

인공지능기반 해상운임 예측 연구 (3차년도)

A Study on AI-based Prediction of Bulk Shipping Freight Rates
(3rd Year)

황수진 · 이호춘 · 최건우 · 김병주



한국해양수산개발원
KOREA MARITIME INSTITUTE

저자	황수진, 이호춘, 최건우, 김병주
내부연구진	황수진 한국해양수산개발원 해운·물류연구본부 전문연구원 이호춘 한국해양수산개발원 해운·물류연구본부 부연구위원 최건우 한국해양수산개발원 해운·물류연구본부 전문연구원 김병주 한국해양수산개발원 해운·물류연구본부 연구원

연구기간 2021. 5. 17. ~ 2021. 10. 31.

보고서 집필내역

연구책임자	황수진 연구총괄, 제1장, 제2장, 제3장, 제4장, 제5장
내부연구진	이호춘 제1장 일부 최건우 제2장 일부, 제5장 일부, 부록 김병주 제2장 일부

발간사

코로나19의 전 세계적 확산세가 지속되면서 경제 활동은 비대면 사회로 빠르게 전환하고 있으며, 그 중심에 정보통신기술을 바탕으로 빅데이터, 인공지능 등과 같은 4차 산업혁명을 대표하는 핵심기술이 있다. 이러한 시대적 흐름에 발맞춰 국내 유일 해양수산정책연구기관 한국해양수산개발원은 2018년 3월 해운빅데이터연구센터를 설립하고 선도적으로 변동성이 커지는 해상운임 전망의 정도를 제고하기 위해 다원적인 시각과 새로운 분석 방법을 도입하기 위한 지속적인 노력을 기울이고 있다.

해운 분야에서 시황에 대한 예측은 과거부터 지속해서 시도되고 있지만, 주식시장을 훨씬 상회하는 변동성을 가지고 있어 구체적인 성과를 거두기 어려웠으며, 특히 건화물선 시장은 코로나19, 기상이변 외에도 과거보다 시장 전체에 영향을 미치는 사건들이 빈번하게 발생함에 따라 실시간으로 시장을 관찰하고 전망할 수 있는 모델에 대한 수요가 증가하고 있다.

이러한 시대적 흐름에 맞춰 2019년 1월부터 대표적인 인공지능 알고리즘인 심층신경망 모델을 구축하고 이를 「KMI 주간해운시장포커스」를 통해 정부나 해운기업에 서비스 중이며, 2020년 5월부터는 VAR/VECM 모델을 추가하였다. 이처럼 전통적인 방법과 최신 전망기법을 총망라한 단기 예측 시스템을 구축하고자 노력 중이며, 본연 구도 그 일환으로 2020년 구축한 LSTM 모델을 개선하고, GRU와 같은 최신 기법을 적용한 모델을 추가하였다. 또한 정부와 해운기업의 정보 활용도를 높이기 위해 모델 예측력 향상과 기간 확장에 주안점을 두어 연구가 진행되었다.

이번 연구 이후에도 지속적인 연구를 통해 한국해양수산개발원이 해운 시황 전망을 위한 다양한 모델을 구축해 나날이 커지는 변동성과 불확실한 해운 환경 변화 속에 정부와 해운기업에 가치 있는 정보를 생산하고 제공할 것으로 기대된다.

끝으로 이 연구의 책임을 맡은 해운·물류연구본부의 황수진 전문연구원과 연구진의 노고에 깊은 감사를 표한다. 아울러 이 연구에 자문과 조언을 아끼지 않은 임남균 목

포해양대학교 교수와 해양수산부 김한울 서기관께 깊은 감사를 드린다. 이 보고서가
출판되기까지 도움을 주신 모든 분께 이 자리를 빌려 감사를 드린다.

2022년 1월
한국해양수산개발원
원장 김 중 덕

목차

요약_i

Executive Summary_xi

01

서론_1

제1절 연구의 배경 및 필요성	1
제2절 연구의 목적	10
제3절 연구의 범위 및 내용	10
1. 연구의 범위	10
2. 연구의 내용 및 방법론	11

02

해상운임예측모형 구축 현황_13

제1절 전통적인 예측 방법	13
1. 회귀모델 기반 예측 모형	15
2. 시계열 모델 기반 예측 모델	18
제2절 인공지능에 기반한 해상운임 예측 방법	24
1. 선행연구	24
2. 시사점	29
제3절 해상운임예측모델 연구진행 현황	32
1. 기존 보유 모델	32
2. 시사점	35
제4절 기존 연구와 차별성	37
1. 데이터 전처리 방식 개선	37
2. 다양한 알고리즘을 활용한 다수의 예측 모델 구축	37
3. 데이터의 활용성 확대를 통한 예측 정확도 향상	38

03 인공지능기반 해상운임 예측 모델 구축 _39

제1절 개요	39
제2절 모델의 구축	40
제3절 데이터 및 전처리	49
1. 데이터 개요	49
2. 데이터 전처리	51
3. 변동성 분석 적용 결과	56
제4절 인공지능 알고리즘	68
1. DNN	68
2. LSTM	71
3. GRU	73
4. 시사점	75

04 모델별 예측 결과 및 평가 _77

제1절 개요	77
1. 모델 체계	77
2. 모델 검증 방법	78
3. 분류모델 성능 평가지표	80
제2절 주간 예측 모델의 결과	82
1. DNN에 기반한 해상운임 예측 모델 결과	82
2. LSTM에 기반한 해상운임 예측 모델 결과	85
3. GRU에 기반한 해상운임 예측 모델 결과	89
4. 시사점	92
제3절 단기 예측 모델의 결과	93
1. 케이프선	93
2. 파나막스선	97
3. 시사점	100

05

결론 및 정책제언_101

제1절 결론	101
1. 하이브리드 단기 예측 모델 구축	101
2. 다양한 데이터 조합 및 알고리즘 기반 복수 모델의 확보	102
제2절 시사점	103
1. 다양한 방법론을 도입한 예측 모델 필요	103
2. 전망 시점에 따른 방법론 적용의 차별화 필요	103
3. 産-學-研 공동연구 필요	103
4. 뉴노멀에 따른 새로운 변수 필요	105

참고문헌_107

부록_113

표 목차

〈표 1-1〉 IMF 세계 경제 성장을 전망치 수정 내용	4
〈표 1-2〉 건화물선 시장의 구분 기준	11
〈표 1-3〉 기존 모델과 제안 모델의 차별성	12
〈표 2-1〉 해운에서 사용한 분석적 기법	14
〈표 2-2〉 선행연구 요약	30
〈표 3-1〉 3장 구성	39
〈표 3-2〉 데이터 전처리에 사용한 기술분석지표 상세	45
〈표 3-3〉 데이터 전처리 비교	48
〈표 3-4〉 데이터 요약	50
〈표 3-5〉 케이프선 기초통계 요약	50
〈표 3-6〉 파나막스선 기초통계 요약	51
〈표 3-7〉 변동성 분석	54
〈표 3-8〉 제안 모델이 사용한 딥러닝 알고리즘 종류	75
〈표 4-1〉 혼동행렬 구성	79
〈표 4-2〉 DNN기반 모델을 활용한 케이프선 차주 운임의 방향성 예측 결과	82
〈표 4-3〉 케이프선 모델의 혼동행렬	83
〈표 4-4〉 성능지표 비교	83
〈표 4-5〉 DNN기반 모델을 활용한 파나막스선 차주 운임의 방향성 예측 결과	84
〈표 4-6〉 파나막스선 모델의 혼동행렬	84
〈표 4-7〉 성능지표 비교	85
〈표 4-8〉 LSTM기반 모델을 활용한 케이프선 차주 운임의 방향성 예측 결과	86
〈표 4-9〉 케이프선 모델의 혼동행렬	86
〈표 4-10〉 성능지표 비교	87
〈표 4-11〉 LSTM기반 모델을 활용한 파나막스선 차주 운임의 방향성 예측 결과	87
〈표 4-12〉 파나막스선 모델의 혼동행렬	88
〈표 4-13〉 성능지표 비교	88
〈표 4-14〉 GRU기반 모델을 활용한 케이프선 차주 운임의 방향성 예측 결과	89
〈표 4-15〉 케이프선 모델의 혼동행렬	90
〈표 4-16〉 성능지표 비교	90
〈표 4-17〉 GRU기반 모델을 활용한 파나막스선 차주 운임의 방향성 예측 결과	91

〈표 4-18〉 파나막스선 모델의 혼동행렬	91
〈표 4-19〉 성능지표 비교	92
〈표 4-20〉 모델의 명칭	93
〈표 4-21〉 성능지표 비교	95
〈표 4-22〉 성능지표 비교	96
〈표 4-23〉 성능지표 비교	98
〈표 4-24〉 성능지표 비교	100

그림 목차

〈그림 1-1〉 건화물선 지수와 코스피 지수의 변동성 비교	2
〈그림 1-2〉 금융위기로 인한 건화물선 지수 변화	3
〈그림 1-3〉 건화물선 지수의 변동성	5
〈그림 1-4〉 인공지능기반 해상운임 예측 모델개발 진행 상황(2019~현재)	6
〈그림 1-5〉 실제 서비스 진행 현황	7
〈그림 1-6〉 한국은행 전망모델시스템	8
〈그림 1-7〉 한국해양수산개발원 수산부문 전망모델 「KMI-FOSiM」 구조도	9
〈그림 2-1〉 BDI 관측치의 시계열 요소분해법 적용결과(1999.11~2018.10)	19
〈그림 2-2〉 2013년 5월부터 2018년 4월까지의 BDI 자료를 활용한 전방예측 결과 예시	20
〈그림 2-3〉 VAR/VECM 프로세스	24
〈그림 2-4〉 순환신경망 모델	26
〈그림 2-5〉 인공지능기반 모델 개발 연구 진행 상황	32
〈그림 3-1〉 일반적인 인공지능 모델 구축과정	41
〈그림 3-2〉 웨이블릿 분석을 활용한 데이터 특징 추출	42
〈그림 3-3〉 EMD를 활용한 데이터 특징 추출	43
〈그림 3-4〉 기술분석을 활용한 데이터 특징 추출	43
〈그림 3-5〉 DFN를 활용한 데이터 특징 추출	44
〈그림 3-6〉 호드릭-프레스콧 필터를 활용한 데이터 특징 추출	46
〈그림 3-7〉 기술 분석 및 시계열 분해를 활용한 데이터 특징 추출	47
〈그림 3-8〉 모델 구축과정 중 데이터 전처리의 상세과정	48
〈그림 3-9〉 케이프선 변수 일별 변화율	57
〈그림 3-10〉 케이프선 운임 5일간 일중변동성($n=5$)	58
〈그림 3-11〉 케이프선 운임 30일간 일중변동성($n=30$)	58
〈그림 3-12〉 케이프선 스팟운임의 역사적 변동성(2012년~2020년)	60
〈그림 3-13〉 케이프선 스팟운임의 기간별 변동폭 비교(2020년)	61
〈그림 3-14〉 케이프선 스팟운임의 기간별 변동비율 비교(2020년)	62
〈그림 3-15〉 케이프선 스팟운임의 변동성이 범위 내에서 움직일 확률(2020년)	63
〈그림 3-16〉 케이프선 스팟운임과 FFA가격간의 베이스스	66
〈그림 3-17〉 케이프선 FFA가격의 스프레드	67

〈그림 3-18〉 DNN 모델의 구조	69
〈그림 3-19〉 LSTM 셀스테이트(Cell state)의 구조	71
〈그림 3-20〉 LSTM과 GRU 모델 구조 비교	74
〈그림 4-1〉 케이프션 운임 방향성 예측 결과(은닉층 2개, 학습 비율 70:30)	94
〈그림 4-2〉 케이프션 운임 방향성 예측 결과(은닉층 3개, 학습 비율 70:30)	94
〈그림 4-3〉 케이프션 운임 방향성 예측 결과(은닉층 2개, 학습 비율 80:20)	95
〈그림 4-4〉 케이프션 운임 방향성 예측 결과(은닉층 3개, 학습 비율 80:20)	96
〈그림 4-5〉 파나마스선 운임 방향성 예측 결과(은닉층 2개, 학습 비율 70:30)	97
〈그림 4-6〉 파나마스선 운임 방향성 예측 결과(은닉층 3개, 학습 비율 70:30)	98
〈그림 4-7〉 파나마스선 운임 방향성 예측 결과(은닉층 2개, 학습 비율 80:20)	99
〈그림 4-8〉 파나마스선 운임 방향성 예측 결과(은닉층 3개, 학습 비율 80:20)	99
〈그림 5-1〉 해운시황 예측모델 개발 및 서비스	104

