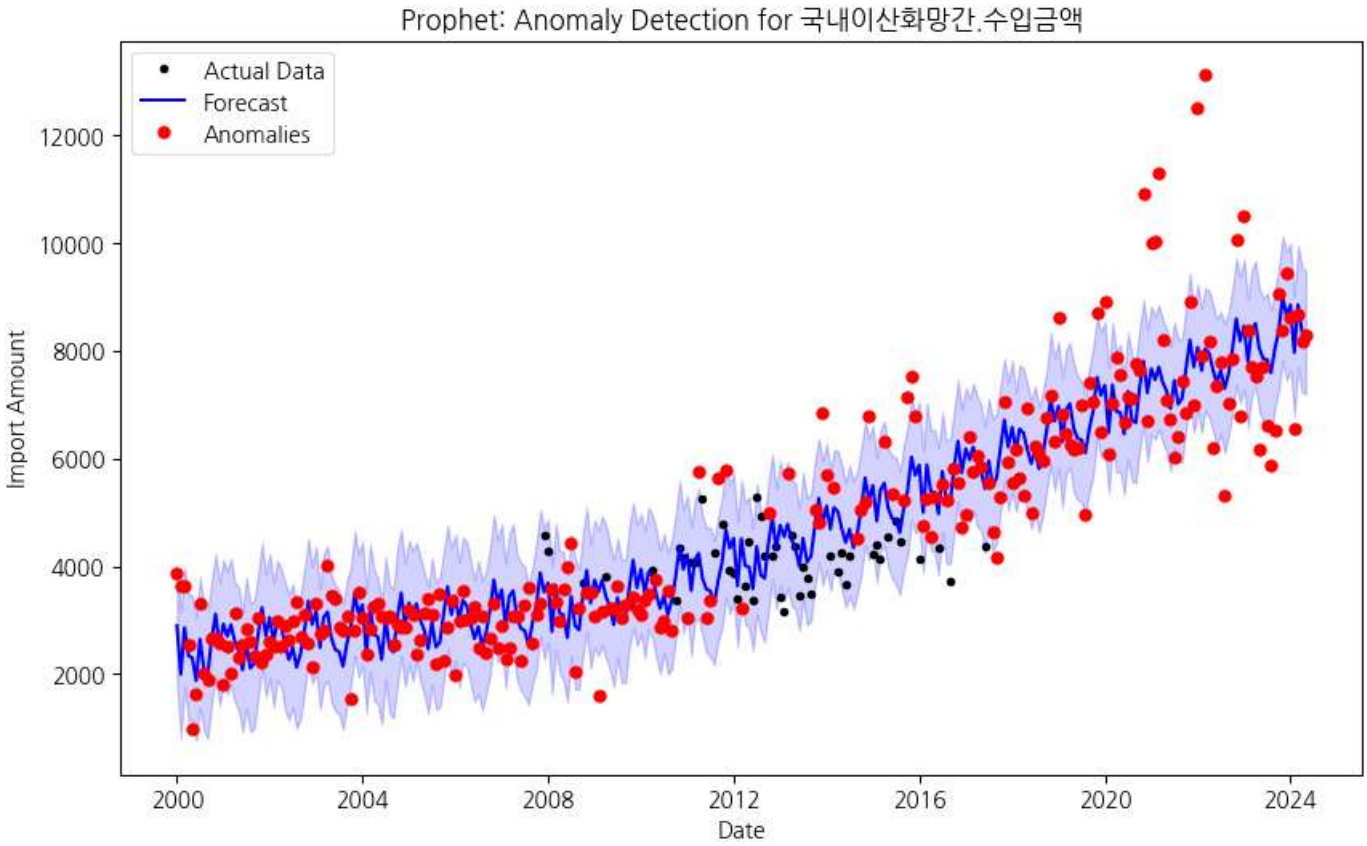
2) Prophet 모형

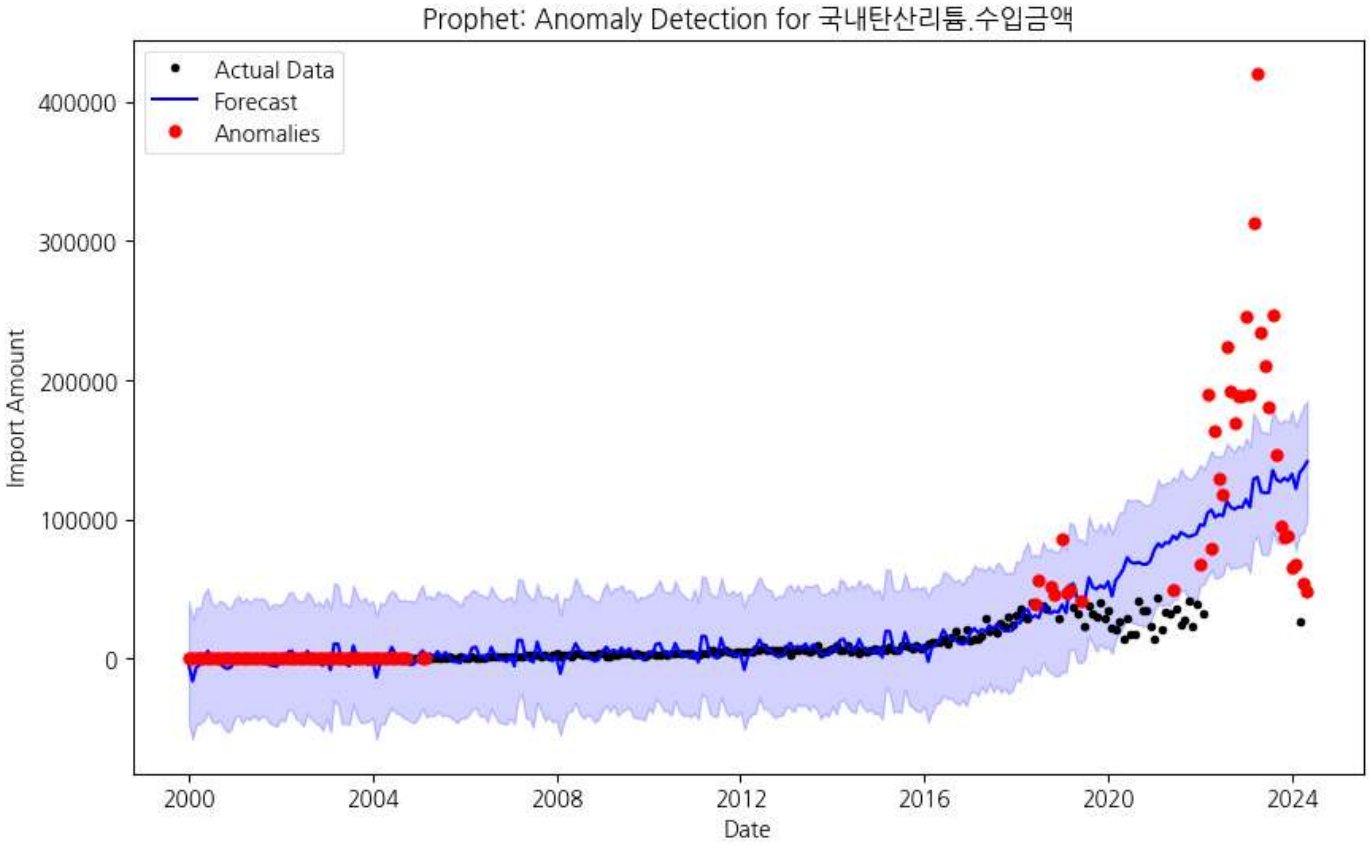
Prophet은 페이스북이 개발한 시계열 데이터 예측을 위한 오픈소스 라이브러리로 비정상적이고 변동성이 큰 시계열 데이터를 효율적으로 예측할 수 있습니다. 위에서 코미스(https://www.komis.or.kr)의 리튬 가격 데이터를 사용하여 이상치를 찾는 시계열 모형으로 설명하였고 추가적인 분석을 위하여 한국무역협회(https://stat.kita.net/newMain.screen)의 국내 주요광물(7종)의 수출금액 데이터를 바탕으로 추가 분석을 진행하였습니다.

