기계학습을 활용한 수출증감률 예측

안채린, 유헌창
고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
cherin13@korea.ac.kr, yuhc@korea.ac.kr

Predicting Export Change Rate using Machine Learning Methods

Chaerin Ahn, Heonchang Yu
Dept. of Computer & Information Technology, Korea University

요 약

수출의존도가 높은 한국은 코로나19 팬데믹, 우크라이나-러시아 전쟁 등 대외환경의 변화에 따른 수출 여건에 민감할 수 밖에 없는 환경이다. 이에 발 빠르게 대응하기 위해 정확한 수출증감률 예측이 필요하며 이를 가장 잘 수행할 수 있는 예측모델을 찾고자 한다. 수출에 영향을 끼치는 주요변수 선정 후, min-max 정규화를 시행하고 변수간 상관계수와 다중공선성 확인을 통해 변수를 축소했다. 그리고 머신러닝 예측모델로 많이 사용되는 Linear Regression, Decision Tree, Gradient Boost Regressor, Random Forest 4가지 모델에 대입하여 수출 증감률 예측 정확도를 비교했다. 그 결과, Linear Regression의 MSE가 0.087로 가장 낮아 제일 우수한 모델이라는 결론에 도달했다.

1. 서론

인구감소, 자원부족 등으로 인해 내수시장 규모가 작은 한국은 GDP 대비 수출입액이 차지하는 비율이 2021년 기준 84.8%에 달할 정도로 높은 무역의존도를 보인다[1]. 특히 최근에는 2020년부터 지속된 코로나19 팬데믹과 우크라이나-러시아 전쟁, 급상승한 금리 등의 여러 변수에 더욱 민감할 수밖에 없는 환경이다. 수출에 주력하는 여러 기업뿐만 아니라 이를 지원 및 관리하는 정부. 기관 모두 발빠르게 수출입변화를 예측하고 대응하는 능력이 그 어느때보다 필요한 시점이다.

따라서 본 연구에서는 수출에 영향을 미치는 여러 경제변수를 바탕으로 수출증감률 예측 성능이 가장 우수한 모델을 찾고자 한다. 다중공선성 확인을 통해 변수 갯수를 압축하고 Linear Regression, Decision Tree, Gradient Boost Regressor, Random Forest 4가지 모델의 평균제곱오차(MSE)값 비교를 통해 가장 우수한 성능의 모델이 무엇인지 도출하고자 한다. 마지막으로 이를 활용하여 추후 경제 변화에 적절히 대응할 수 있는 방안을 제시한다.

2. 관련 연구

관련 연구를 살펴보면, 최근에는 머신러닝을 기반으로 수출입을 예측하는 연구가 증가하고 있으며, 과거 전통적 방식인 통계기법과 성능을 비교하는 연구도 등장하고 있다. 장나원, 한희준(2022)은 보루타 알고리즘을 통해 중요 변수 순위를 도출하고 랜덤 포레스트를 통해 수출입 증가율을 예측하는 연구를 시행했다[2]. 이를 통해 최적의 변수 개수로 축소한 뒤 랜덤 포레스트를 실시했을 때 예측력이 가장 우수하다는 결론에 도달했다. 남성휘(2021)은 시계열분석 모형과 LSTM 등의 머신러닝 모형들의 수출증가율 예측 성능을 비교했다[3]. 그 결과, 시계열분석이 타 머신러닝 분석보다 예측 성능이 우수하게 도출되었다. 본 연구에서는 가장 쉽게 접근할 수 있는 Linear Regression, Decision Tree, Gradient Boost Regressor, Random Forest 4가지 모델 성능 비교에 집중하여 살펴보려고 한다.

① Linear Regression

Linear Regression은 선형 회귀분석으로 종속 변수 y와 한 개 이상의 독립 변수 x와의 선형 상관관계를 모델링하는 회귀분석 기법이다[4].

② Decision Tree

Decision Tree 분석기법은 나무 구조로 도식화하여 의사결정 과정을 나타내고, 이를 활용해 분류와 예측을 수행하는 분석방법이다. 이는 비모수적 방법으로 선형성, 정규성 또는 등분산성 등의 통계적 가정을 필요로 하지 않는다는 장점이 있다[5].