약어 목록

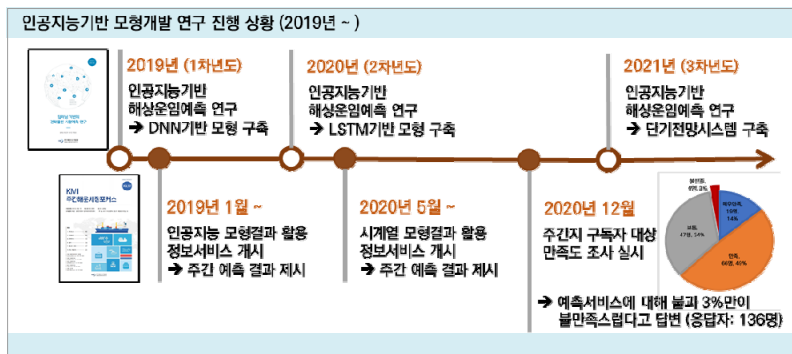
- 3SLS: 3단계 최소제곱법(Three stage least squares)
- AR: 자기회귀모델(Autoregressive Model)
- DNN: 심층신경망(Deep Neural Network)
- EMD: 경험적 모드 분해법(Empirical Mode Decomposition)
- GRU: 게이트 순환신경망(Gated Recurrent Unit)
- IMF: 국제통화기구(International Monetary Fund)
- IMF: 내재모두함수(Intrinsic Mode Function)
- IV: 도구변수(Instrumental Variable)
- LSTM: 장단기메모리 신경망(Long Short Term Memory)
- MA: 이동평균(Moving Average)
- SVM: 서포트 벡터 머신(Support Vector Machine)

요약

1. 연구 목적

- 본 연구에서는 해운기업과 정부가 데이터 기반 의사결정을 할 수 있도록 고품질의 해운 시황 전망 정보 서비스를 제공하기 위해 다양한 인공지능 알고리즘을 활용한 복수 모델을 구축하고자 함
- 한국해양수산개발원 해운빅데이터연구센터는 2013년 7월 22일부터¹⁾ 「KMI 주간해운시장포커스」를 발간하기 시작하여 8년 이상 정부, 해운업계, 금융 분야 등 종사자에게 해운 시황과 관련한 정보를 제공 중임

〈인공지능기반 모델 개발 연구 진행 상황(2018-현재)〉



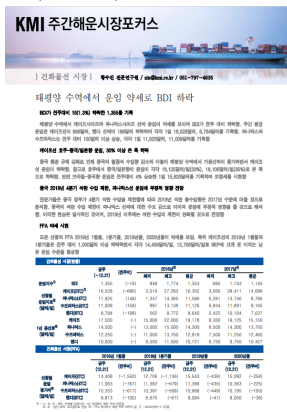
자료: 저자 작성

1) KMI 홈페이지>정기간행물>동향정보>KMI 주간해운시장포커스 게시일 기준

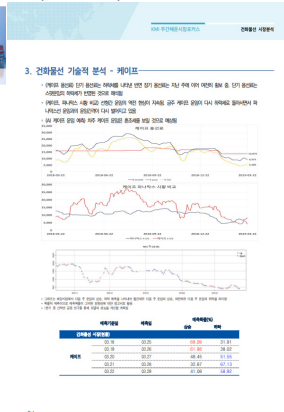
- 2019년 1월에 1차 연도 연구 수행을 통해 개발한 DNN기반 차주 운임 예측 모델 결과를 주간지를 통해 제공하기 시작함
- 2020년 5월부터 2020년 연구 결과로 전통적 계량 분석 방법을 활용한 VAR/VECM 모델 결과를 추가하여 제공 중임

〈「KMI 주간해운시장포커스」 서비스 개편 현황〉

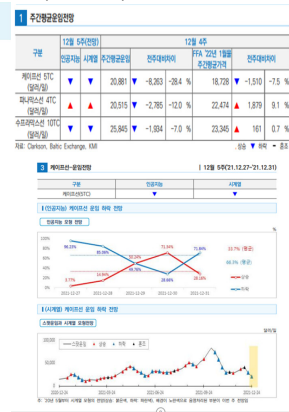
(a) 기존 서비스
(2013.7~)



(b) DNN 모델 서비스 시작
(2019.1~)



(c) 계량 모델 서비스 시작
(2020.5~)



자료: 저자 작성

- 다양한 인공지능 알고리즘을 활용한 모델을 구축하여 기존에 보유하고 있는 모델 품질을 개선하고 새로운 방법론을 도입하여 예측력을 향상하고 단일 모델에 대한 의존도를 낮추어 고품질 서비스를 제공하고자 함
- 인공지능의 대표 알고리즘인 DNN, LSTM, GRU를 활용해 모델을 개발함
- 현재 주간지는 차주의 운임 전망을 상승과 하락 등 분류 형태로 제공 중이며, 이 점을 고려하여 모든 모델을 분류모델로 구축해 기존 모델보다 예측력이 향상되었는지 검증을 진행함

- 본 연구를 통해 모델의 예측 기간을 최대 분기까지 확장하여 정부, 해운기업 등 사용자 관점에서 정보 활용도를 보다 높이고자 함

2. 연구특징 및 방법

1) 다양한 알고리즘을 활용한 복수의 모델 구축

- 모델은 기본적으로 분류 형태로 인공지능 대표 알고리즘인 DNN, LSTM, GRU를 사용하여 구축함

〈제안 모델이 사용한 딥러닝 알고리즘 종류〉

항목	개념		특징
	공통점	개념	
DNN	• 대표적인 딥러닝 알고리즘 계열임 • 알고리즘 구조가 입력층과 출력층, 그리고 다수의 은닉층으로 구성됨	• 다수의 층 구조로 이루어져 있으며, 순환하는 개념은 포함하지 않음	• 가장 먼저 접하게 되는 인공지능 알고리즘임
LSTM		• 순환신경망 형태로 이전 시점의 정보 활용	• 시계열 데이터 학습을 위해 구조가 설계
GRU		• LSTM과 유사하나, 계산 성능 향상으로 실제 사용에 적합	

자료: 저자 작성

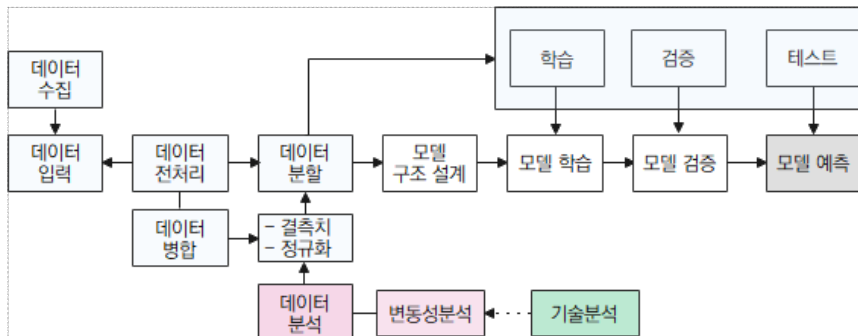
2) 데이터 전처리 과정에 변동성 분석을 추가

- 데이터 전처리란 분석하기 앞서 데이터를 적합한 형식으로 처리하는 과정을 말하며, 데이터의 결측치, 이상치 등 정제하는 과정을 통해 모델 성능을 향상시키는 필수 단계임²⁾³⁾

2) 이우성·오성준·안덕호(2018), p. 27.

- 인공지능 모델을 구축하는 과정에 데이터가 표준화된 형태로 존재하지 않기 때문에 기본적으로 이상치, 결측치 등 데이터를 정제하고 형태를 통일시키는 데이터 전처리 작업을 필수적으로 진행함
- 인공지능 알고리즘은 변수 가중치와 편향 크기를 조정하여 실제값과 예측값과의 오차범위를 줄여나가며, 이 과정에서 알고리즘이 자체적으로 데이터의 특징을 추출하거나 분석하지 못함
- 본 연구에서는 데이터 전처리 과정에서 데이터 특징을 추출하는 과정을 변동성 분석으로 모델의 예측성능을 높이하고자 함
 - 2019년부터 개발한 모델에는 이와 같은 알고리즘의 작동 방식을 고려해 데이터 전처리 과정에 데이터의 변동성을 분석하는 단계를 추가함
- 자체적으로 보유하고 있는 모델은 기술분석 방법을 이용해 데이터의 특징을 추출했으나, 본 연구에서는 전하물선 시장 변동성이 크다는 점에 주목하여 변동성 분석을 새롭게 도입함

〈제안 모델의 데이터 전처리 과정 상세〉



자료: 저자 작성

〈데이터 전처리의 차별화〉

구분	데이터 전처리		
	기술분석	시계열 성분분해	변동성 분석
방법론	이동평균비율, MACD, Signal, 오실레이터, 모멘텀, 변화율, 상대강도지수, SLOW%K, SLOW%D, 백분위, 회귀계수, 로그수익률, 표준편차, 윌리엄스 %R,	시계열의 데이터 성분 분해 추세, 주기, 계절성, 랜덤	로그수익률, 기간 일중변동성 역사적 변동성, 일간변동성, 변동폭, 변동비율, 변동성 범위 내에서 움직일 확률, 변동 범위, 확률밀도함수, 베이스스, 스프레드, 윌리엄스 %R,
2019	운임	-	-
2020	-	운임	-
2021	-	-	운임, FFA

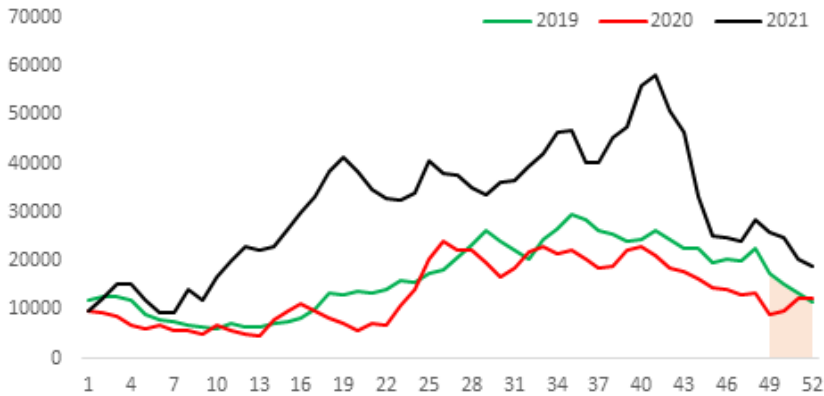
자료: 저자 작성

3) FFA 데이터의 활용성 제고

- 모델의 예측성능 향상을 위해 또 다른 접근법으로 FFA 데이터 활용성을 대폭 강화하여 차주 운임과 기간 확장 시 예측력 향상을 도모함
 - FFA 가격은 미래 시장에 대한 해운시장 참여자에 기대 심리 변화를 분석하는데 유용한 데이터임
 - 예시로 제시한 케이프선 FFA 가격 추이 변화를 보면, 2019년 1월 브라질 댐 붕괴사고와 12월 코로나19 확산 시점에 가격이 즉각적으로 변화하는 과정을 확인할 수 있음
- 데이터 전처리 방법에서 FFA 특징을 활용할 수 있는 베이스스와 스프레드 분석을 추가했으며, 변동성 분석을 통해 특징을 추출함
 - 기존 모델은 운임에만 기술분석과 같은 데이터 전처리를 적용함

〈케이프선 FFA 익월물 가격 변화(2019-2021)〉

(a) FFA 1개월 후의 주간 평균 가격 변화 추이



(b) 특정 시점에서의 FFA 개월물의 일일 가격 변화 추이



자료: 저자 작성

4) 모델 예측 결과 안정성 검증을 위한 다양한 검증 방법 수행

- 정확도를 포함하여, 재현율, 정밀도와 F1을 통해 결과에 안정성이 확보된 모델을 최종 선택함
- 모델 구축과정에서 사용하는 학습, 검증, 테스트 셋 데이터 외에 사용하지 않은 추가적인 테스트 셋 데이터를 사용하여 모델의 성능을 한 번 더 검증하여 과적합과 과소 적합을 사전에 방지함
 - 일반적으로 모델 검증은 검증과 테스트 셋을 이용해 구축된 모델의 성능을 평가하는 데에서 마무리되어 실전에 사용하면 예측성능이 떨어지는 경우가 발생함
 - 검증 과정을 한 번 더 수행하여 실제 모델을 활용해 운임을 예측하고 서비스하는 과정에서 결과에 품질을 일정 수준으로 확보하고자 함

2. 연구 결과

1) 연구 결과 요약

- 예측성능과 기간을 확장한 복수 모델을 구축하여 제시함
 - DNN, LSTM, GRU의 차주 운임 예측 모델을 구축함. 특히 DNN은 1차 연도에 개발한 모델보다 높은 예측력을 보여 줌
- 제안한 차주 운임 예측 모델은 케이프선이 80.0%, 파나막스가 86.67%로 기존 모델보다 예측성능이 향상됨
 - 1차 연도 모델의 정확도는 케이프선이 70.9%, 파나막스선이 72.3%임
 - 변동성 분석이라는 새로운 데이터 전처리 방법을 도입하고 FFA 가격 활용을 확대한 것도 예측성능을 향상한 원인인 것으로 분석함

-
- 다수의 변수를 사용해 모델을 구축할 때는 DNN이 적합한 것으로 판단함
 - 주어진 데이터에 반복적 학습을 통해 변수별 가중치 크기를 조정하는 DNN은 작동 방식으로 인해 다수에 변수 중 변수 차원을 감소시키는 역할을 한 것으로 판단함
 - 반면 시계열 알고리즘은 알고리즘의 작동 방식에서 주어진 입력변수를 순차적으로 학습하여 속도가 상대적으로 느리며, 학습 횟수별 동일 크기에 가중치가 적용되면서 예측 결과가 DNN과 비교해 낮은 것으로 분석함
 - DNN에 기반한 모델을 통해 운임 예측 기간을 12주까지 확장함
 - 구축 모델을 활용해 케이프션 12주 후 운임을 예측하였을 때 60.67%에 정확도를 보였으며, 파나마스선은 82.22%로 나타남
 - 정확도를 비롯하여 전체적인 모델 성능지표로 평가했을 때 파나마스선 모델 정확도는 82.22%로 높게 나타났지만, F1의 비율이 현저히 낮게 나타나 실제 서비스로 진행하기에 추가적인 보완이 필요함
 - 예측 기간을 7주까지 늘렸을 때 높은 예측 정확도와 안정된 예측 결과를 도출함