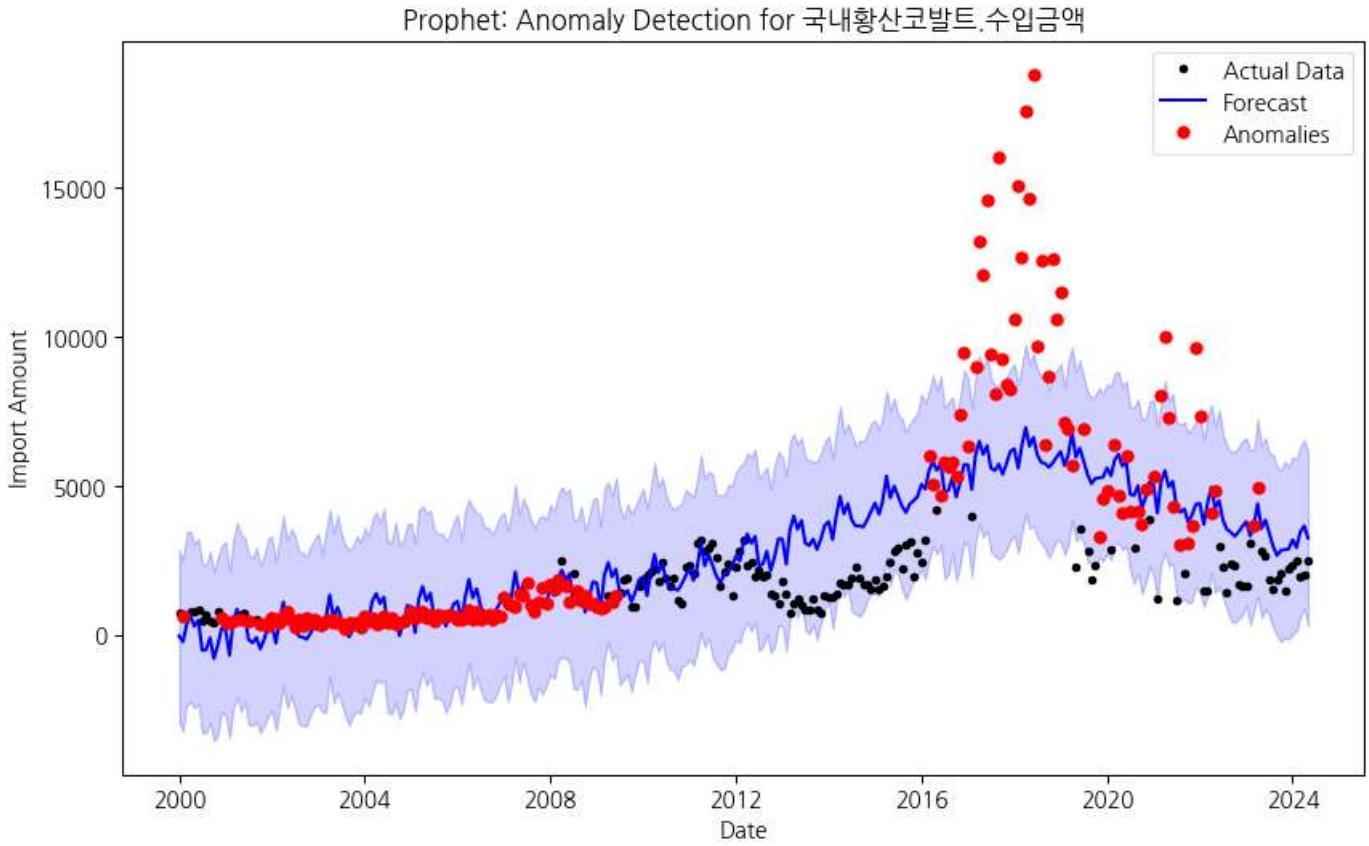


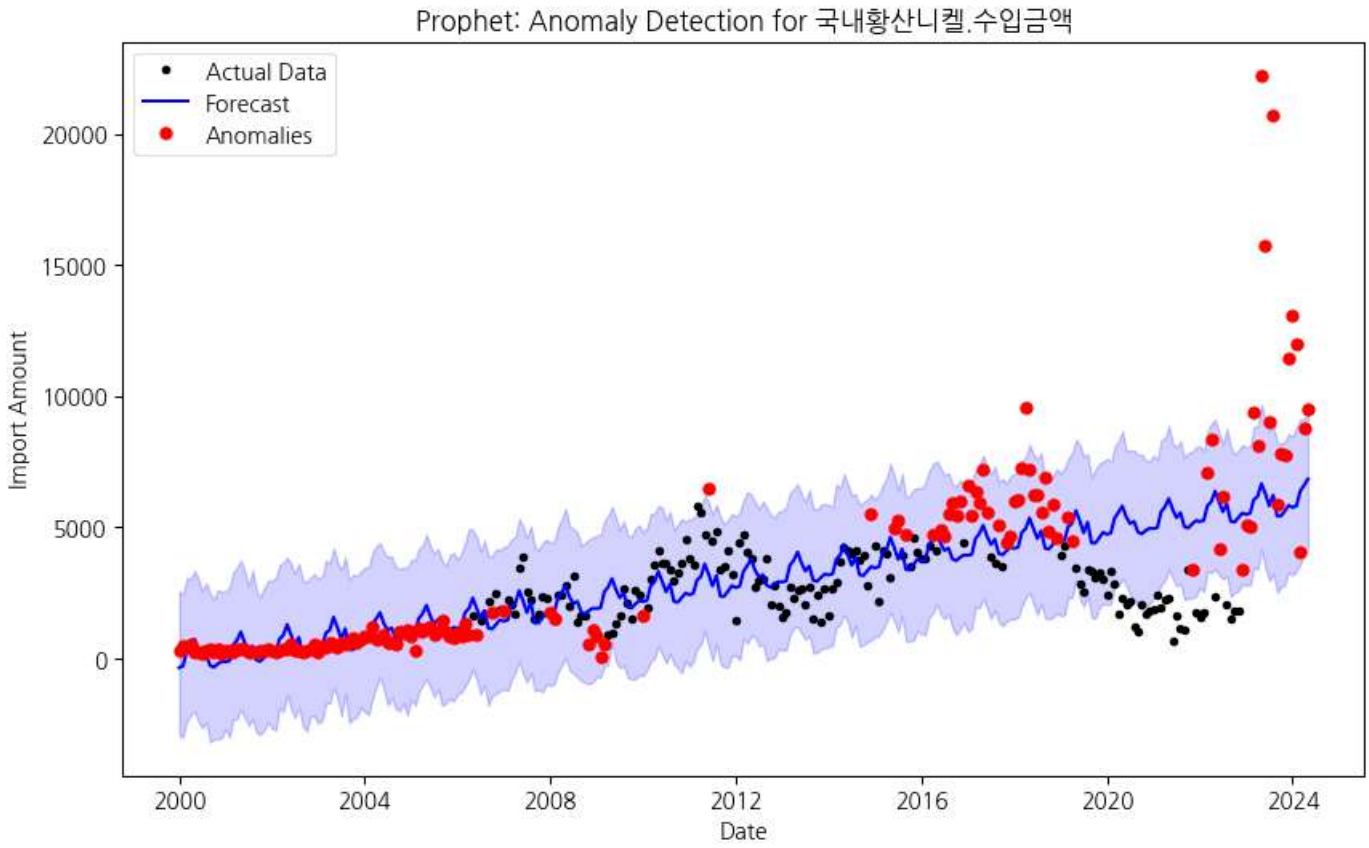

Prophet: Anomaly Detection for 국내산화리튬.수입금액
Actual Data
Forecast
Anomalies
Import Amount
800000
600000
400000
200000
0
2000
2004
2008
2012
2016
2020
2024
Date


Prophet: Anomaly Detection for 국내산화코발트.수입금액
Actual Data
Forecast
Anomalies
Import Amount
100000
80000
60000
40000
20000
0
2000
2004
2008
2012
2016
2020
2024
Date


Prophet: Anomaly Detection for 국내이산화망간.수입금액
Actual Data
Forecast
Anomalies
Import Amount
12000
10000
8000
6000
4000
2000
2000
2004
2008
2012
2016
2020
2024
Date


Prophet: Anomaly Detection for 국내탄산리튬.수입금액
Actual Data
Forecast
Anomalies
Import Amount
400000
300000
200000
100000
0
2000
2004
2008
2012
2016
2020
2024
Date


Prophet: Anomaly Detection for 국내황산코발트.수입금액
Actual Data
Forecast
Anomalies
Import Amount
15000
10000
5000
0
2000
2004
2008
2012
2016
2020
2024
Date


Prophet: Anomaly Detection for 국내황산니켈.수입금액
Actual Data
Forecast
Anomalies
Import Amount
20000
15000
10000
5000
0
2000
2004
2008
2012
2016
2020
2024
Date

이러한 방식으로 각각의 국내 수입금액에 대한 이상치 데이터를 찾아 낼수 있으며 가격에 대한 예측또한 가능합니다.