③ Gradient Boost Regressor

Gradient Boosting Tree는 기존 Decision Tree가 단일 분류자를 사용했던 것과 달리 분류자들의 예측을 종합함으로써 정확성을 높이는 앙상블기법의 하나이다. 이를 변형한 XGBoost, LightGBM, CatBoost 등도 자주 사용되고 있다[6].

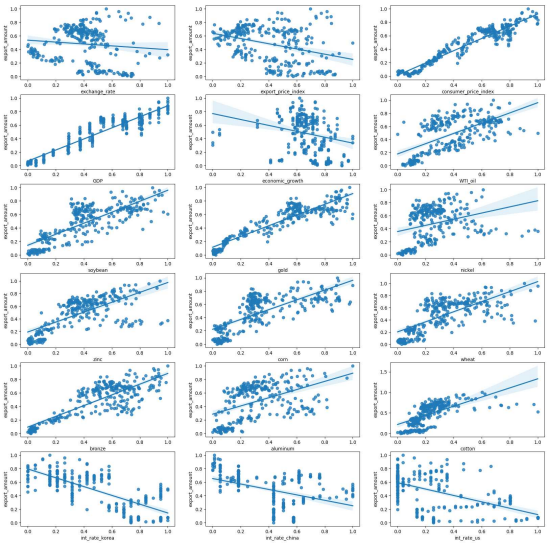
④ Random Forest

Random Forest는 반응변수를 예측하기 위해 여러 Decision Tree를 결합하는 과정을 통해 기존 모델의 예측력을 높이고 안정성을 강화시킨다. 부스트랩 과정을 통해 하나의 데이터로도 다양한 데이터 모델링을 하고 이를 종합한 후에 최종모형을 만드는 bagging 방식이다[7].

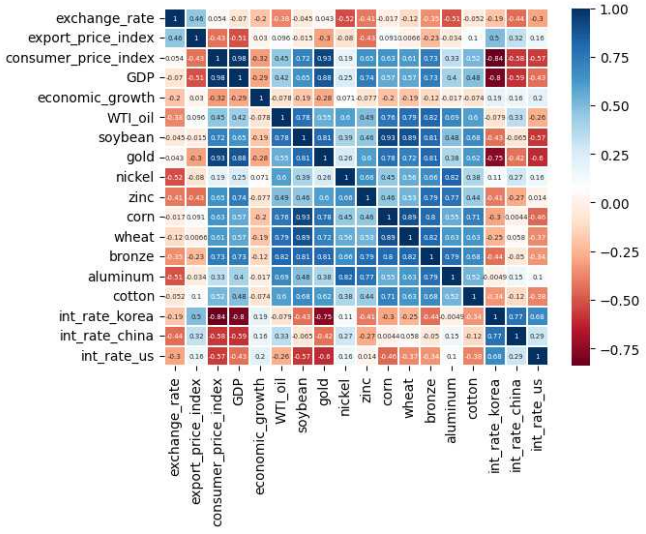
3. 데이터 전처리

본 연구에서는 수출액 예측을 위해 가장 주요하다고 생각되는 변수 18개(원달러 환율, 수출물가지수, 소비자물가지수, GDP, 경제성장률, 주요 10대 원자재 가격, 한·중·미 금리)를 선정하여 2001~2022년 월별 데이터를 활용하였다. 국내총생산(GDP)와 같이 연도별 데이터만 존재하는 변수의 경우 매달 같은 금액을 적용시켰다. 변수별 값의 편차를 줄이기 위해 최솟값 0, 최댓값 1로 변환하는 min-max 정규화를 수행하였으며 소숫점 넷째자리에서 반올림하였다.

그림 1은 독립변수 18개와 수출액 간의 관계를 그래프로 표현한 것이다. 대다수의 독립변수들이 수출액과 강한 선형성을 보이고 있다. 또한, 그림 2는 각 독립변수끼리의 상관계수를 확인한 히트맵이다. 변수간 상관계수도 높아 다중공선성을 줄이기 위해 VIF (Variance Inflation Factor) 값을 계산했다. 그 중 가장 높은 값을 띄는 변수인 소비자물가지수, 한국금리, bronze 변수를 제거했다. 마지막으로 학습용 데이터셋과 테스트용 데이터셋은 가장 일반적으로 머신러닝에서 적용하는 7:3 비율로 나눠 진행하였다.