〈2018년 이후 연구 진행 상황과 이번 연도 제안 모델의 차별성〉

구분	1차 연도 (2019년)	2차 연도 (2020년)	3차 연도 (2021년)
개발모델	DNN을 활용한 해상운임예측 모델 (분류모델)	LSTM을 활용한 해상운임예측 모델 (회귀모델)	인공지능기반 복수 모델 구축 (분류모델)
특 징	모델	해운 데이터의 시계열적 특 성을 이용한 모델 구축	① 복수모델 개발 ② 예측 기간 확장
	데이터	• 가격 데이터 활용 (단위: day) • 현물 : 운임, 용선료. 선물 : FFA 가격	• FFA 가격 +2월물 추가
	전처리	운임만 기술분석 활용	변동성 분석 도입
		① 운임의 시계열 형태 (components of time series)를 분해 해 특징 추출	① 데이터변동성 추출 ② FFA도 변동성 분석을 통해 데이터 특성 추출
	알고 리즘	DNN(Deep Neural Network)	LSTM(Long-Short Term Memory)
의의	해상운임예측 분야에 인공 지능을 적용한 모델을 개발 하고 그 결과를 서비스함	해운데이터의 시계열적 특 성을 활용해 예측모델을 개 발하고 추가하였음	복수모델에 기반한 운임 예 측 능력 향상 및 기간 확장
한계	• 단일 모델	• 인공지능 모델이기는 하 지만, 시계열 모델의 한 계를 포함함	• 실제 모델을 구축하고 서 비스하는 과정에서 나타 난 한계점(결과 상이, 예 측 기간) 보완
활용	• 기존 주간지에 주간 예측 부분을 추가하여 제공	• 2020년 5월 두 개의 모 델(인공지능, 시계열)에 대한 예측 결과를 서비스 하기 시작했으며, 예측 결과가 다른 경우 발생	• 복수모델 구축을 통해 단 일 모델의 편향 발생 가 능성을 보완 • 예측 기간 확장 • 향후 기존 분기 보고서에 예측 부분을 추가하여 제공

자료: 저자 작성

2) 제언

- 다양한 방법론을 도입한 예측 모델
 - 실제 모델을 예측에 적용하여 활용하면 실험과 달리 예측 성능이 크게 떨어지는 경우가 비일비재함
 - 최신 기법과 더불어 전통적인 방법론을 기반으로 한 다수의 모델을 확보하여 예측 성능을 수시로 확인하고 성능을 향상할 수 있도록 지속적인 노력을 기울여야 함
- 전망 시점에 따른 방법론 적용 차별화
 - 데이터 생성 주기와 인공지능 및 전통 방법론의 알고리즘 작동 방식에 장·단점에 따라 현실 반영을 높이고 전망 시점에 맞는 복수 모델 구축이 필요함
- 산·학·연 공동연구 추진
 - 코로나19 시대로 인해 정부 및 해운기업의 적극적인 소통이 어려웠던 점을 고려했을 때 다양한 형태로 정부와 해운기업에 요구를 파악할 수 있는 추진 체계 마련이 필요함
 - 현장의 소리와 함께 새롭게 개발되고 있는 기술과 개선되고 있는 전통적인 방법론 등 이론적인 부분에 전문지식을 공유할 수 있는 공동연구가 추진되어야 함

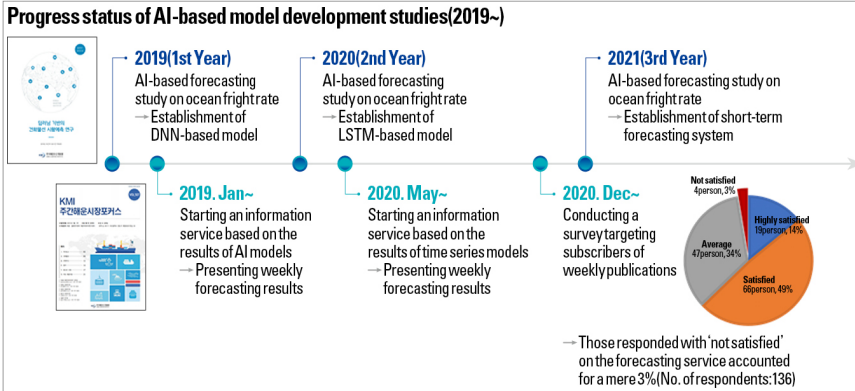
EXECUTIVE SUMMARY

1. Purpose

- This study aims to establish multiple prediction models with various AI algorithms in order to provide the high-quality forecasting information service for shipping market conditions, allowing shipping companies and the government to make data-based decisions.
- Since its first publication of 「KMI Weekly Shipping Market Focus」 on July 22, 2013⁴⁾, the Maritime Big-data Analysis Center of the Korea Maritime Institute has been providing information pertaining to shipping market conditions for various market players such as the government, shipping industry and financial area for more than 8 years.

4) KMI Homepage>Periodic publications>Trend information>KMI Weekly Shipping Market Focus, Based on posting date

〈Progress status of AI-based model development studies (2018~Present)〉



Source: Prepared by the author

- The weekly publication began to provide the result of subsequent week's freight rate prediction model based on DNN, which was developed through the 1st year study conducted in January of 2019.
- Since May 2020, the weekly publication has included the result of the VAR/VECM model, the traditional quantitative analysis method, which was established in the 2020 research.

(*) Γ is a $\mathbb{Q}$ -linear combination of the following elements:

[illegible][illegible]

11 주안평균종가상행진

구분	12월 5주(12월)		12월 4주	
	주당 수익률	주당 평균종가	12월 19일 수익률	12월 18일 종가
코스피지수 STC (코스피)	▼	20.881	-8.363 -28.4 %	18.725 -1.510 -7.5 %
코스닥지수 STC (코스닥)	▲	20.915	-12.785 -12.0 %	22.244 1.879 9.1 %
주요지수 STC (KOSPI, KOSDAQ)	▼	25.845	-1.804 -7.0 %	23.365 ▲ 161.0 7.0 %

RISK: American, Batic, Exchange, EMR

 12월 1주 : 12월 4주 : 12월 5주

3. 주안평균종가상행진 12월 5주(12.17.21.27.31.12.31)

구분	주당 수익률	주당 평균종가
코스피지수(STC)	▼	

1. 코스피지수(코스피)를 중심으로 주안 평균종가상행진

12월 1주 : 12월 4주 : 12월 5주

1. 코스피지수(코스피)를 중심으로 주안 평균종가상행진

12월 1주 : 12월 4주 : 12월 5주

2. 코스피지수(코스피)를 중심으로 주안 평균종가상행진

12월 1주 : 12월 4주 : 12월 5주

-
- This study aims to enhance the applicability of information from the perspective of users such as the government and shipping companies, by extending the prediction period of models up to the quarter.

2. Methodology and Feature

1) Establishing multiple models utilizing various algorithms

- Models were established basically in the classification form by using DDN, LSTM and GRU, the representative AI algorithms.

〈Type of deep-learning algorithms used by suggested models〉

Item	Category		Characteristics
	Common features	Concept	
DNN	• A representative deep-learning algorithm • The algorithm structure consists of input layer, output layer and a number of hidden layers.	• This model consists of a multiple layers but excludes the concept of recurrent	• The first AI algorithms to come across
LSTM		• In the form of recurrent neural network, this model utilizes information at the prior time	• The structure is designed to learn time-series data
GRU		• Similar to LSTM, but this model is more suitable for actual utilization due to improved performance in calculation	

Source: Prepared by the author

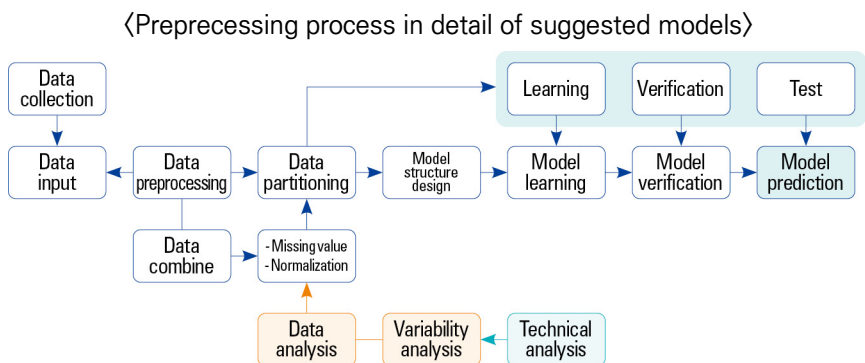
2) Adding the variability analysis to the data preprocessing process.

- Data preprocessing refers to the process of processing data into an appropriate format prior to analysis and is the essential step to improve the data analysis performance of models through the process of filtering missing values and abnormalities, etc.⁵⁾⁶⁾
- Since data is not present as a standardized format even in the process of establishing AI models, it is essential to carry out the data preprocessing work which filters data such as abnormalities and missing values and unifies formats.
- AI algorithms adjust the weight of variables and the size of bias to reduce the margin of error between actual values and predicted values, and in this process the algorithms fail to extract or analyze the features of data.
- This study aims to enhance the predictability of models by introducing a new methodology called the variability analysis to the process of extracting the features of data in the data preprocessing process.
- With such operation method of the algorithms considered in all the models developed since 2019, the step of extracting the features of data was added to the data preprocessing process.

5) Lee Woo-sung, Oh Seong-Jun, Ahn Deok-Ho (2018). pp. 27.

6) Yang Hyun-Sik, Im Yoo-Bin, Lee Jin-Young, Kim Dong-Hyo, Kim Se-Jung, Kim Hying-Joo (2019), (pp. 438-439.)

- While our existing models extracted the features of data by the technical analysis method, this study newly introduced the variability analysis with a focus on the high variability of the dry bulk shipping market.



자료: 저자 작성

〈Differentiation of data preprocessing〉

Classification	Data Preprocessing		
	Technical Analysis	Time-Series Decomposition	Variability Analysis
Methodology	Moving average, MACD, Signal, Oscillator, Momentum, Change rate, Relative strength index, SLOW%K, SLOW%D, Percentile, Recession coefficient, Log returns, Standard deviation, Williams %R	Time-series data decomposition trend, cycle, seasonality, random	Log returns, Period, Daily variability, Historical variability, Width of fluctuation, Variable ratio, Probability of moving within the range of variability, Range of fluctuation, Probability density function, Basis, Spread, Williams %R
2019	Freight rate	-	

Classification	Data Preprocessing		
	Technical Analysis	Time-Series Decomposition	Variability Analysis
2020	Freight rate		–
2021	–		Freight rate, FFA

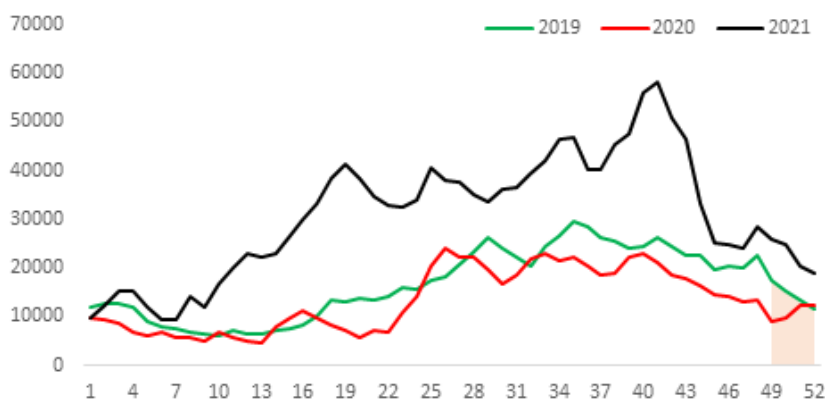
Source: Prepared by the author

3) Improving the applicability of the FFA data

- As another approach to improve the predictive performance of models, this study sought to improve the prediction capability when predicting freight rates for the subsequent week and extending the prediction period, by significantly enhancing the applicability of the FFA data.
- The FFA price is useful data to analyze the change of shipping market players' expectations for the future market.
- Looking at changes in the FFA price trend of Capesize ships presented as an example, it can be found that the price changed immediately both at the time of the Brazil dam collapse accident in January, 2019 and at the time of the spread of the COVID-19 pandemic in December of the same year.
- This study added the Basis and Spread analysis which can utilize the features of FFA in the data preprocessing method, and extracted the features through the variability analysis.
- In the existing models, the data preprocessing method such as the technical analysis was applied only to freight rates.

〈FFA price trend of Capesize ships for the following month contracts
(2019~2021)〉

(a) Trend of FFA weekly average price change after one month



(b) Trend of daily price change of FFA monthly contract at a specific point



Source: Prepared by the author

3) Utilizing various verification methods to verify the stability of model prediction results

- This study finally selected the models in which the stability of results was secured via recall, precision and F1 as well as accuracy.
- This study verified the performance of models once again with additional untapped test-set data, other than the learning, verification and test-set data used in the process of establishing models to prevent both overfitting and underfitting in advance.
- Since the verification of models is generally completed by evaluating the performance of the established models with the verification and test set, when used in practice, the predictive performance of models may be deteriorated.
- This study sought to secure the quality of prediction results to a certain level in the process of predicting and servicing freight rates with the actual models by carrying out the verification process once again.