Prophet 모델은 추세와 계절성, 휴일 및 이벤트를 수식으로 반영하고 있으며, 베이지안 시계열 기반으로 오차를 추정합니다.

ARIMA와 다르게 비정상성 데이터도 비교적 예측 성능이 높은 것으로 알려져 있습니다.

하지만 Prophet 모델 조차도 예측하지 못하는 Anomalies에 대한 예측은 위에서 언급하였던 모델들과 분석 방법을 병행하여 분석해야 될 것으로 생각됩니다.

(7) 앙상블 모형

앙상블 모형이란 한가지 이상의 머신러닝 모델을 결합하여 하나의 모델보다 더 나은 예측 성능을 얻는 방법입니다. 앙상블 기법은 배깅[9], 부스팅[10], 스태킹[11], 심플 앙상블[12] 등이 있으며 여기서는 심플 앙상블의 형태로 LSTM과 Random Forest을 합친 앙상블 모델을 살펴보려합니다.

데이터는 앞에서 활용하였던, UNcomtrade(https://comtradeplus.un.org/)의 전세계 주요광물(니켈과 리튬, 망간, 텅스텐, 코발트)의 수입중량과, 수입금액, 수출중량과, 수출금액과 한국무역협회(https://stat.kita.net/newMain.screen)의 주요광물의 국내 수입금액과 수입중량, 국내 수출금액과 수출중량의 데이터를 활용하였습니다.

LSTM과 Random Forest 모형 모두 종속변수인 국내산화리튬의 수입가격을 예측하기 위하여 모형을 구성하였습니다. LSTM은 시계열 데이터를 바탕으로 예측값을 생성하였고, Random Forest 는 각 시점의 독립 변수(국내산화리튬의 수입가격을 제외한 모든 변수)를 사용하여 타겟값을 예측하였고 두 예측값의 평균울 내어 최종 예측값으로 산정하였습니다.

결과는 아래와 같습니다.

```
Ensemble Model MSE: 0.003750195642395263
Ensemble Model R2 Score: 0.7374566197795465
```

이 모델의 R2 값은 약 0.737로, 이는 모델이 약 73.7%의 분산을 설명할 수 있음을 의미합니다.

R2 값이 0.7 이상이면 모델이 데이터를 상당히 잘 설명한다고 볼 수 있습니다. 따라서 이 모델은 매우 양호한 예측 성능을 가지고 있다고 해석할 수 있습니다.

---

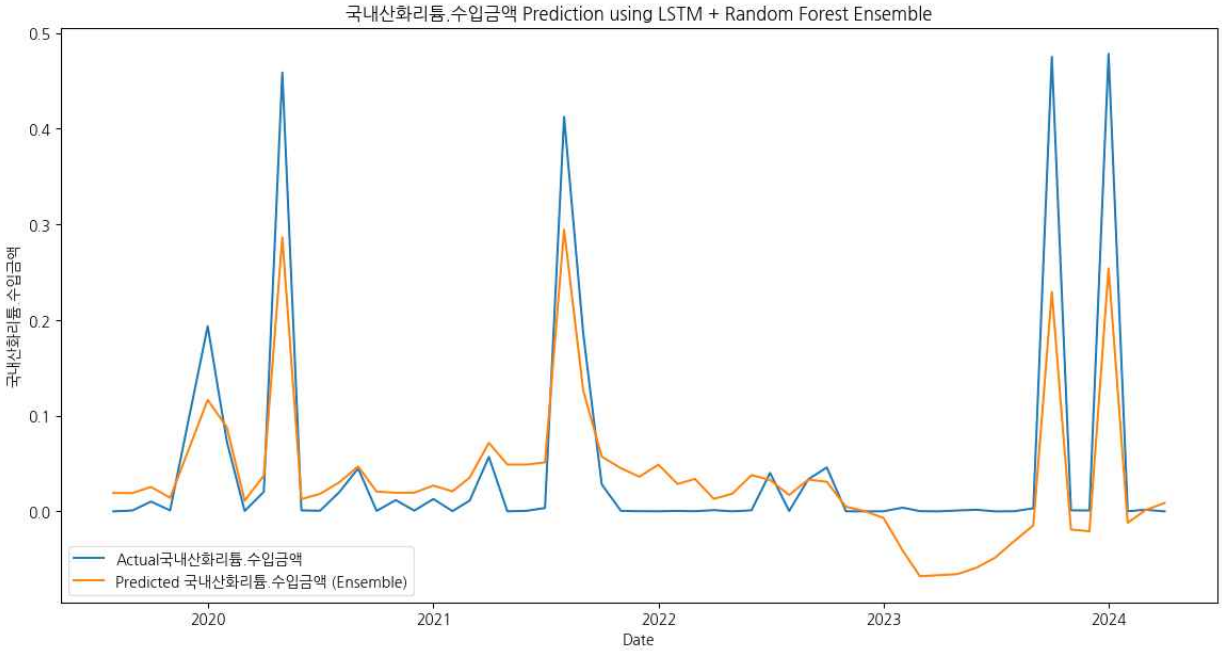
9) 배깅(Bagging): 배깅은 동일한 모델을 여러 개 훈련하되, 각각의 모델이 훈련되는 데이터 샘플이 서로 다름. 각 모델의 예측 결과를 결합하여 최종 예측을 만듬.

10) 부스팅 (Boosting): 부스팅은 순차적으로 모델을 훈련시키며, 각 모델이 이전 모델의 오류를 보완. 마지막에 모든 모델의 예측을 결합하여 최종 예측을 만듬.

11) 스태킹 (Stacking): 스태킹은 서로 다른 종류의 모델을 훈련시키고, 이들의 예측을 새로운 데이터로 사용하여 최종 메타 모델을 훈련시키는 방법. 메타 모델이 최종 예측을 만듬.

12) 심플 앙상블 (Simple Ensemble): 여러 모델의 예측을 단순히 평균내거나 가중 평균하여 최종 예측을 만듬.

실제와 예측값의 그래프는 아래와 같습니다.