(그림 1) 변수와 수출액 선형관계



(그림 2) 변수간 상관계수 히트맵

4. 예측 모델 성능 비교

여러 연구에서 예측 모델로 많이 사용되는 Linear Regression, Decision Tree, Gradient Boost Regressor, Random Forest 네 가지를 활용하여 수출액을 예측해보았다. 예측모델의 성능을 비교하는데 가장 많이 사용되는 평균제곱오차(MSE)를 각 모델별로 구하여 비교했다. 이 값이 낮을수록 우수한 모델로 평가된다.

그 결과, 표 1과 같이 Linear Regression의 MSE는 0.087, Decision Tree은 0.119, Gradient Boost Regressor는 0.106, Random Forest는 0.103으로 나타났다. MSE 값의 관점에서만 봤을 때는 Linear

Regression의 성능이 가장 우수하게 도출되었다고 볼 수 있다.

그러나 위와 같이 도출된 결과는 독립변수와 수출금액간의 높은 선형관계가 큰 영향을 미쳤기 때문이며, 이에 따라 4가지 모델 모두 MSE 값이 0.08~0.12 사이 정도로 낮게 도출되었다. 각 모델간의 편차 또한 크지 않은 만큼 근소한 차이로 모델의 우수성을 학정적으로 판단하기에는 한계가 있으며, 후속 연구를 통해 여러 관점을 통한 다각적 접근이 필요할 것으로 생각된다.

<표 1> 예측 모델별 MSE 비교

예측 모델	MSE
Linear Regression	0.087
Decision Tree	0.119
Gradient Boost Regressor	0.106
Random Forest	0.103

5. 결론

본 연구에서는 수출 증감률을 예측하기 위해 주요변수 선정 후, min-max 정규화하고 변수간 상관계수와 다중공선성 확인을 통해 갯수를 축소했다. 그리고 머신러닝 예측모델로 많이 사용되는 Linear Regression, Decision Tree, Gradient Boost Regressor, Random Forest 4가지 모델에 대입하여 수출 증감률 예측 정확도를 비교했다. 그 결과, Linear Regression의 MSE가 0.087로 가장 낮아 위 비교조건 하에서는 제일 우수한 모델이라는 결론에 도달했다. 이와 같은 결론이 도출된 이유는 그림1에서 알 수 있듯이 대다수의 변수들이 수출액과 선형관계에 있어 Linear Regression의 성능이 가장 높게 나타난 것으로 추측된다.

각 모델별 MSE 값 자체가 낮게 나왔으며 편차도 작은 만큼 모델간 명확한 비교우위를 확인하기 위해 추후 연구에서는 다각적인 시도로 접근해보고자 한다. 다른 머신러닝 모델들의 hyperparameter 조정을 통해 예측성능을 더 높일 수 있는지 확인해보고자 한다. 또한, 선형성이 낮은 다른 변수들을 활용하여 예측하는 경우와 성능을 비교하고, 종합적으로 가장 수출 증감률을 잘 예측할 수 있는 설명력 높은 변수들의 조합을 도출해보고자 한다. 이를 활용하여 빠르게 변화하는 수출 여건 속에서 이에 대응할 수 있는 정부 대책을 마련하고, 각 산업에서의 대응 방안을 모색할 수 있을 것이다.

참고문헌

[1] 국가지표체계 Kindicator ‘GDP 대비 수출입비율’ 자료 (www.index.go.kr)

[2] 장나원, 한희준, “머신러닝 방법을 활용한 한국의 수출입 증가율 예측 및 분석”, 국제경제연구 제28권 제 4호, 2022. 12.

[3] 남성휘, “시계열 분석 모형 및 머신 러닝 분석을 이용한 수출 증가율 장기예측 성능 비교”, 무역학회지 제 46권 제 6호, 2021.

[4] B. Kim and J. Kim, “An Optimal Design Method of a Linear Generator for Conversion of Wave Energy,” Journal of the Korea Institute of Electronic Communication Sciences, Vol.16, No.6, Dec. 2021, pp. 1195-1204.

[5] 김신곤, 박성용. “의사결정트리 알고리즘의 성과 비교에 관한 연구,” 한국경영정보학회 학술대회, pp. 371-383, 1999.

[6] Friedman, J. H. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp. 1189-1232, 2001.

[7] Breiman. L, "Random Forests," Machine Learning, Vol.45, pp. 5-32, Oct. 2001.