2. Results

1) Summary

- This study established and provided multiple models with the extended predictive performance and periods.

-
- It established the subsequent week's freight rate prediction models using DNN, LSTM and GRU. Specifically, DDN showed the higher prediction capability than the models developed in the 1st year.
 - With 80.0% for Capesize ships and 86.67% for Panamax ships, the proposed prediction models were shown to have their predictive performance improved, compared to the existing models.
 - The accuracy of the 1st-year models was 70.9% for Capesize ships and 72.3% for Panamax ships.
 - It was also analyzed that introducing the new data preprocessing method called the variability analysis and expanding the utilization of the FFA price led to the improvement of predictive performance.
 - This study determined that DDN is appropriate when establishing models with present and multiple variables in the process of establishing models using various algorithms.
 - It was determined that DNN, which adjusted the size of weight by variables via the iterative learning of given data, played a role in reducing the dimension of variables among multiple variables due to its operation method.
 - On the other hand, it was analyzed that the time-series algorithm had a relatively low speed by learning the given input variables in its operation method and its prediction results were lower than those of DDN as the weight of the same size was applied by the number of learning times.

- In this study, the freight rate prediction period was extended to 12 weeks through the DNN-based model.
- When predicting freight rates after 12 weeks for Capesize ships, the accuracy of the established models was shown to be 60.67%, and 82.22% as for Panamax ships.
- As a result of evaluation with the performance indicators of overall models, including accuracy, the accuracy of the Panamax model was shown to be high at 82.22%, but additional supplementation is required for such accuracy to be realized as actual service.
- Higher accuracy and stable prediction results were created when the prediction period was extended up to 7 weeks.

〈Research progress status after 2018 and differentiation of this year's suggested models〉

Classification		1 st Year (2019)	2 nd Year (2020)	3 rd Year (2021)
Development model		An ocean freight rate prediction model using DNN (Classification model)	An ocean freight rate prediction model using LSTM (Regression model)	AI-based multiple models (Classification model)
Features	Model	Construct a model based on technical analysis by utilizing the fact that shipping data is highly vulnerable	Establish a model by utilizing time-series characteristics of shipping data	① Develop multiple models ② Extend prediction period
	Data	• Utilize price data (unit: day) • Spot: freight rates, chartered frights, Futures: FFA price		• FFA price + add 2 months contract

Classification		1 st Year (2019)	2 nd Year (2020)	3 rd Year (2021)
	Preprocessing	Utilize freight rates for technical analysis	Utilize freight rates for technical analysis and time-series decomposition	Introduce variability analysis
			① Extract characteristics by decomposing the components of time series of freight rates	① Extract data variability ② Extract characteristics of data for FFA through variability analysis
	Algorithm	DNN(Deep Neural Network)	LSTM(Long-Short Term Memory)	DNN, LSTM , GRU
Significance		Develop AI-applied models to ocean freight rate forecasting and provide the results	Develop and add to the model by utilizing time series characteristics of shipping data	Boost the predictability based on multiple models and extend the period
Limitation		• A single model	• AI-based model but it contains limitations of time-series model	• Supplement the limitations (difference in results, prediction period) arising in the process of construction and service of the model
Utilization		• Provide the service after adding weekly forecasting section in existing weekly publications	• Forecasting results have been provided since May of 2020 generated from two models (AI and time series) but there were cases of showing different prediction results	• The establishment of multiple models can make up for the potential issues resulting from distortion from a single model.

Classification	1 st Year (2019)	2 nd Year (2020)	3 rd Year (2021)
			• Extend the prediction period • Add 'prediction' section in existing quarterly reports in the future

Source: Prepared by the author

2) Policy suggestions

- Prediction models introducing various methodologies
 - When the actual models are applied and utilized for prediction, it is not uncommon that their predictive performance becomes significantly lower than what is shown in the experiment.
 - It is required to verify the predictive performance by securing the large number of models based on the conventional methodologies along with the latest techniques, and to make continuous effort to improve the performance of the models.
- Differentiation of methodology application with the point of forecast time
 - It is required to establish multiple models appropriate for the point of forecast time which reflects actual situations more, depending on the data creation cycle, the algorithms of AI and conventional methodologies and the pros and cons of their operation methods.

-
- Promoting industry-academia-institute joint researches
 - Given that active communication with the government and shipping companies has been difficult due to the COVID-19 pandemic, it is required to prepare a system to identify the needs of the government and shipping companies in various forms.
 - It is required to promote joint researches which could provide opportunities to share specialized knowledge in theoretical areas such as new technologies being developed and the conventional methodologies being improved, along with voice from the field.

01

서론

제1절 연구의 배경 및 필요성

모든 산업 전반에 걸쳐 데이터 중심의 의사결정에 대한 중요성이 그 어느 때 보다 강조되고 있다. 이는 최근 정보통신기술 발달로 데이터의 실시간 수집 및 분석 가능한 환경이 조성되었기 때문이다.

Foster and Tom(2013)¹⁾은 의사결정자가 하나의 방식에만 의존하지 않고 경험, 직관과 데이터 분석과 같은 다양한 방식을 조합하여 의사를 결정하는 행위를 데이터 중심 의사결정(DDDM, Data-driven decision-making)이라고 정의한다. 이 방법이 특정 현상의 패턴 분석과 전망을 가능하게 하므로 기업 경영을 위한 전략 수립 및 예외적인 상황 발생 시에 적절한 대비 및 대응책 마련에 도움을 준다는 사실을 실증분석을 통해 입증한 바 있다.

이처럼 데이터 중심 의사결정은 데이터 분석을 통해 확보된 객관적 자료를 경험과 직관을 이용해 해석함으로써 기업이 다양한 시나리오와 그에 따른 전략 수립을 가능하게 한다. 이는 비단 기업에만 긍정적인 영향을 주는

1) Foster & Tom(2013), pp. 51-59.

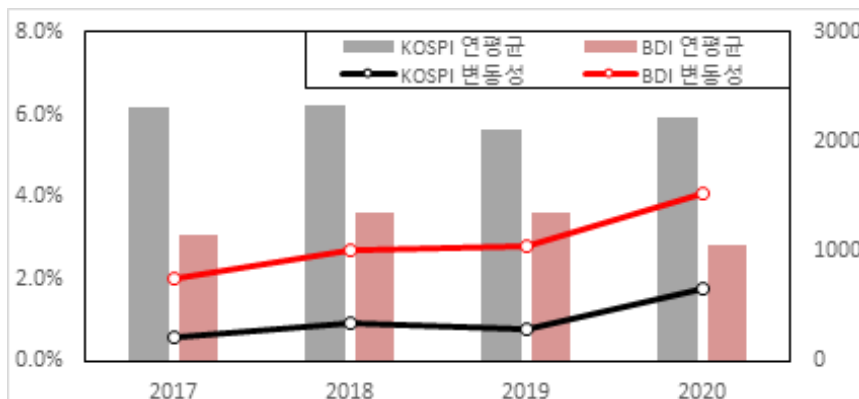
것은 아니며, 정부, 개인 등이 합리적 의사결정을 내리는 데에도 도움을 준다.

이 같은 의사결정 방식은 해운 분야에서도 중요하게 인식되며, 특히 데이터 분석을 위한 기초자료로 활용하기 위하여 해운 시황 전망과 관련한 정보 수요가 증가하면서 해상운임 예측 모델 개발 수요가 증가하고 있다.

해운 기업은 향후 수익 실현 가능성을 추정한 결과를 바탕으로 용대선 비율, 선박 투자 여부 및 기간 조정과 같은 기업의 미래수익과 직결되는 시장 포트폴리오 구축에, 정부는 해운 청사진에 따른 정책 이행 시점과 강도를 판단할 때 해운 시황 전망과 관련한 정보를 활용하기 때문이다.

특히 건화물선 시장은 <그림 1-1>에서 보는 것처럼 주식시장보다 변동성이 큰 시장으로 운임 예측이 쉽지 않다. 큰 변동성으로 인해 예측이 쉽지 않았던 대표적인 사례로 금융위기를 꼽을 수 있는데, 2008년 리먼 브라더스 투자은행 파산을 계기로 촉발된 금융위기가 전 세계 경기를 한순간에 침체 국면에 들어서게 했으며, 당시 초호황을 누리던 해운시장도 예외는 아니었다.

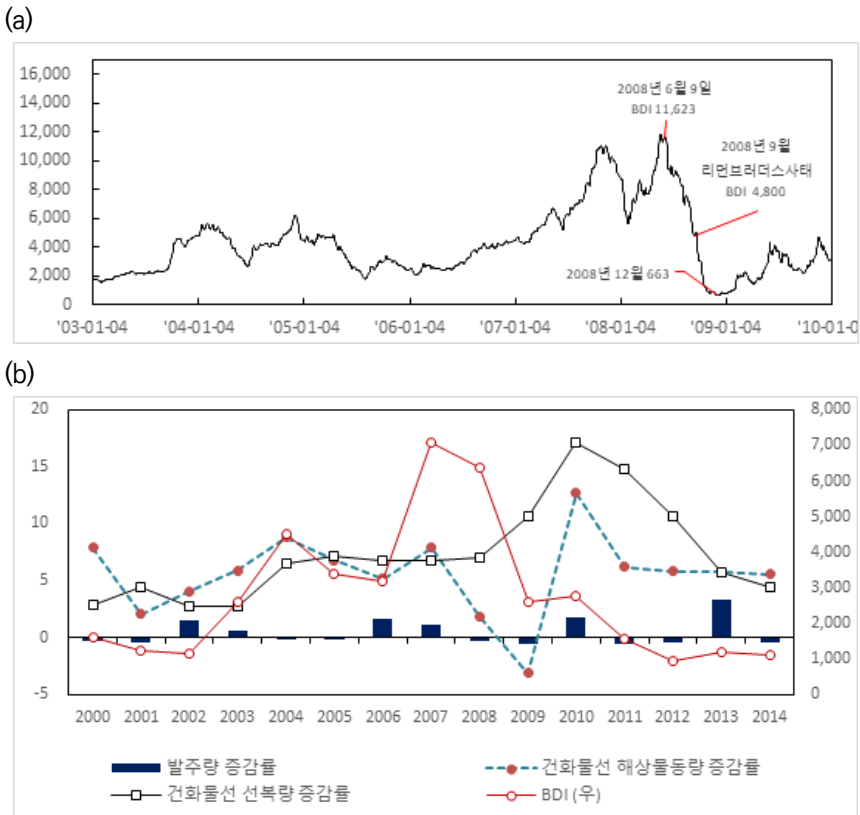
<그림 1-1> 건화물선 지수와 코스피 지수의 변동성 비교



주: 연별 변동성은 매년 일별 지수 변화율의 표준편차임
자료: 클락슨, 야후(2021. 7. 1)의 자료를 바탕으로 저자 작성

〈그림 1-2〉에서 확인할 수 있듯이, 세계 금융위기가 발생하기 직전인 2000년부터 2007년까지 BDI가 연평균 24% 상승할 정도로 건화물선 시장은 초호황기였다. 대부분 기업은 이러한 운임 상승세가 이어갈 것으로 전망하고 최고가를 갱신 중이던 운임과 용선료, 선가 수준에 맞춰 용선 비용을 늘리거나 선대 확충계획을 진행했다.

〈그림 1-2〉 금융위기로 인한 건화물선 지수 변화



주: 클락슨의 건화물선 운임지수를 이용하여 변동성을 계산하여 나타내었으며, 연별 변동성은 매년 일별 지수 변화율의 표준편차임
 자료: 클락슨(2021. 7. 1)의 자료를 바탕으로 저자 작성

2008년 6월 9일 BDI가 11,623을 기록할 정도로 강세시장이었던 건화물선 시장은 같은 해 12월 최고치 대비 94.3% 폭락한 663을 기록할 정도로 시황이 악화하였다. 이처럼 세계 금융위기로 건화물선 시장은 공급이 수요를 상회하는 상태가 장기가 지속되었고, 결국 해운기업들은 경영 위기를 맞게 된다.

IMF(국제통화기금, International Monetary Fund)는 매년 세계경제성장률 전망치를 발표한다. <표1-1>에서 보는 것처럼 2008년 실제 세계경제성장률은 3.0%, 2009년은 -0.6% 기록했으나, IMF가 2007년 7월에서 2008년 1월까지 4%대의 경제성장률을 전망했었다는 점을 미루어보면 당시 세계 금융위기가 어느 정도 예상 밖의 사태였는지 확인할 수 있다.