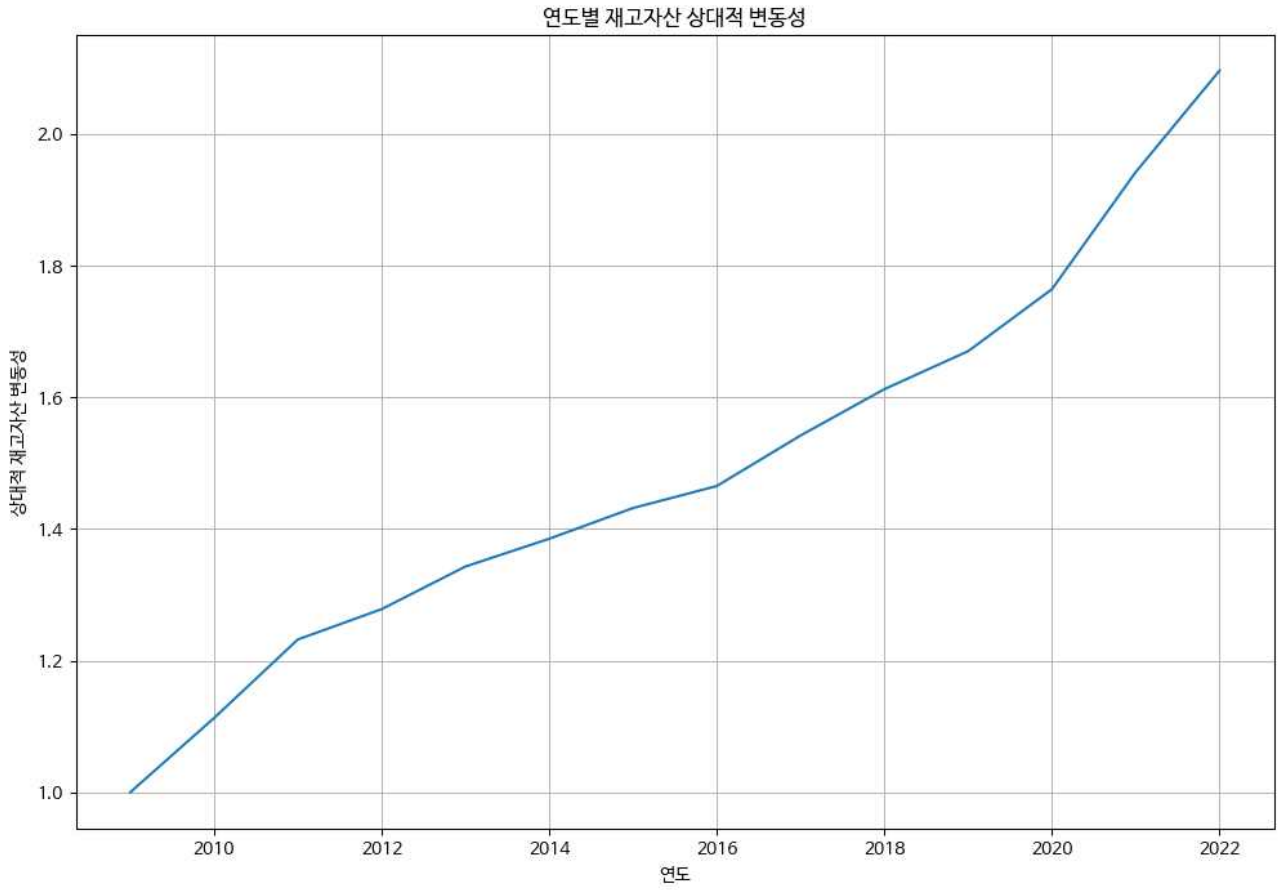
위의 단일 모형에서 예측하지 못했던 단점을 거의 모두 극복하고 이상치를 포함한 예상값을 매우 높은 정확도로 예측하는 것을 볼 수 있습니다.

앙상블 모형은 여러 모형들의 콜라보로 이루어지기 때문에 많은 성능 개선의 가능성이 남아있습니다. 따라서 위의 모형들을 기반으로 한 조금 더 다양한 앙상블 모형을 개발해야 한다고 생각합니다.

V. 사업화 방안 및 기대효과

1. 서론

한국 경제는 대기업에서 중견, 중소, 소기업에 이르기까지 긴밀하게 연결된 공급 및 생산체계를 갖추고 있습니다. 대기업은 다양한 사업을 통해 리스크를 분산시키고 대비하는 반면, 중소기업은 자본이 부족하고 사업 다각화가 미흡하여 공급망 리스크에 더욱 취약합니다. 공급망 리스크는 재고자산의 변동성 증가와 함께 중요한 요소로 부상하고 있습니다. 한국은행의 통계데이터에 따르면 2009년 부터 2022년까지 주요 광물이 필수적으로 쓰이는 전자 부품, 반도체, 자동차, 그리고 베터리 제조에 이르는 다양한 업종에서 재고자산의 변동성이 꾸준히 증가해왔습니다.



10년 새 두 배 가까운 변동성 증가는 원자재 공급 위기 관리에 상대적으로 취약한 중소기업에게 더 큰 부담으로 다가올 수 밖에 없습니다. 만약 중소기업에게 원자재 수급에 대한 정확한 정보를 제공할 수 있다면 사업 불안정성을 완화시킴으로써 경제전반에 긍정적인 영향을 미칠 수 있을 것입니다. 따라서 본 사업의 목표는 중소 및 중견 기업들을 위한 공급망 리스크 관리 솔루션을 개발하는 것입니다.

2. 주요 광물의 위기 대응 방안의 구체적 사례 및 그 한계

주요 광물의 공급 위기 대응 솔루션은 이미 해외 여러 이니셔티브에 의해 적극 추구 되고 있습니다. 예를 들어, The Responsible Business Alliance (RBA)[13]에서는 Responsible Minerals Initiative (RMI)를 운영하고 있습니다.[14] RMI는 분쟁 영향을 받거나 고위험 지역에서 광물을 조달하는 것과 관련된 리스크를 평가하고 관리하기 위한 도구와 자원을 제공합니다. 이 이니셔티브는 공급망의 잠재적 리스크를 식별하기 위해 공급업체로부터 데이터를 수집하는 분쟁 광물 보고 템플릿(Conflict Minerals Reporting Template, CMRT)을 사용합니다. 그러나 이것은 산업간 데이터 교류를 촉진하기 위한 프로토콜 개발에 가깝고 데이터를 사용한 공급 위기 예방에서는 부족한 면이 있습니다.

또 다른 예로는 OECD의 Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas[15]가 있습니다. 이 지침은 인권 침해, 부패, 분쟁 자금 조달과 관련된 리스크를 식별하고 평가하기 위한 프레임워크를 제공합니다. 그러나, 이 지침은 주로 인권 및 윤리적 문제 해결을 위한 가이드라인으로 작용하며, 공급망의 물리적 차질을 사전에 감지하고 예방하는 데에는 한계가 있습니다.
또한, 이 지침은 주로 기업들이 자사의 공급망에서 발생할 수 있는 인권 및 환경 리스크를 평가하고 관리하는데 중점을 두고 있으며, 실시간 데이터 모니터링이나 머신러닝을 통한 예측 분석과 같은 기술적 접근 방법은 포함되지 않습니다. 이로 인해, 공급망의 실제 차질을 선제적으로 예방할 수 없습니다.

한편, 미국의 미국 지질조사국(U.S. Geological Survey, USGS)은 글로벌 광물 자원, 생산 및 무역 데이터를 수집 및 분석하여 잠재적 공급 리스크를 식별하고 정책 결정을 위한 정보를 제공합니다. 이 데이터는 국가적 또는 산업적 수준에서 중요한 인사이트를 제공할 수 있지만, 개별 기업들이 특정 광물의 공급 리스크를 구체적으로 이해하고 대응하는 데에는 한계가 있습니다.

3. 사업화 솔루션의 제안 방안

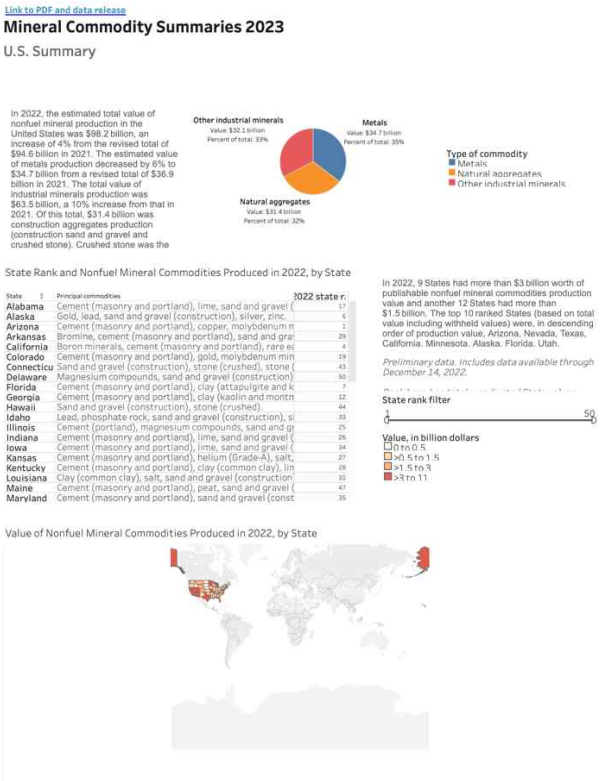
이에 우리가 제안하는 솔루션은 데이터 모델을 통해 이상을 초기 감지하고 기업들이 위험을 예방할 수 있게 한다는 점에서 우위를 갖습니다. 공급망 리스크 관리 솔루션은 웹 또는 모바일 플랫폼의 형태로 제공됩니다. 실시간 데이터 모니터링, 예측 분석, 알림 서비스, 커스터마이징 가능한 대시보드를 제공하여 사용자들이 필요한 정보를 즉시 확인할 수 있도록 합니다.
필요한 정보란 본 보고서에서 살펴보았던 머신러닝 모델(ARIMA, LSTM, Prophet 등)을

13) 글로벌 공급망에서 책임있는 비즈니스 관행을 촉진하기 위해 설립된 세계 최대의 산업 연합체
14) https://www.responsiblebusiness.org/initiatives/rmi/
15) https://www.oecd.org/daf/inv/mne/mining.htm

활용한 예측 분석 및 이상 탐지 모델을 통해 생산됩니다. 또한 중소기업 및 중견기업을 대상으로 하는 바, 월 구독형 서비스 또는 분석 당 과금 모델을 도입하여 접근성을 높이고, 데이터 분석 컨설팅 및 맞춤형 보고서를 제공하여 부가가치를 창출할 수 있습니다.