<표 1-1> IMF 세계 경제 성장률 전망치 수정 내용

단위: %

전망연도	전망치					실제	
	2007		2008				
전망시기	'07.7	'07.10	'07.7	'07.10	'08.1	'08	'09
세계	5.2	5.2	5.2	4.8	4.1	3.0	-0.6

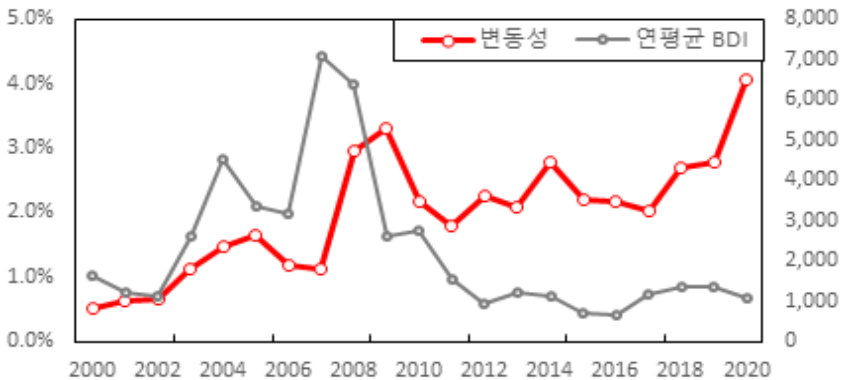
자료: IMF July 07; October 07; January 08; April 08; October 08(2021.12.31.)의 자료를 바탕으로 저자 작성

이외에도 유럽 재정위기, 중국 금융 불안, 브렉시트 등과 같은 대내외 경제 변화의 여파로 건화물선 지수의 변동성이 확대되었으며 국내외 해운업체는 몇 차례 위기 상황을 경험하게 된다.

이처럼 위기를 통해 정부와 기업은 해운시장의 향후 전개 방향을 전망하고 상황에 따른 대응 방안 마련으로 시장에 노출되는 위험을 줄이고자 노력하고 있다. 더욱이 최근에는 <그림 1-3>에서 보는 바와 같이, 2019년 말부터 예상치 못하게 발생한 코로나19가 전 세계적으로 확산되면서 시장의 불확실성도 함께 확대되고 있다.

이렇듯 불안정하고 불확실한 상황 속에 정부와 기업은 데이터 기반한 의사결정에 대한 필요성이 커지고 있으며, 해상운임 예측 모델에 대한 수요가 증가하고 있다.

〈그림 1-3〉 건화물선 지수의 변동성



주: 클락슨의 건화물선 운임지수를 이용하여 변동성을 계산하여 나타내었으며, 연별 변동성은 매년 일별 지수 변화율의 표준편차임

자료: 클락슨(2021. 7. 1)의 자료를 바탕으로 저자 작성

또 하나의 시대적 흐름은 2016년부터 4차 산업혁명이라는 용어와 함께 인공지능, 빅데이터 등이 주요 핵심기술로 부상하고 있다는 점이다. 한국해양수산개발원 해운빅데이터연구센터는 2019년부터 인공지능기반 해상운임 예측 모델을 구축하기 위한 연구를 수행 중이며(〈그림 1-4〉 참고), 연구 결과를 일반연구보고서 형태로 매년 발간하여 연구 내용을 공유하고 있다.

〈그림 1-4〉 인공지능기반 해상운임 예측 모델개발 진행 상황(2019~현재)

인공지능기반 운임예측 연구 진행 상황



자료: 저자 작성

본 연구는 2019년부터 계속된 사업으로, 해운 시황 전망 모델 구축에 인공지능 방법론을 적용하고 그 결과를 실제 정책실무자 및 해운 기업이 활용 가능한지 가능성을 모색하고 있다.

1차 연도(2019년)에 구축한 모델은 DNN 알고리즘을 활용한 분류모델이며, 차주의 운임을 예측하도록 개발한 후 주간지를 통해 매주 실제 서비스를 진행 중이다. 1차 연도에 구축한 모델은 기술 분석을 활용해 데이터의 특성을 추출하는 단계를 추가하도록 설계했으며 그 결과 DNN을 활용해 예측했을 때보다 향상된 예측성능을 보였다.

2차 연도(2020년)에는 1차 연도에 구축한 DNN 알고리즘이 해운데이터가 가지고 있는 시계열적 특성을 충분히 활용하지 못하였다는 한계를 보완하기 위해 LSTM을 적용하여 차주 운임을 예측하는 회귀모델을 구축하였다.

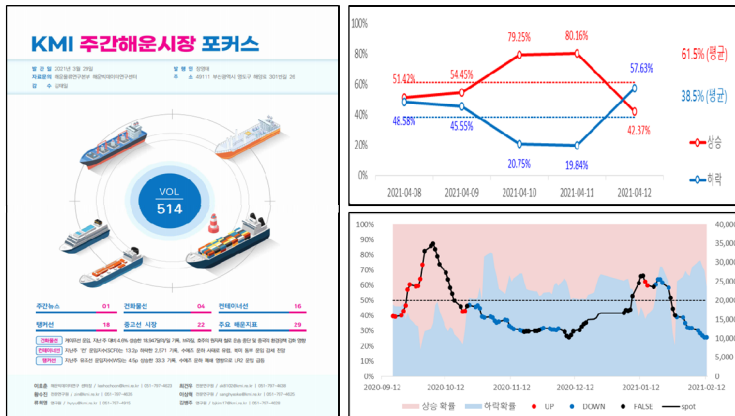
데이터 전처리에 시계열 분해법을 추가하여 해운 데이터 특징을 충분히 활용하고자 모델을 설계하였다. 그 결과 기술분석과 함께 시계열 분해로

데이터 특징을 추출해 입력변수로 활용한 모델이 기술분석만 사용한 모델보다 우수한 예측성능을 보이는 것을 확인하였다. 다만, DNN과 LSTM 모델 모두 예측성능에는 큰 차이를 보이지 않았다.

실제 서비스에 활용하는 모델은 1차 연도에 개발한 DNN 알고리즘 기반 분류모델로 차주 시황의 예측 방향을 전망하여 제시하고 있다. 2차 연도 연구에서 개발한 회귀모델은 직접 운임을 예측하는 회귀분석을 채택하여 운임의 변동범위를 제시하는 만큼 서비스를 진행하기에 앞서 추가적인 검토가 필요하며, 본 연구에서 같은 분류모델로 구축함으로써 보다 비교를 용이하도록 하였다. 모델의 예측 결과물은 <그림 1-5>에서 처럼 본 센터에서 매주 발간하고 있는 주간지인 'KMI 주간해운시장포커스'를 통해 매주 시황의 예측 결과서비스를 제공 중이다.

〈그림 1-5〉 실제 서비스 진행 현황
운임예측모형 활용 현황

» 정부, 해운업 종사자 등을 대상으로 서비스 제공 중



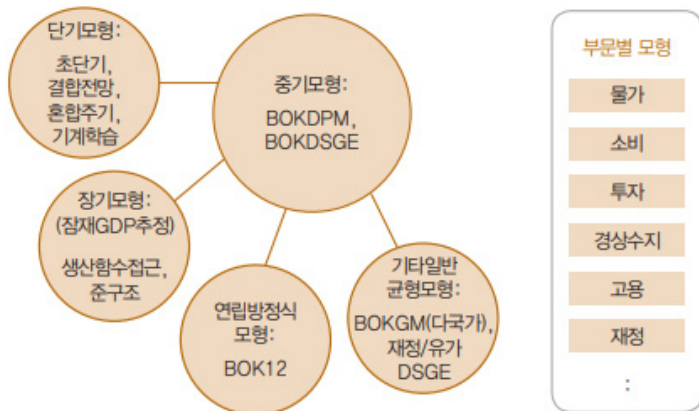
자료: 저자 작성

지금까지 1, 2차 연구는 인공지능 방법론을 해운 분야에 적용과 활용 가능한지 판단하기 위해 진행되었다. 이를 위해 DNN과 LSTM 등 인공지능 알고리즘에 기반해 차주 해상 운임을 직접 또는 상승과 하락으로 분류하여 예측하는 모델을 구축하였으며, 그 결과를 정부와 해운기업 등에 서비스하였다.

본 연구는 모델 결과의 활용성을 높이는 데 주안점을 두고 예측 기간 확장 및 예측 정확도 향상을 목적으로 하며, 이를 위해 복수의 모델을 확보하고자 한다. 복수의 모델이란 다양한 방법론에 기반한 모델을 구축하는 것을 의미하며, 전통적인 방법론, 기존 방법론의 개선과 재발견, 새로운 방법론의 등장 시에 이를 활용하여 모델을 추가로 구축한다. 이러한 복수 모델 확보는 모델 예측 정확도 향상, 기간 확장과 기존 모델의 유지 및 보수를 가능하게 한다. 여기서 몇 가지 사례를 소개하고자 한다.

이승윤·이은경·한진현(2017)²⁾은 한국은행의 기존 전망모델시스템에 대한 설명과 추가로 새롭게 개발한 모델을 제안하였다.

〈그림 1-6〉 한국은행 전망모델시스템



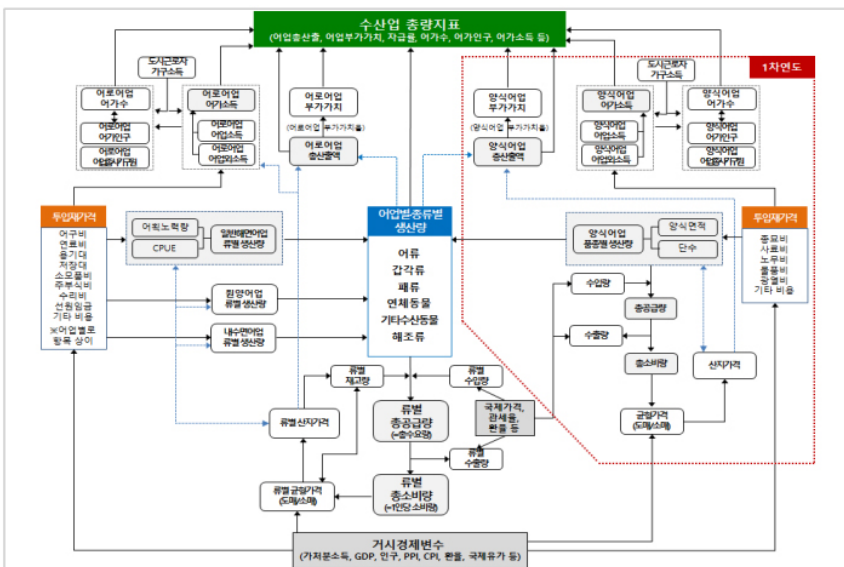
자료: 이승윤·이은경·한진현(2017), p. 17

2) 이승윤·이은경·한진현(2017), pp. 16~41.

여기서 주목할 부분은 한국은행이 전망모델시스템을 구축하는 방식이다. <그림 1-6>에서 보는 바와 같이 예측 기간, 다양한 방법론에 따라 다양한 모델을 개발한 것을 통칭하여 전망모델시스템이라고 부르며, 모델을 개발한 이후에도 수시로 예측력을 점검하고 새로운 방법론을 도입해 모델의 수를 늘리는 것이 중요하다고 언급한다.

이현동·하현정·허수진(2019)³⁾은 한국해양수산개발원의 수산부문 전망모델에 대한 설명과 함께 기존의 모델을 개선하고 새롭게 제안하였다. 기존의 모델은 2003-2004년과 2013년에 개발되었으며, 이후 2018년부터 2019년까지 2개 연도에 걸쳐 새로운 체계의 모델을 구축하여 <그림 1-7>과 같이 제안하였다.

<그림 1-7> 한국해양수산개발원 수산부문 전망모델 「KMI-FOSiM」 구조도



자료: 이현동·하현정·허수진(2019), p. 71

3) 이현동·하현정·허수진(2019), pp. 1~255.

제2절 연구의 목적

이번 연구는 운임의 예측 기간을 주간에서 한 주씩 연장하여 최장 12주 까지 연장하여 현재의 주간 단위 서비스를 분기 단위로 기간을 늘려 모델의 활용도를 높이하고자 하였다.

실제 서비스에서 구독자의 혼란을 초래한 부분이었던 단일 모델기반 운임서비스에서 탈피하기 위해 복수의 모델을 구축하여 특정 방법론에 대한 의존도를 낮추고자 하였다. 또한 단일 모델 활용 시 발생할 수 있는 편향 문제를 해결하고자 하였다.

이를 통해 실제 해상운임 예측 모델 결과를 서비스하는 과정에서 파악한 모델과 서비스의 한계 및 문제점을 개선하여 향상된 품질의 서비스를 해운 기업 및 정부 등에 제공하고자 한다.

제3절 연구의 범위 및 내용

1. 연구의 범위

본 연구는 기존 모델을 이용해 실제 예측 서비스를 진행하는 과정에서 도출된 문제점과 예측 서비스에 대한 구독자의 수요를 반영해 활용성을 높인 해상운임 예측 모델을 개발하고자 하였다. 특히 변동성이 큰 건화물선 시장에서 케이프와 파나막스에 집중하여 모델을 구축하였다(〈표 1-2〉 참고).

〈표 1-2〉 건화물선 시장의 구분 기준

항목	재화중량톤수	주요 운송 화물 종류
케이프선	• 10만 DWT 이상	철광석, 석탄
파나막스선	• 7만 DWT 이상 99,999 DWT 이하	석탄, 곡물

자료: 클락슨 내용을 바탕으로 저자 작성

인공지능에 대표 알고리즘인 신경망 계열 알고리즘인 DNN, LSTM, GRU를 활용한 복수 모델을 확보하고, 모델의 변수 및 변수 간의 정의를 재정립하고 데이터 전처리 방식을 개선하여 예측력을 높이고, 예측 기간을 최대 12주까지 확장한 후 그 결과를 제시하였다.

2. 연구의 내용 및 방법론

본 연구는 정확한 시황 전망 정보를 제공하기 위해 다양한 알고리즘을 활용한 모델을 구축하고 예측 기간을 확장하는 것이다. 이러한 목적을 달성하기 위해 세 가지 세부 목표를 정하고 이를 달성하는데 초점을 두었다. 예측 기간 확장은 단순히 기간을 늘린다고 해결되는 문제가 아니며, 본 연구에서는 FFA 데이터의 활용성 확대와 데이터 전처리 방법 개선을 통해 해운시장 변동성과 시장참여자의 미래 시장에 대한 기대 심리변화를 모델에 반영하여 문제를 해결하고자 하였다.

2장에서는 선행연구를 통해 최신 연구 동향을 파악하고 현재 보유하고 있는 모델의 개선사항을 도출하였다. 제3장에서는 제안하고자 하는 모델의 구조와 특징 및 사용하는 인공지능 알고리즘에 관해 자세하게 설명하여 제시하고, 제4장에서 구축한 모델을 활용해 시장별 운임 방향성을 추정한 결과를 제시한 후 예측 정확도를 평가한다. 마지막으로 제5장에서는 연구를 요약하고 이를 통해 얻은 결론과 시사점을 도출하였다. 본 연구를 통해 현

재 한국해양수산개발원 해운빅데이터연구센터의 기존 모델과 제안 모델 간의 차별성을 <표1-3>에서 정리하여 제시하였다.

〈표 1-3〉 기존 모델과 제안 모델의 차별성

구분		1차 연도 (2019년)	2차 연도 (2020년)	3차 연도 (2021년)
개발모델		DNN을 활용한 해상운임예측 모델 (분류모델)	LSTM을 활용한 해상운임예측 모델 (회귀모델)	인공지능기반 복수 모델 구축 (분류모델)
특 징	모델	해운 데이터의 변동성이 크 다는 점을 이용하여 기술분 석을 활용한 모델 구축	해운 데이터의 시계열적 특 성을 이용한 모델 구축	① 복수모델 개발 ② 예측 기간 확장
	데이터	• 가격 데이터 활용 (단위: day) • 현물 : 운임, 용선료. 선물 : FFA 가격		• FFA 가격 +2월물 추가
	전처리	운임만 기술분석 활용	운임만 기술분석 및 시계열 성분 분해 활용	변동성 분석 도입
			① 운임의 시계열 형태 (components of time series)를 분해 해 특징 추출	① 데이터변동성 추출 ② FFA도 변동성 분석을 통해 데이터 특성 추출
	알고 리즘	DNN(Deep Neural Network)	LSTM(Long-Short Term Memory)	DNN, LSTM, GRU
의의		해상운임예측 분야에 인공 지능을 적용한 모델을 개발 하고 그 결과를 서비스함	해운데이터의 시계열적 특 성을 활용해 예측모델을 개 발하고 추가하였음	복수모델에 기반한 운임 예 측 능력 향상 및 기간 확장
한계		• 단일 모델	• 인공지능 모델이기는 하 지만, 시계열 모델의 한 계를 포함함	• 실제 모델을 구축하고 서 비스하는 과정에서 나타 난 한계점(결과 상이, 예 측 기간) 보완
활용		• 기존 주간지에 주간 예측 부분을 추가하여 제공	• 2020년 5월 두 개의 모 델(인공지능, 시계열)에 대한 예측 결과를 서비스 하기 시작했으며, 예측 결과가 다른 경우 발생	• 복수모델 구축을 통해 단 일 모델의 편향 발생 가 능성을 보완 • 예측 기간 확장 • 향후 기존 분기 보고서에 예측 부분을 추가하여 제공

자료: 저자 작성

02

해상운임 예측 모델 구축 현황

해상운임은 정성·정량적인 방법을 활용하여 예측한다. 정성적 방법은 델파이(delphi), 시나리오 분석 등이 있으며, 정량적 방법은 다음과 같이 세 가지로 구분할 수 있다. 첫째, 수요와 공급 변수를 회귀모델이나 연립방정식모델에 포함하여 예측하는 방법이 있다. 동 방식을 사용하면 종속변수가 독립변수에 미치는 영향력을 측정하고 이를 기반으로 미래를 예측하는 순으로 진행된다. 둘째, 시계열 모델을 이용해 예측하는 방법이다. 이 방법은 데이터의 시계열적 특성(추세, 계절성, 주기성 등)을 이용하여 예측하는 방법이다. 셋째, 최근에 시도되고 있는 인공지능·빅데이터를 이용해 예측하는 방법이다. 동 방식은 기존 정형 자료에서 비정형 자료까지 이용 가능한 데이터를 확장하고 학습을 통해 미래를 예측하는 방법이다.

제1절 전통적인 예측 방법

Martin stopford(2015)⁴⁾는 해운시장 예측을 위한 다양한 방법론을 정리했다. 분석 방법에 따라 4가지 대분류와 14개 소분류로 구분하였는데,

4) 마틴 스토포드, 양창호 외 역(2015), 박영사. pp. 1~955.

설문조사 방법은 해운에 종사하는 전문가들이 새로운 트렌드를 파악하는데 유용한 방법이다. 다만 설문조사 방법 자체가 가지는 조사방식, 표본추출에 따른 편향성 등은 문제점으로 남는다.

추세분석은 시계열 자료를 활용해 예측하는 방법이며, 상대적으로 간단하나 예측 시점에 따라 추세가 변경되어 결과가 크게 바뀔 수 있다는 단점이 있다. 수학적 모델은 일반적인 회귀모델을 기반으로 예측하는 방식으로 독립변수와 종속변수의 관계를 통해 미래를 예측하는 방식이다. 다만 동모델은 미래를 예측하기 위해서는 독립변수에 대한 미래값을 추정해야 한다. 확률분석은 시나리오 분석과 유사한데 미래에 발생 가능한 이벤트의 확률을 계산하는 분석 방법이다.

〈표 2-1〉 해운에서 사용한 분석적 기법

분석기법	주요특성
1.설문조사	
델파이 기법	전문가 그룹이 합의된 예측을 도출할 수 있도록 하는 논의회합
2. 추세분석	
단순추세	단순규칙. 예를 들어 ‘무변화’ 또는 만일 수익이 2배 이상 된다면 OPEX는 하락할 것임
추세외삽 (extrapolation)	몇 가지 방법을 사용하여 추세를 추정하고 과거의 추세를 대입하여 미래를 예측
평활법 (smoothing)	평균 변화를 얻기 위해 변동치를 평균화하여 예측
분할 (Decomposition)	추세, 계절성, 주기성과 무작위 변동을 분리하고, 분리한 각각을 추정
필터 (Filters)	예측치가 과거 실제값 혹은 잔차들(errors)과 선형조합으로 표시됨
자기회귀 (Autoregressive)	예측치가 과거 실제값의 선형조합으로 표시함
박스젠킨스모델	인정성 문제를 처리하는 규칙을 가진 자기회귀(ARMA)모델의 변량
3. 수학적모델	
단일회귀	목표변수를 예측하기 위해 한 가지 설명변수로 추정된 방정식

분석기법	주요특성
다중회귀	목표변수를 예측하기 위해 한 가지 이상의 설명변수로 추정된 방정식
계량경제 모델	목표변수를 예측하기 위한 회귀방정식시스템
수급 모델	구성요소로써 수급을 추정하고 균형변화량을 예측
민감도분석	다른 가정으로 예측치의 민감성을 분석
4. 확률분석	
몬테카를로	특정한 결과가 일어날 가능성을 계산하기 위하여 확률분석이 사용됨

자료: 마틴 스토포드, 「해운경제학」, 양창호 외 역(2015), 박영사. p 878

1. 회귀모델 기반 예측 모델

Beenstock and Vergottis(1989)⁵⁾는 외부 충격으로 인하여 유조선 시장의 반응을 분석하기 위해 운임, 용선, 중고선, 폐선 등 시장을 세분화한 모델을 구축하였다. 세부시장 변화를 파악하기 위하여 변수선정작업을 선행하고 특정 방법론(3SLS)⁶⁾을 이용해 모델을 정립하였다.

유조선 시장을 예로 들면, 수요와 공급 변수를 이용하여 가설을 세운 후, 실증분석을 통해 모델을 구축하는 방식이다. 이를 통해 유조선 시장 계선률(λ)이 운임과 계선 비용에 반비례하고 운항비와 고정비에 비례한다고 추정했으며, 아래와 같은 식을 도출하였다.

$$\text{수요함수: } M^s = K^*(FR/\alpha PB)^{1/(\alpha-1)}$$

여기서,

M 은 톤-마일 수요(demand for ton-miles)

5) Beenstock & Vergottis(1989), pp. 265-280.

6) Beenstock & Vergottis(1989), p. 270.

FR 은 운임(freight rates)

PB 는 선박의 단위비용(Unit voyage costs)을 말한다.

$$\text{공급함수: } K^* = K(1 - \lambda) - ST + COM$$

여기서,

K 는 유조선 선박량(tanker fleet)

X 는 유조선 계선 비율(proportion of tankers laid up)

ST 는 유조선 해상원유저장(tankers in storage)

COM 은 겸용선(combined carriers)을 말한다.

Beenstock, and Vergottis(1989)⁷⁾는 운임시장 외에 선박 건조, 폐선, 중고선 시장도 위와 같은 과정을 통해 모델을 구축하고 세부시장별 시나리오(해상물동량 10% 증가, 연료유가 25% 상승)에 따른 영향을 도출하였다.

Luo, Fan, and Liu(2009)⁸⁾는 컨테이너 운임을 예측하기 위해 운송 수요와 공급 간 관계를 이용하여 다음과 같은 가정을 설립하였다. 컨테이너 공급의 변화는 시장 운임과 운송 수요, bunker 가격의 함수로 구성되며, 발주와 인도 시점 간에는 차이가 발생한다. 또한 가격의 변화는 수요와 공급에 따라 변동한다. 이러한 가정을 근거로 세계 컨테이너 선대에 대한 방정식과 수급 변화에 따른 운임의 변화를 측정하기 위한 2개의 방정식을 아래와 같이 정립하였다.

$$\ln M_t = 0.947 + \ln K_t + 0.31 \times \ln (FR/PB)_t$$

$$\ln (\lambda / (1 - \lambda))_t = 3.673 - 1.435 \times \ln (FR/TQ)_t - 2.096 \times \ln (FR/TQ)_t - 1$$

7) Beenstock & Vergottis(1989), pp. 263-280.

8) Luo, Fan, & Liu(2009), pp. 507-523.

식을 유도하기 위해 1980년부터 2008년까지 컨테이너 시장 자료를 3단계 최소제곱법(3SLS: three stage least squares)으로 분석했다. 이 방법은 독립변수와 오차항이 상관관계가 나타날 때 도구 변수를 이용해 이를 보정하는 2단계 최소제곱법과 연립방정식 간 오차항의 상관관계를 고려하는 무관 회귀(SUR: seemingly unrelated regression)가 합쳐진 방법론으로 추정 결과 수요와 운임의 90% 이상을 설명할 수 있다.

$$\begin{aligned}\Delta X_t &= \eta P_{t-\theta} Y_{t-\theta} - \eta c_1 X_{t-\theta} - \eta c_2 OIL_{t-\theta} + \epsilon_{1t} \\ &= \alpha_1 P_{t-\theta} Y_{t-\theta} - \alpha_2 c_1 X_{t-\theta} - \alpha_3 c_2 OIL_{t-\theta} + \epsilon_{1t}\end{aligned}$$

$$\Delta P_t = \delta \Delta Y_t - \delta \varphi \Delta X_t + \varepsilon_{2t} = \alpha_4 \Delta Y_t - \alpha_5 \Delta X_t + \varepsilon_{2t}$$

여기서,

Y : 컨테이너 운송 수요

P : TEU당 시장 운임

X : 컨테이너 공급(선대량)

OIL : 병커 가격

θ : 발주와 인도시점 차이를 말한다.

Beenstock and Vergottis(1989)⁹⁾와 Luo, Fan, and Liu(2009)¹⁰⁾ 연구 모두 3단계 최소제곱법을 사용했는데 이는 2단계 최소제곱법을 기반으로 한다. 2단계 최소제곱법은 도구 변수를 이용한 추정 방법인데 이는 내생적인 설명변수가 오차항과 상관관계를 가지면서 이를 해결하기 위해 사용하는 방법이다. 해운에서는 수요와 공급을 이용해 운임을 예측하는데 다음과 같이 수급 함수가 있다고 가정한다.

9) Beenstock & Vergottis(1989), pp. 263-280.

10) Luo, Fan, & Liu(2009), pp. 507-523.

$$\text{수요 : } Q^d = \alpha_0 + \alpha_1 P + \alpha_2 x + e^d$$

$$\text{공급 : } Q^s = \beta_0 + \beta_1 P + e^s$$

민인식·최필선(2009)¹¹⁾은 이들 수요-공급 함수에서 수요와 공급량에 영향을 미치는 오차항은 균형 수량뿐만 아니라 균형 가격에도 영향을 미치게 되어 가격 변수는 내생성(endogeniety)을 갖게 된다고 설명한다.

회귀모델에 내생 변수가 있을 때는 모델의 구성 변수와 오차항 간 상관관계가 없지만, 내생적 변수와는 존재할 때에 도구변수(IV: instrumental variable)를 이용해 추정할 수 있다. 2단계 최소제곱법을 추정하기 위해서는 1단계로 내생성이 있는 가격 변수를 종속변수로 하며 연립방정식 모델 내 모든 외생변수에 대한 회귀식으로 추정한다. 1단계에서 추정된 P의 적합값($\hat{P}$)을 회귀 모델에서 P를 $\hat{P}$ 로 대체하여 OLS로 추정한다.