솔루션을 활용한 중소기업들은 재고관리 및 공급망 리스크관리에 대한 정확한 정보에 접근함으로써 리스크를 효과적으로 분산하고 핵심 광물공급 차질에 대비할 수 있게됩니다. 이로써 단기적으로는 중소기업의 경영 안정성을 높이고, 장기적으로 국내의 중소기업 생태계 구조 다변화까지 기대할 수 있습니다.

참고) Figure 1 USGS주요광물공급망리스크대시보드



4. 향후 시스템 구축 방안

우리가 구상하고 있은 모델의 구체적 모습은 다양한 데이터 소스를 통합한 데이터 통합 플랫폼이며, 우리는 글로벌 시장 데이터, 환율 정보, 정치적 리스크 지표 등을 분석함으로써 보다 정확한 변동성 예측을 수행할 것입니다.

우리는 실시간 예측 모니터링 시스템을 구축하여 자원 가격 변동 추이를 시각화하고, 주요 변동성을 실시간으로 감지하여 신속한 대응이 가능하게 할 것입니다.
GARCH, LSTM, Prophet 등의 예측 모델을 지속적으로 최적화하고 머신러닝 기법과 다양한 모델들과의 앙상블 조합을 적용하여 예측 성능을 주기적으로 개선 하도록 하겠습니다.

기업 사용자들을 대상으로 정기적인 워크숍과 온라인 교육을 제공하여 예측 시스템 사용법과 데이터 해석 방법에 대한 전문성을 강화하고, 예측 결과를 효과적으로 활용할 수 있도록 지원할 계획입니다.

이러한 시스템 구축은 변동성 예측의 정확성을 높이고, 기업의 리스크 관리와 비용 절감을 극대화할 수 있다고 생각합니다.

5. 기업들의 모델 적용의 실제적 활용방안

수혜 기업들은 모델의 변동성 예측을 활용하여 자원 관리의 효율성을 극대화할 수 있습니다. 먼저, 비축 전략을 통해 자원 가격 변동성이 높은 시기에 대비할 수 있습니다. 예를 들어, 리튬 가격이 급등할 것으로 예측되면, 해당 시점 이전에 충분한 양의 리튬을 비축하여 가격 상승에 따른 비용 증가를 방지합니다. 이는 자원 확보의 불확실성을 줄이고, 경영 안정성을 강화할 것으로 생각합니다.

또한, 수혜 기업들은 구매 계약 최적화를 통해 자원 구매 시기를 최적화할 수 있습니다. 니켈 가격이 하락할 시점을 예측하여 장기 계약을 체결하거나 단기 구매로 대응함으로써 자원 구매 비용을 절감하고 안정적인 공급을 유지합니다. 최적의 구매 시점을 예측하고 이에 맞춘 구매 전략을 수립하여 자원 확보 비용을 최소화하고 경쟁력을 유지할 수 있도록 활용가능합니다.

기업들의 재고 관리 효율화도 중요한 전략입니다. 가격 변동성이 높은 기간에는 재고를 높게 유지하여 급격한 수요 증가나 공급 차질에 대비하고, 안정적인 시기에는 재고를 낮게 유지하여 관리 비용을 최소화합니다. 이는 수혜기업들의 자원 수입의 변동성을 효과적으로 통제하고 비용 효율성을 높이는 데 기여할 수 있습니다.

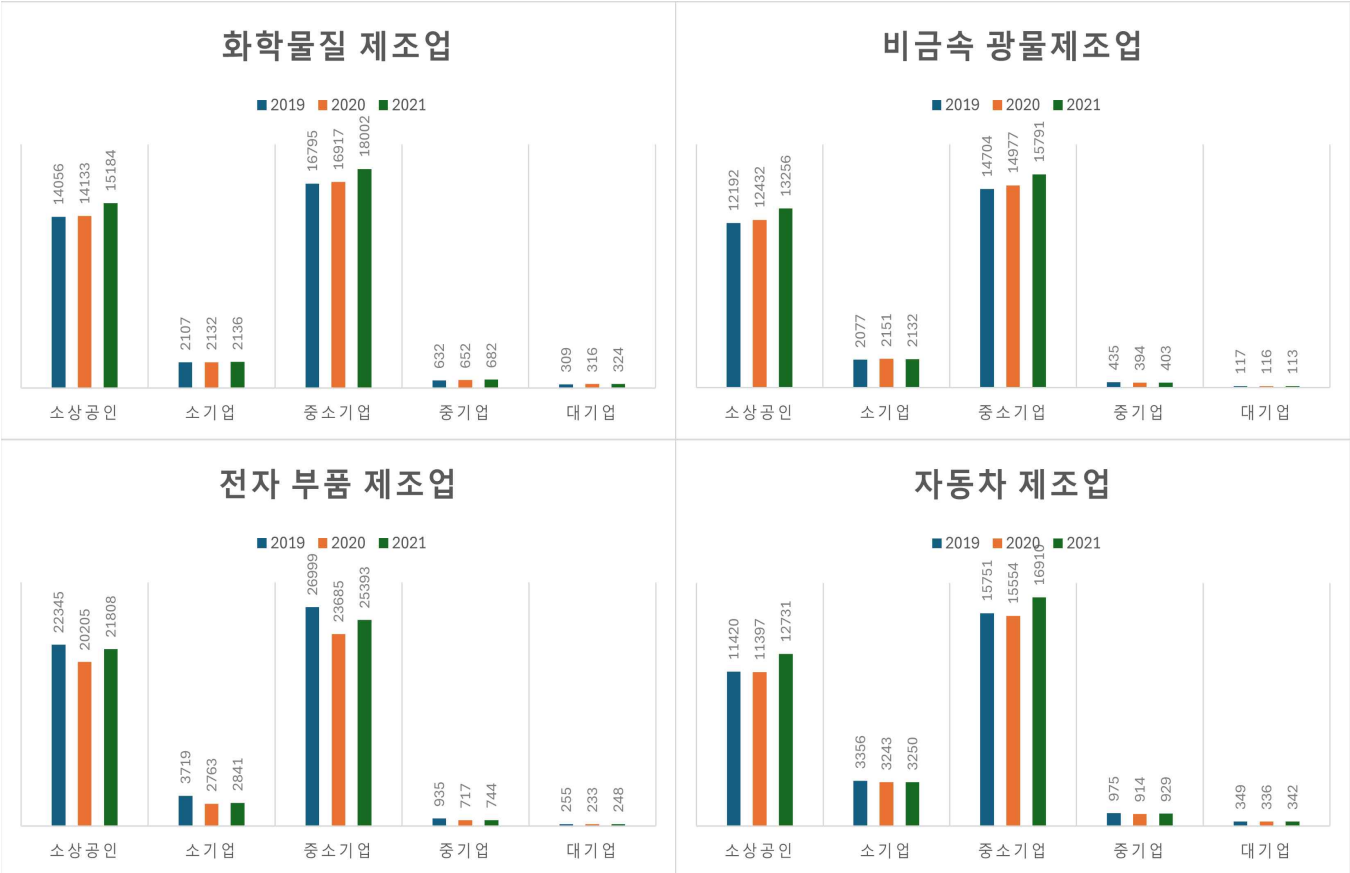
수혜 기업은 위기 대응 계획을 통해 자원 수입에서 발생할 수 있는 위기를 사전에 감지하고 대응 방안을 마련할 수 있습니다.
만약 특정 국가의 정세 불안정으로 인한 공급 차질을 알수 있다면 기업은 대체 공급원을 확보하거나 비상 계획을 수립하여 공급망 리스크를 최소화 할 수 있습니다. 이는 예상치 못한 상황에 신속하게 대응하고 공급망의 안정성을 유지하는 데 중요합니다.

마지막으로, 가격 헤징 전략을 통해 자원 가격 변동에 대한 재무적 리스크를 관리할 수 있습니다. 기업이 리튬 가격이 상승할 것으로 예측되면, 리튬 선물 계약을 통해 가격 리스크를 헤징하여 재무적 불확실성을 줄입니다. 이는 자원 가격 변동에 대한 대비책을 마련하고 재무 리스크를 관리하는 데 중요한 역할을 합니다.

6. 국내 사업체들의 영향 평가

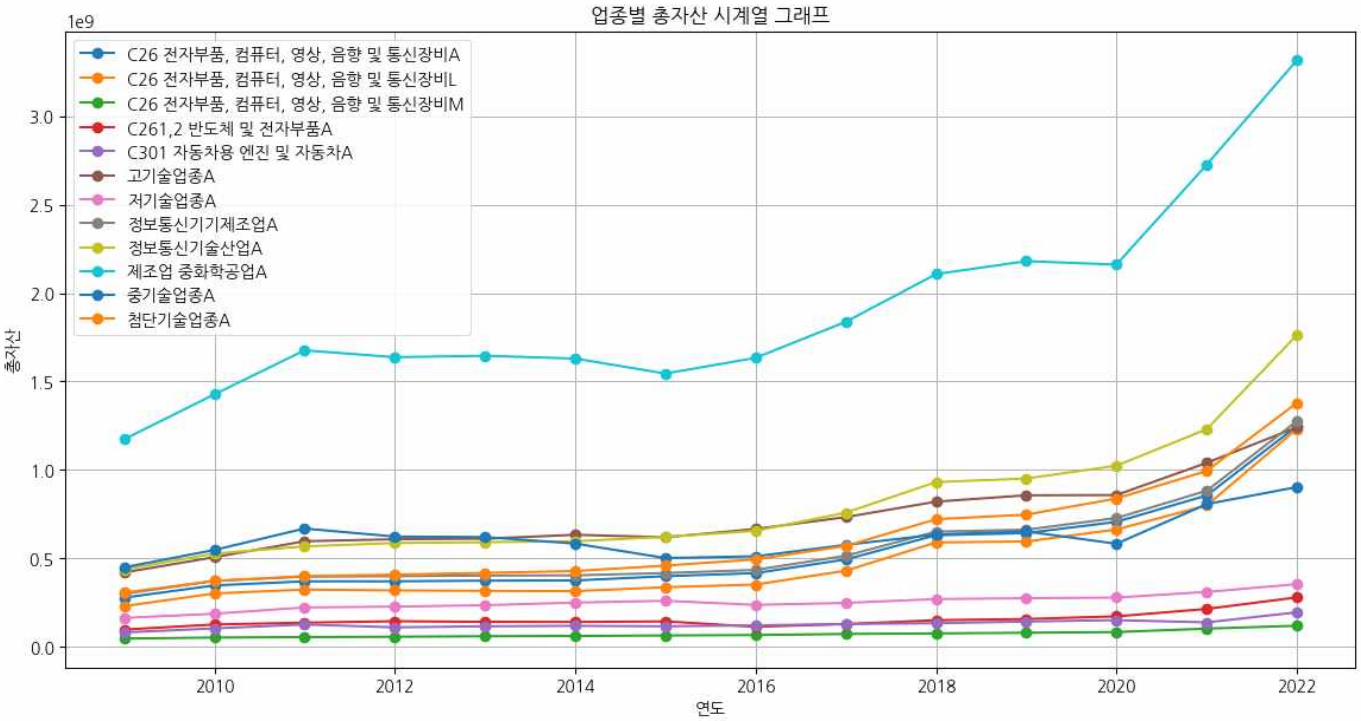
우리는 국내 주요광물을 원재료로 구입하여 제조하는 기업들과 연관된 솔루션 모델의 영향 평가를 위하여 현재 그와 관련된 업종의 기업들을 통계청(KOSIS)과 한국은행 통계자료를 활용하여 평가를 수행하였습니다.

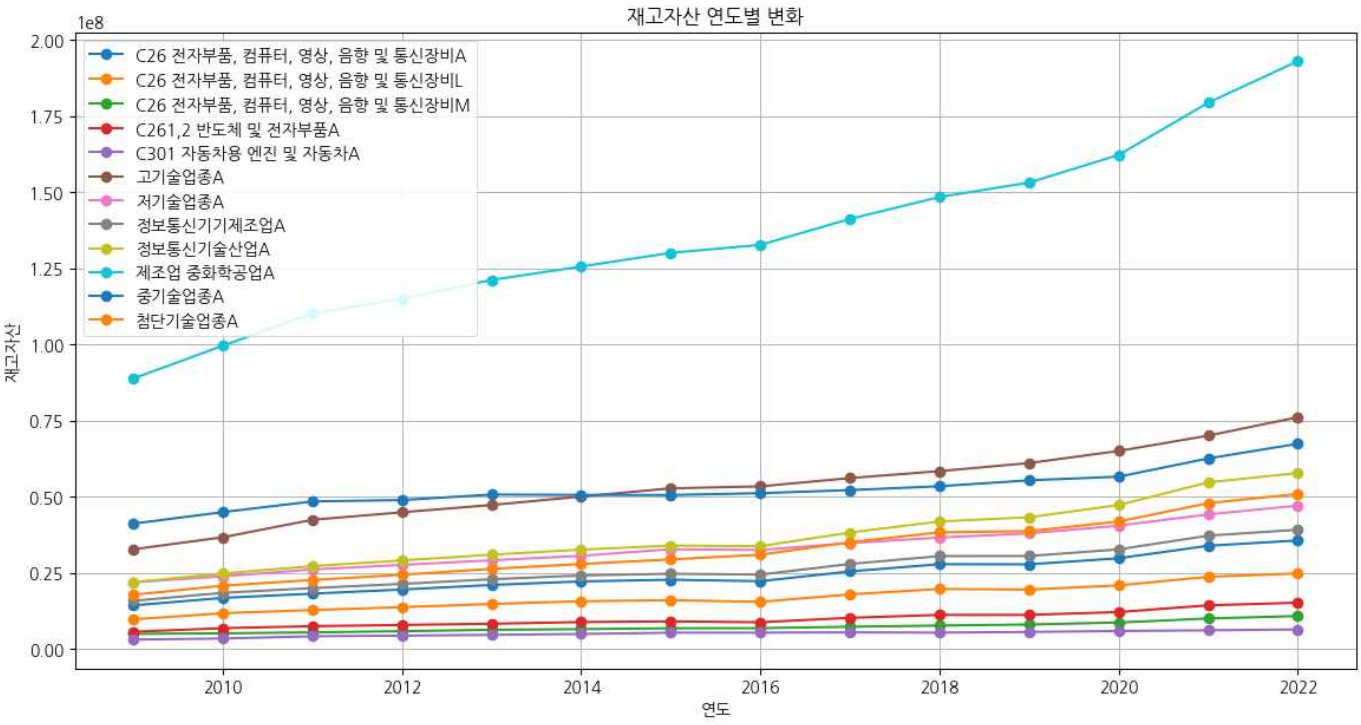
주요광물과 직접 혹은 간접적으로 연관되어 있는 기업들의 업종은 1. 화학물질 제조업, 2. 비금속 광물제조업, 3. 전자 부품제조업, 4. 자동차 제조업 이며, 이는 통계청 업종분류에 따른 자료입니다.



21년 자료를 보면 소상공인 사업체는 화학물질 제조업에 15184개, 비금속 광물제조업에 13256개, 전자 부품 제조업에 21808개, 자동차 제조업에 12731개 이며, 중소기업 사업체는 화학물질 제조업에 18002개, 비금속 광물제조업에 15791개, 전자 부품 제조업에 25393개, 자동차 제조업에 16910개 였습니다. 모든 업종에서 중소기업의 비중이 가장 높았으며 다음으로 소상공인, 소기업, 중기업, 대기업 순으로 비중을 차지하였습니다. 그 중에서 소상공인과 중소기업이 전체 비중에 대다수를 차지하였으며 전체적으로 중소기업 이하 규모의 기업체수는 증가하는 추세에 있습니다. 우리는 해당 사업체들이 우리가 기획한 솔루션 서비스의 혜택을 받는 잠재적 수혜집단이라고 생각하고 있습니다.