2. 시계열 모델 기반 예측 모델

1) 시계열분해법

김형준·류두진·조훈(2019)¹²⁾, 이석용(2020)¹³⁾의 연구에서는 원자료(Y_t)를 추세성분(T), 순환변동성분(C), 계절변동성분(S), 불규칙변동성분(I)로 분해하여 예측하는 시계열 요소분해법을 이용해 운임을 예측하였다. 김형준·류두진·조훈(2019)은 로그 변환 된 해상운임지수로부터 추세선분과 계절변동성분을 분해하고 이를 바탕으로 단기 예측을 시행하는 승법 모델을 적용하였다.

11) 민인식·최필선(2009), p. 229.

12) 김형준·류두진·조훈(2019), pp. 715~731.

13) 이석용(2020), pp. 597~618.

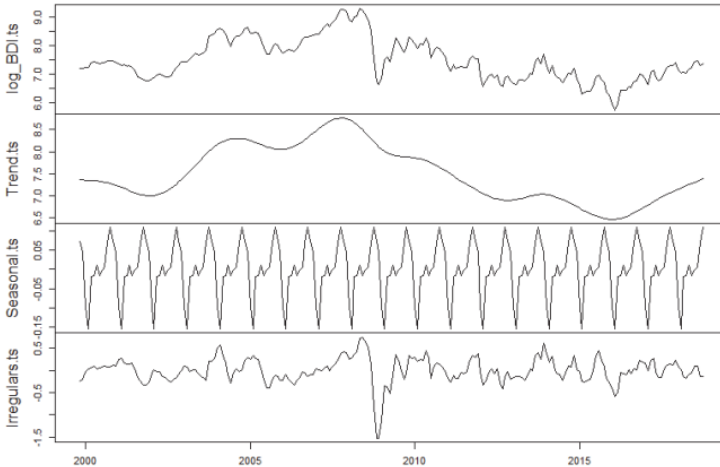
$$Y_t = T_t \times C_t \times S_t \times I_t$$

추세순환성분을 추출하기 위해 호드릭-프레스콧 필터(Hodrick-Prescott filter)를 사용한다. 호드릭-프레스콧 필터는 아래 수식의 제곱합을 최소화 하는 $\tau_{t=1}^T$ 을 계산하는 방법이다. λ 값이 클수록 평활성이 커지며 작을수록 원시계열 자료와 유사하게 된다(김형준·류두진·조훈, 2019).¹⁴⁾ 이 방법으로 추세(T)와 순화변동(C)을 구분한다.

$$\min_{\tau_{t=1}^T} t \left[\sum_{t=1}^T (x_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} (\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})^2 \right]$$

계절성은 원자료로부터 추세순환성분을 추출한 후 단순평균을 사용하였으며 상기에서 나타나지 않은 변화는 불규칙변동성분으로 분리했다.

〈그림 2-1〉 BDI 관측치의 시계열 요소분해법 적용결과(1999.11~2018.10)

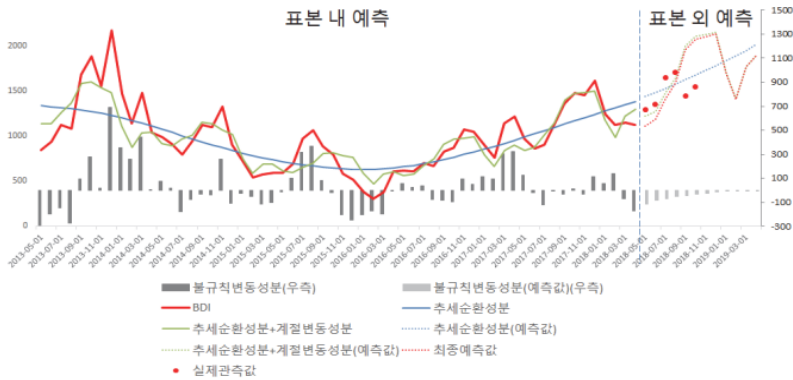


자료: 김형준·류두진·조훈(2019), p.722

14) 김형준·류두진·조훈(2019), p. 720.

이후 선형외삽법으로 추정된 추세순환성분에 계절변동성분을 더하고 ARIMA로 추정된 불규칙변동성분을 더하면 예측이 가능하다. 김형준·류두진·조훈(2019)의 연구는 다음과 같은 결과를 도출했다.

〈그림 2-2〉 2013년 5월부터 2018년 4월까지의 BDI 자료를 활용한 전방예측 결과 예시



자료: 김형준·류두진·조훈(2019), p.724

2) ARIMA

손상기(1995)¹⁵⁾, 모수원(2010)¹⁶⁾은 ARIMA 모델을 이용해 해운 운임을 예측했다. 동 모델은 시계열자료 안에 내재되어 있는 특성을 이용하는 방식으로 자기회귀모델(AR: Autoregressive Model)과 이동평균(MA: Moving Average)을 결합(Integrated)한 형태이다(손상기, 1995).

$$y_t = \alpha + \sum_{s=1}^p \gamma_s y_{t-s} + \varepsilon_t : AR(p) \text{모형}$$

$$y_t = \mu + \varepsilon_t + \sum_{s=1}^q \theta_s \varepsilon_{t-s} : MA(q) \text{모형}$$

15) 손상기(1995), pp.197~212.

16) 모수원(2013), pp. 67~76.

시계열 자료를 사용하기 위해서는 몇 가지 가정이 선행되어야 하는데 첫째, 확률 과정은 선형 모델을 따른다. 둘째, 설명변수와 다른 시점에 확률적 오차는 서로 상관관계가 없다. 셋째, 시계열 자료의 분산은 모든 기간 일정하다. 넷째, 오차항의 평균은 0이고 일정한 분산을 갖되 서로 독립적인 정규분포를 따른다. 이를 백색오차라고하며, 이는 t 기의 충격과 $t+1$ 간의 충격이 서로 독립적이라는 의미이다.

시계열이 불안정할 경우 독립변수와 설명변수 간의 가성회귀(Spurious regression) 문제가 발생할 가능성이 높아 자료의 안정성을 검증해야 한다. 일반적으로 단위근검증을 이용해 자료 안정성을 확인한다. 대부분 시계열 자료는 비정상성을 가지며, 이는 차분을 통해 해결한다. 원자료가 정상성을 가지고 있는 시계열이면 아래의 ARMA 모델을 활용하여 분석할 수 있다.

$$\begin{aligned}\Delta^d y_t &= \Delta y_t - \Delta y_{t-d} \\ &= \sum_{i=1}^p a_i \Delta^d y_{t-i} + \Delta^d \alpha_t + \sum_{j=1}^q \theta_j \Delta^d f_{t-j} = (1-B)^d y_t\end{aligned}$$

불안정한 ARMA 모델이 차분을 통해 안정화되면 다음과 같이 표현할 수 있다.

$$\text{ARIMA}(p,d,q) : a_p(B)(1-B)^d y_t = \theta_0 + \theta_q(B)f_t$$

이재길(2017)¹⁷⁾은 수식에서 f_t 는 평균이 0이고 분산이 σ_f^2 인 백색잡음이며 θ_0 는 추세모수로 0이다. 추세모수가 제거된 1차 차분의 경우 시계열 자료는 차분된 시계열 자료 z_t 의 누적형태로 나타난다고 설명한다.

17) 이재길(2017), p. 314.

$$y_t = (1 - B)^{-1} \Delta y_t = \left(\frac{1}{1 - B} \right) z_t = \sum_{k=-\infty}^t z_k$$

ARIMA모델에서 차수를 탐색하고 결정되면 적합성을 진단하고 모델을 확정해 예측한다. 손상기(1995)¹⁸⁾의 연구에서는 ARIMA(1,1,0)이 최적의 모델로 확정되었으며 모수원(2010)¹⁹⁾은 3개의 ARIMA모델과 2개의 개입-ARIMA모델을 사용해 운임을 예측하였다.

3) 벡터자기회귀모델/벡터오차수정모델

Chen and Meersman(2012)²⁰⁾, 이성윤·안기명(2018)²¹⁾, 정상국·김성기(2011)²²⁾는 다변량 시계열 방법론인 벡터자기회귀모델(VAR: Vector Autoregressive)이나 벡터오차수정모델(VECM: vector error correction models)을 이용해 해운 운임을 예측하였다.

벡터자기회귀모델은 자기회귀모델(AR)을 다변량으로 확장시킨 모델로 2개 이상의 내생변수의 동적변화를 내생변수들의 과거 값을 이용하여 모델화한다(민인식·최필선, 2014)²³⁾. 이를 수식으로 표현하면 다음과 같다.

$$AR: y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + u_t$$

$$VAR: y_t = \sum_{m=1}^p A_m y_{t-m} + u_t$$

18) 손상기(1995), pp. 197~212.

19) 모수원(2013), pp. 67~76.

20) Chen & Meersman(2012), pp. 498~537.

21) 이성윤·안기명(2018), pp. 546~554.

22) 정상국·김성기(2011), pp. 217~240.

23) 민인식·최필선(2009), p. 165.

VAR에서 A_m 은 k 차원일 경우 $k \times k$ 행렬이 되며 u_t 는 k 차원의 백색잡음이 된다(이재길, 2017)²⁴⁾. 벡터자기회귀모델과 벡터오차수정모델의 프로세스는 다음과 같다.

첫째, 자료의 변수를 탐색한 후 로그 변환하여 분석한다. 둘째, 자료의 정상성을 확인하기 위해 단위근 검정을 실시한다. 이를 통해 안정성을 갖는 시계열 자료로 판단되면 벡터자기회귀모델로 분석한다. 다만 경제적인 인과관계가 발생할 때는 구조적 벡터자기회귀(S-VAR)로 모델을 구축한다. 셋째, 불안정한 자료이지만 공적분 관계를 갖는 경우에는 벡터오차수정방법을 활용하고, 공적분 관계가 없을 때는 자료를 차분하여 안정화한 후 벡터자기회귀모델로 구축한다. 넷째 최적 시차와 변수 순서(외생성이 강한 변수부터)를 정한다. 마지막으로 충격반응과 분산분해 등을 통해 분석을 마무리한다.

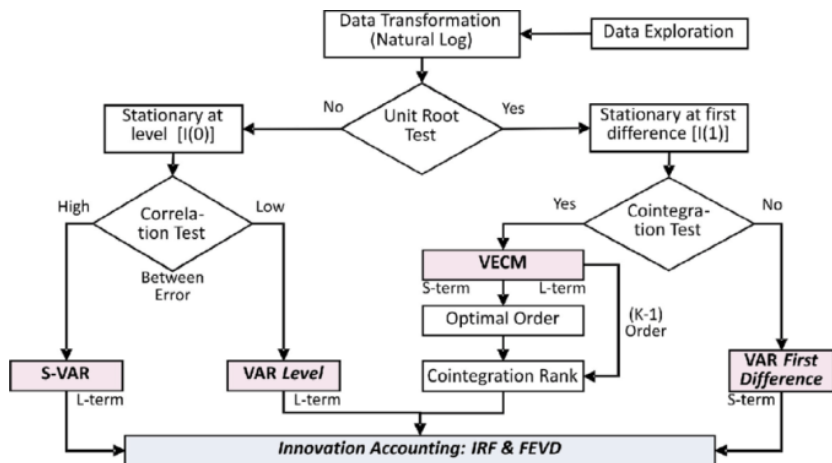
이성윤·안기명(2018)²⁵⁾은 금리, 환율 등의 거시경제 변수와 금융위기와 같은 경제적 충격요인, 발주량 등이 건화물선 운임지수에 영향을 미친다는 결론을 제시하였다.

Chen *et al.*(2012)은 케이프, 파나막스, 수프라, 핸디선에 대한 세부 항로 운임과 함께 거시경제변수(달러 인덱스)를 사용해 추정하였다.

24) 이재길(2017), p. 399.

25) 이성윤·안기명(2018), pp. 546-554.

〈그림 2-3〉 VAR/VECM 프로세스



자료: Ascarya (2017), p. 512

제2절 인공지능에 기반한 해상운임 예측 방법

1. 선행연구

최근 들어 변동성이 크고 불규칙한 해운시황을 전망하기 위한 방법으로 딥러닝을 이용한 연구들이 활발히 진행되고 있다. 딥러닝은 머신러닝의 한 종류로 입력데이터에 대한 유형 분류나 회귀를 수행하는 기법이며(아마시타 타카요시, 2017)²⁶⁾, 딥러닝 모델 분석은 크게 5단계로 진행된다.

1단계에서 초 모수(hyperparameter)를 지정한다. 초 모수는 연구자가

26) 아마시타 타카요시, 심효섭 역(2017), 제이펍. pp. 1~224.

직접 결정할 수 있는 변수로 레이어 수, 학습 기간, 배치 사이즈 등 다양하다. 2단계는 최적화로 예측한 값과 실제값의 차이를 최소화하게 만드는 모델 내의 매개변수를 찾는 과정이다. 이를 위해 경사하강법(gradient descent)을 사용한다. 3단계는 드롭아웃(dropout)으로 테스트 셋의 결과 가 트레이닝 셋에 미치지 못하는 과적합 문제를 해결하는 방법이다.