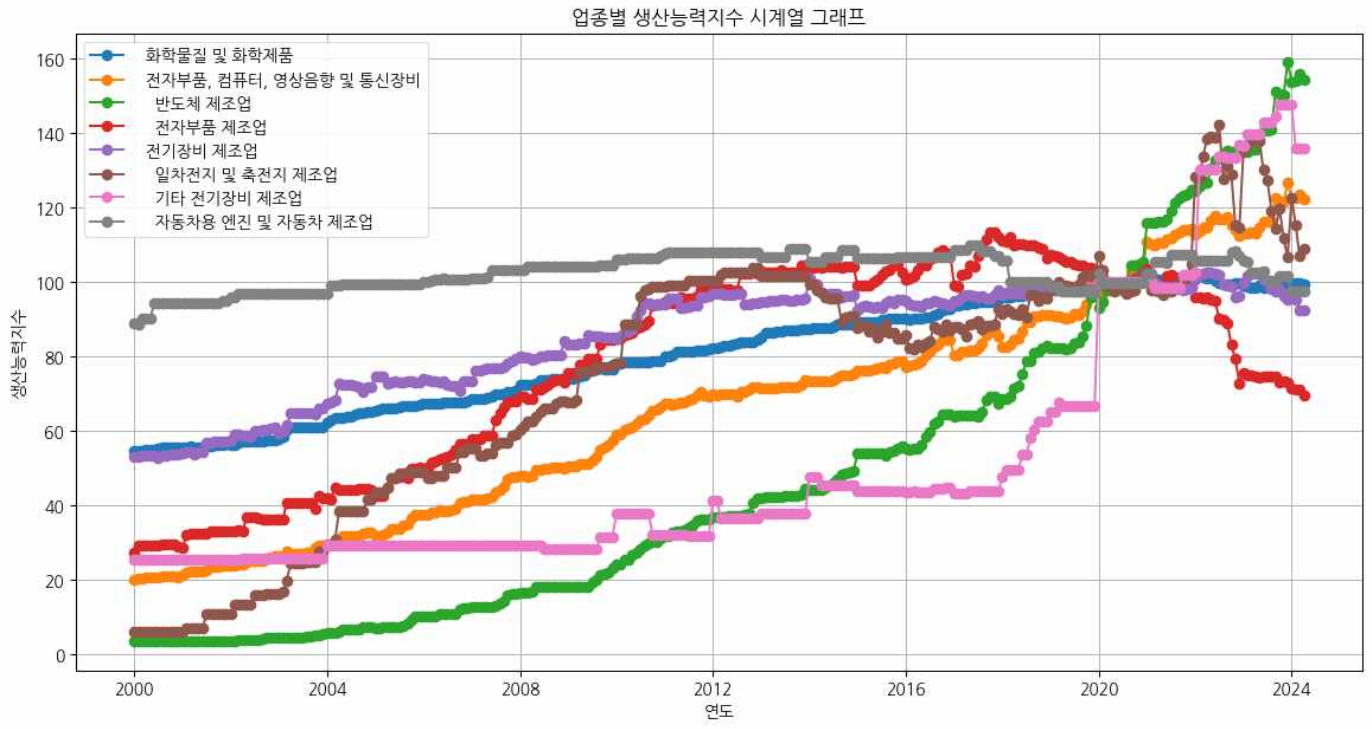
또한 해당 업종들의 총자산은 증가하고 있으며, 재고자산 또한 증가하는 추세에 있습니다.

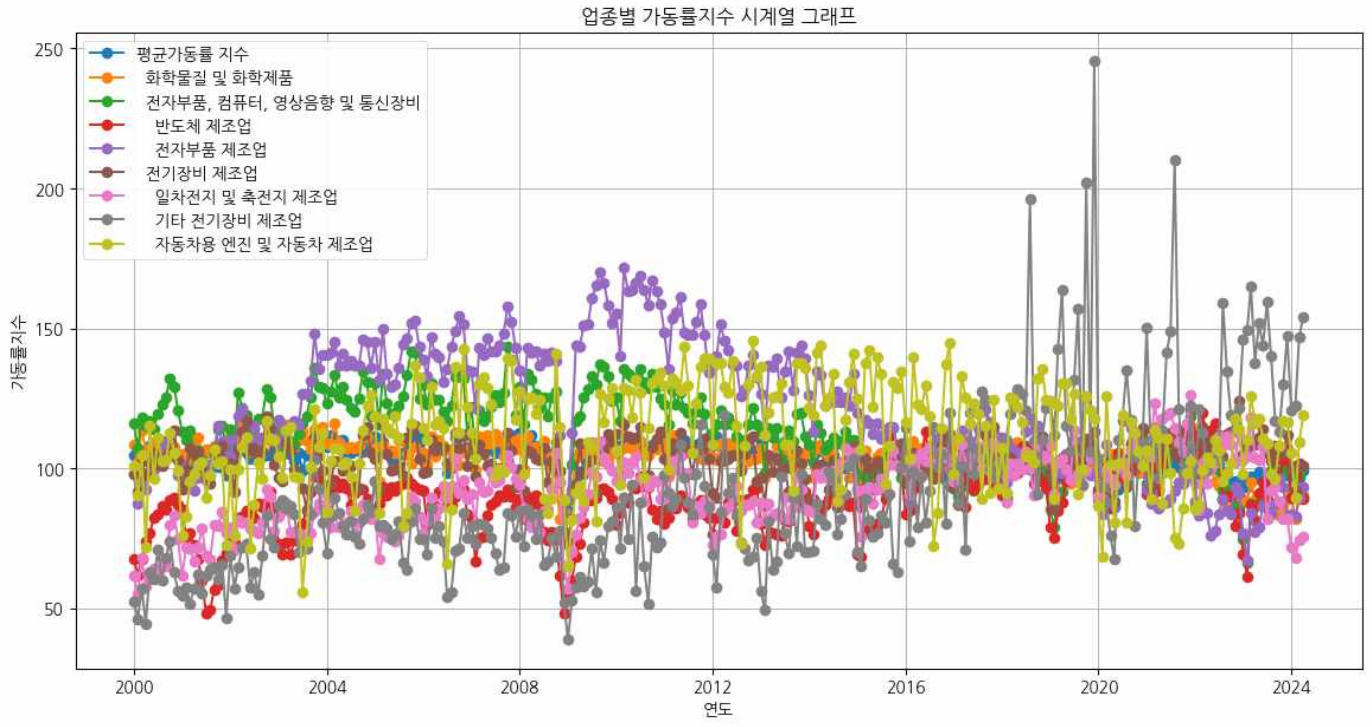




그리고 관련 업종에 따라 생산능력지수[16]와 가동률지수[17]까지 고려하여 종합적으로 판단하여 보았을 때 일부 정체된 부분이 있지만 전체적으로 시장은 증가하는 추세에 있다고 판단되어집니다.

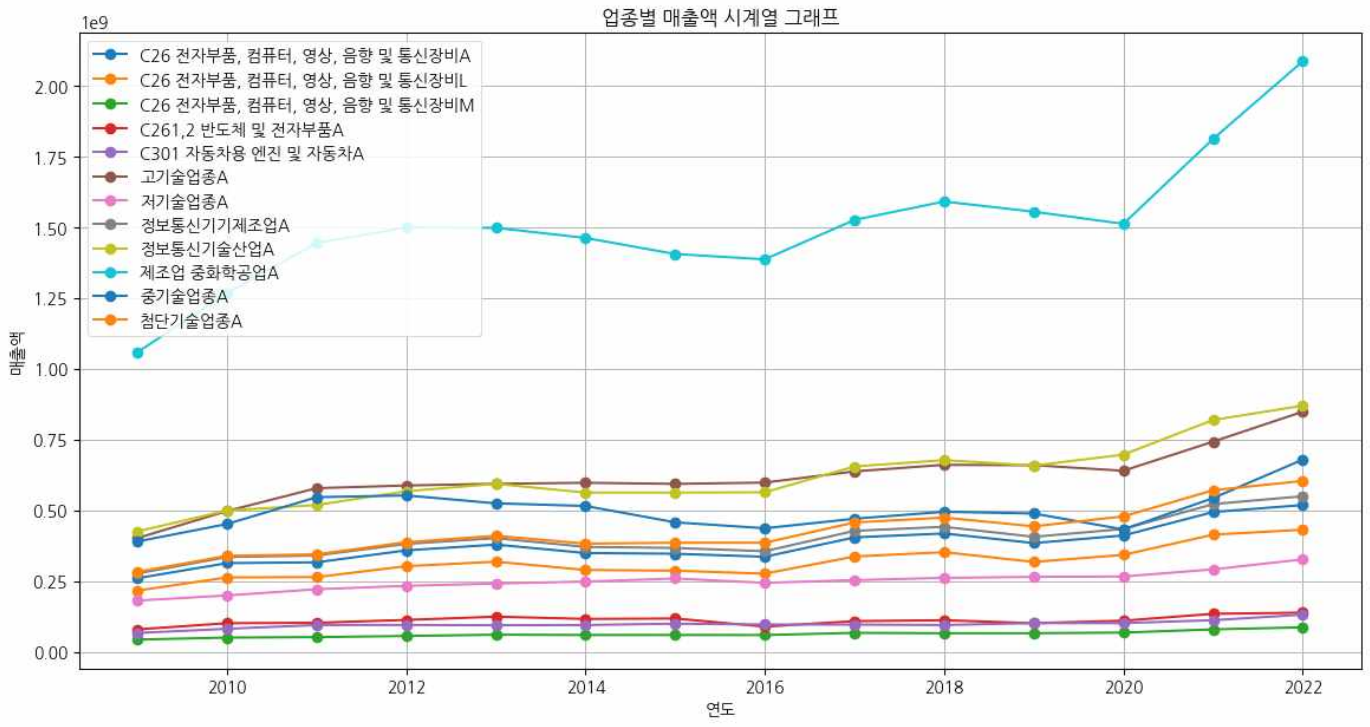
16) 생산능력지수 (Capacity Index): 산업의 설비와 자본이 최대한 가동되었을 때의 생산 능력
17) 가동률지수 (Operating Rate Index): 실제 생산된 생산량을 최대 생산 가능량과 비교하여 나타낸 지수



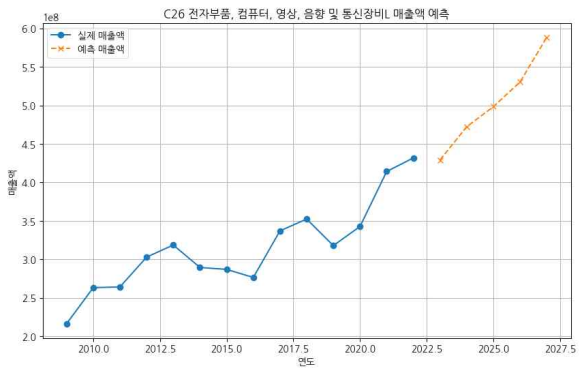


또한 관련 업종에 대한 매출액도 증가하는 추세에 있어 관련 솔루션의 필요성은 꾸준히 증가한다고 생각되어집니다.

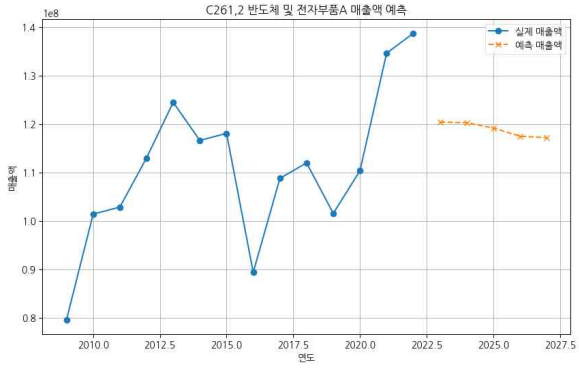
다음은 업종별 매출액 시계열 그래프와 LSTM으로 예측한 업종별 매출액 예측수치입니다.

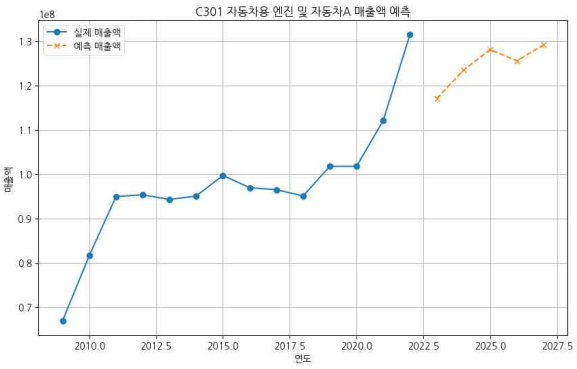

업종별 매출액 시계열 그래프
1e9
C26 전자부품, 컴퓨터, 영상, 음향 및 통신장비A
C26 전자부품, 컴퓨터, 영상, 음향 및 통신장비L
C26 전자부품, 컴퓨터, 영상, 음향 및 통신장비M
C261,2 반도체 및 전자부품A
C301 자동차용 엔진 및 자동차A
고기술업종A
저기술업종A
정보통신기기제조업A
정보통신기술산업A
제조업 중화학공업A
중기술업종A
첨단기술업종A
매출액
연도


C26 전자부품, 컴퓨터, 영상, 음향 및 통신장비A 매출액 예측
실제 매출액
예측 매출액
매출액
연도


C26 전자부품, 컴퓨터, 영상, 음향 및 통신장비L 매출액 예측
실제 매출액
예측 매출액
매출액
연도


C26 전자부품, 컴퓨터, 영상, 음향 및 통신장비M 매출액 예측
실제 매출액
예측 매출액
매출액
연도


C261,2 반도체 및 전자부품A 매출액 예측
실제 매출액
예측 매출액
매출액
연도


C301 자동차용 엔진 및 자동차A 매출액 예측
실제 매출액
예측 매출액
1e8
매출액
연도


고기술업종A 매출액 예측
실제 매출액
예측 매출액
1e9
매출액
연도


저기술업종A 매출액 예측
실제 매출액
예측 매출액
1e8
매출액
연도


정보통신기기제조업A 매출액 예측
실제 매출액
예측 매출액
1e9
매출액
연도


정보통신기술산업A 매출액 예측
실제 매출액
예측 매출액
1e9
매출액
연도


제조업 중화학공업A 매출액 예측
실제 매출액
예측 매출액
1e9
매출액
연도


중기술업종A 매출액 예측
실제 매출액
예측 매출액
1e8
매출액
연도


첨단기술업종A 매출액 예측
실제 매출액
예측 매출액
1e9
매출액
연도

반도체 및 중기술업종은 계절성을 띄고 있어 매출증가가 단기적 예측치에서는 보이지 않았지만, 대부분의 업종에서 매출액이 증가하는 경향을 보였습니다.

따라서 관련 중소, 소상공인 사업체들의 솔루션 필요성은 존재한다고 생각하고 주요 광물에의 리스크에 대한 사전에 정보를 제공하는 서비스는 국내 중소 사업체들에게 미치는 긍정적 효과가 존재한다고 생각합니다.

참고문헌 .

- 현안과과제_세계 2차전지 공급망 구조 현황과 시사점_23-14(현대경제연구원) - 한재진 연구위원
- Using Oil and Lithium Carbonate to Predict New Energy Vehicle Sales in China by Linear Regression Analysis(TPCEE 2022) - Yilin Wang
- Tracing of lithium supply and demand bottleneck in China's new energy vehicle industry—Based on the chart of lithium flow(frontiers) - Linchang Zheng, Ge Chen, Litao Liu, and Yuqi Hu
- Tracing of lithium supply and demand bottleneck in China's new energy vehicle industry—Based on the chart of lithium flow(MDPI) - Zhiyong Zhou, Jianhui Huang
- 공급망 분석을 통해 살펴본 한·중 무역구조 변화와 시사점(한국무역협회) - 김나율, 강내영, 김민우
- 이차전지 수출 변동 요인과 향후 전개방향(한국무역협회) - 도원빈
- 글로벌 공급망 블록화에 따른 중국의 전략과 우리의 대응_이차전지산업을 중심으로(KIET 산업연구원) - 조은교, 심우중, 서동혁
- 비축품목 다변화를 위한 연구(KIET 산업연구원) - 이재윤, 정은미, 송명구
- 희소금속의 유동구조 및 산업수요 조사 연구용역(한국생산기술연구원) - 서석준, 김택수, 심재진
- 원자재시장 동향 및 시사점(BNK 금융경영연구소) - 정성국, 권민지
- 해외경제 포커스_제 2015-14호(한국은행) - 조사국 국제경제부
- 한 눈에 보는 6대 핵심광물 이슈 분석(KIGAM 한국지질자원연구원) - 미래전략연구센터