이 방법은 은닉층에 있는 셀을 무작위로 삭제하는 방법으로 특정 가중치에 대해 민감도가 하락해 일반화가 가능하다. 4단계는 반복 횟수(Epoch) 결정이다. 반복 횟수가 커지면 과적합이 될 가능성이 커 테스트 셋의 손실 함수 값은 커질 수 있다. 따라서 적절한 반복 횟수를 지정하는 것이 중요하다. 5단계는 예측오차평가로 신경망 모델의 성능을 평가하기 위해서는 예측값과 실제값 간 차이를 이용한다.

김경환(2021)²⁷⁾은 일반적으로 MAPE(Mean Absolute Percentage Error), RMSE(Root Mean Squared Error), MAE(Mean Absolute Error) 등 다양한 방법을 이용한다고 설명한다.

Uyara, İlhan, and İlhan(2016)²⁸⁾, 한민수·유성진(2017)²⁹⁾은 전화물 선 운임 예측을 위해 RNN(Recurrent Neural Networks)을 이용했다. RNN 모델은 연속되는 자료에 대한 학습능력이 뛰어나 추세나 시간을 이용해 예측하는데 사용된다. 동 모델은 인공신경망과 구조는 유사하나 입력층과 출력층 사이에 데이터 일부가 순환하는 경로(loop)를 가지고 있다는 특징이 있다.

RNN 모델은 학습하는 과정에서 과거의 정보를 기억하는 동시에 최신의 데이터로 정보를 갱신할 수 있게 되었다. 즉, 은닉계층(hidden layer)이

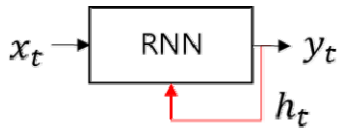
27) 김경환(2021), pp. 1~120.

28) Uyara, İlhan & İlhan(2016), pp. 642~647.

29) 한민수·유성진(2017), pp. 50~53.

일종의 메모리 역할을 하게 되는 것이다. 이를 아래와 같이 식으로 나타낼 수 있다(황수진 외, 2020).³⁰⁾

〈그림 2-4〉 순환신경망 모델



자료: 황수진 외(2020), p.37

$$y_t = \tanh(h_{t-1} W_h + x_t W_i + b)$$

여기서,

t 는 시간

x_t 는 입력값을 말한다.

Lyridis *et al.*(2004)³¹⁾는 원유 수요, 선대규모, 원유 생산량, 원유 가격, 용선료, 신조선가, 중고선가, 병커 유가, 해체가 등 변수를 활용하여 인공 신경망(ANN: Artificial Neural Networks) 모델을 활용해 VLCC 운임 예측모델을 연구하였다. 1개월, 3개월, 6개월, 9개월, 12개월 후 시장에 대한 예측을 시도하여 예측 기간이 장기화할수록 예측이 불안정해지는 결과를 보여준다.

Li *et al.*(2006)³²⁾은 ANN 기반 유조선 운임 예측 모델을 개발하고 전통적인 시계열 모델인 ARMA 모델과 그 성능을 비교하였다. 그 결과 단기 예측에는 시계열 모델이 우수했으나 예측의 기간이 길어질수록 ANN 기반

30) 황수진 외(2020), p. 37.

31) Lyridis *et al.*(2004), pp. 93-108.

32) Li *et al.*(1997), pp. 9-30.

모델이 ARMA 모델과 비교해 예측력이 우수하다는 결론을 제시하였다.

모수원(2010)³³⁾은 다변량 모델보다 단일변량 모델이 더 우수한 예측성
과를 보임과 BDI를 설명하는 변수들이 BDI 자료보다 시차를 두고 발표되
는 것을 근거로 단일변량분석기법인 자기회귀-이동(ARIMA)모델과 장기순
환 가정을 도출하는 호드릭-프레스콧 필터(Hodrick-Prescott)필터 기법
을 활용하여 분석을 진행했다.

Han *et al.*(2014)³⁴⁾은 하이브리드 기반 BDI 예측 모델을 제안하였는데
이는 데이터 전처리 과정에 신호분석법인 웨이블릿으로 BDI를 분해해 특
징을 추출하고 이를 모델의 입력변수로 활용한 방식을 말한다.

인공지능 알고리즘은 SVM을 사용했으며, 비교 모델로 VAR, ARMA 모
델을 제시하였다. 모델간 비교 결과 하이브리드형 SVM 모델의 예측성이
좋았다는 결론을 제시하였다.

Zeng *et al.*(2016)³⁵⁾은 하이브리드형 BDI 예측 모델을 개발하였다. 여
기서 하이브리드형이란 데이터 전처리 과정에서 EMD를 이용해 BDI 데이
터 특징을 추출하여 모델 학습과정에 입력변수로 활용하도록 설계된 모델
을 말한다. 인공지능 알고리즘은 ANN을 사용했으며, EMD-ANN 모델의
예측 성능이 우수함을 보여주었다.

Bao, Pan, and Xie(2016)³⁶⁾는 SVM기반 BDI 예측 모델을 개발하고
비교 모델로 NN모델을 제시한 뒤 두 모델의 성능을 비교하였다. 그 결과
1개월 후 BDI 예측에 SVM이 적합하다는 결과를 제시하였다.

Shynkevich *et al.*(2017)³⁷⁾는 기술지표와 머신러닝 기법을 활용하여 주

33) 모수원(2010), pp. 222~233.

34) Han *et al.*(2014), pp. 1~12.

35) Zeng *et al.*(2016), pp. 192~210.

36) Bao, Pan, & Xie(2016), pp. 65~69.

가 방향성을 예측하는 모델을 개발하였다. 주가는 기술분석을 통해 이동평균, 윌리엄스 %R, 스톡캐스틱 등으로 변환하여 입력변수로 사용하였고, 인공지능 알고리즘으로는 인공신경망, SVM, KNN 알고리즘을 이용해 주가의 1일부터 30일 이후의 주가 방향성을 예측하는 분류모델을 제시하였다.

Eslami *et al.*(2017)³⁸⁾은 이전 소개한 하이브리드 모델과는 다른 접근 방식을 통해 모델의 예측력을 높이하고자 했다. 대부분 하이브리드 모델이 데이터 전처리에 새로운 방법을 추가했다면, 이 모델은 유전자 알고리즘의 특징을 이용해 학습 횟수를 최소화하여 해를 찾아내는 방식을 적용하였다. 이를 통해 과대 적합과 과소 적합을 피하고 모든 경우의 수를 계산하는 전체 탐색영역 방법이 아닌 최소한의 횟수만으로도 최적화된 ANN 모델을 개발하는 방법을 제안하였다.

이우식(2017)³⁹⁾은 코스피 주가지수 방향성 예측 문제에 기술적 분석 지표와 의사결정나무(decision tree), 서포트 벡터 머신(SVM: Support Vector Machine), 인공신경망 알고리즘을 결합한 모델을 비교하였다. 본 연구에서는 상대강도지수, 모멘텀, 변화율, 지수이동평균 등 기술적 지표를 활용하여 75%의 정확도를 보였다.

Yang and Mehmed(2019)⁴⁰⁾는 ANN을 이용하여 파나막스선 운임지수를 예측하는 모델을 개발하였다. 이 모델은 FFA를 사용하여 데이터 부분에서도 기존 모델과 차별성을 강조하였다. 데이터는 파나막스선 운임지수와 FFA 두 가지를 이용했으며, 데이터 전처리와 조합을 기준으로 두 개 그룹으로 구분한 후 ANN 모델을 구축하였다.

전자는 파나막스선 운임지수 자체 시차 차이를 이용한 그룹을 입력변수

37) Shynkevich *et al.*(2017), pp. 71-88.

38) Eslami *et al.*(2017), pp. 538-550.

39) 이우식(2017), pp. 287-295.

40) Yang & Mehmed(2019), pp. 390-414.

로 하는 ANN 모델이며, 후자는 파나마스선 운임지수와 FFA 두 가지 변수를 입력변수로 하는 ANN 모델이다. 두 모델의 성능을 비교하여 제시하였는데, FFA 가격을 입력변수로 이용한 ANN 모델의 예측 성능이 우수했다는 결론을 제시하였다.

Zhang *et al.*(2019)⁴¹⁾는 하이브리드형 BDI 예측 모델을 제안하였다. DFN을 활용해 데이터 전처리 과정에서 운임 데이터가 가지고 있는 정보 중 불필요한 정보를 제거하여 모델을 구축했다는 특징을 갖는다. DFN을 활용하여 BDI에서 추세, 주기성, 오차항 등을 추출하고 이를 ANN에 적용하여 모델을 개발하였다. 결과적으로 DFN을 활용해 운임 노이즈를 제거한 모델의 예측력이 기존 ANN보다 우수하다는 결과를 제시하였다.

2. 시사점

해상운임 예측에도 모델을 구축하기 위해 인공지능 방법론을 활용하려는 시도의 연구가 증가하고 있다. 특히 최근에는 하이브리드형 인공지능 모델을 제안하는 연구가 늘어나고 있다.

하이브리드형은 데이터의 전처리 방식에서 새로운 방법론을 도입하여 모델의 학습에 입력변수로 활용되기 전에 특징을 추출해 변환하는 과정을 추가한 형태와 서로 다른 인공지능 알고리즘을 결합하여 학습의 효율성을 향상하게 시키는 등 서로 상호보완하는 효과를 얻게 만드는 방법이 있다.

이번 연구는 데이터 전처리 방식에서 기존의 연구자들이 해운 분야에서 사용하지 않은 변동성 분석 방법을 도입했으며 특히 FFA 가격 변동성 추출을 통해 시장심리 변화를 수치화하여 활용하였다 점에서도 차이가 있다.

41) Zhang *et al.*(2019), pp. 499-516.

〈표 2-2〉 선행연구 요약

구 분	선행연구와의 차별성		
	연구목적	연구방법	주요 연구내용
주 요 선행 연구	• 과제명: Forecasting Tanker Market Using Artificial Neural Networks • 연구자(년도): Lyridis 외 (2004) • 연구목적: 인공신경망을 활용한 탱커 스팟운임 예측	• 인공신경망(ANN)을 활용한 예측 분석 • 수요, 공급, 운임데이터 활용	• 비정상 및 비선형 시계열 처리에 적합한 인공신경망을 활용하여 스팟운임 예측 (1개월, 3개월, 6개월, 9개월, 12개월) • 예측 개월마다 다양한 변수의 조합에 따른 모델 설정으로 최적의 예측 모델 선택 • 인공신경망 모델과 시계열 모델 예측력 비교
	• 과제명: 딥러닝 분석과 기술적 분석 지표를 이용한 한국 코스피 주가지수 방향성 예측 • 연구자(년도): 이우식(2017) • 연구목적: 기술적 지표와 딥러닝분석 기법을 활용한 주가 방향성 예측	• 의사결정나무, 서포트 벡터 머신, 인공신경망을 활용한 주가 예측 분석 • 상대강도지수, 모멘텀, 변화율 지표 등의 기술적 지표 활용	• 주가의 추세와 강도를 나타내는 기술적 지표를 바탕으로 주가의 변화 예측 모델 학습 • 예측 모델의 성능 분석 및 평가
	• 과제명: Forecasting price movements using technical indicators: Investigating the impact of varying input window length • 연구자(년도): Shynkevich 외(2017) • 연구목적: 기술적지표와 머신러닝 기법을 활용한 주가 방향성 예측	• 인공신경망, 서포트 벡터 머신, k-최근접 이웃을 활용한 주가 방향성 예측 • 이동평균, 윌리엄스 %R, 스토캐스틱 등 기술적 변수 활용	• 기술적 지표와 머신러닝을 통해 1일~30일 후 주가 방향성 예측 모델 개발 • ‘상승’, ‘하락’ 두 가지 범주를 가지는 모델과 ‘상승’, ‘보합’, ‘하락’ 세 가지 범주를 가지는 모델 개발
	• 과제명: Forecasting price movements using technical indicators: Investigating the impact of varying input window length	• 인공신경망, 서포트 벡터 머신, k-최근접 이웃을 활용한 주가 방향성 예측 • 이동평균, 윌리엄스 %R, 스토캐스틱 등 기	• 기술적 지표와 머신러닝을 통해 1일~30일 후 주가 방향성 예측 모델 개발 • ‘상승’, ‘하락’ 두 가지 범주를 가지는 모델과 ‘상승’, ‘보합’, ‘하락’ 세 가지 범주를

구분	선행연구와의 차별성		
	연구목적	연구방법	주요 연구내용
	- 연구자(년도): Shynkevich 외(2017) - 연구목적: 기술적지표와 머신러닝 기법을 활용한 주가 방향성 예측	술적 변수 활용	가지는 모델 개발
	- 과제명: A new approach for Baltic Dry Index forecasting based on empirical mode decomposition and neural networks - 연구자(년도): Qingcheng Zeng 외 (2016) - 연구목적: 하이브리드형 BDI 예측 모델 개발	- 대상 : BDI - 데이터 : 일별, 주간 데이터, 1999년 11월 1일부터 2013년 11월 29일까지 - 하이브리드형 - EMD(경험적 모드 분해, Empirical Mode Decomposition)법은 신호처리 분야에서 개발된 기술임	- 데이터 전처리과정에서 EMD방법을 활용해 BDI를 장기추세와 오차항, IMF(고유모드함수)로 분해함 - ANN에 입력변수로 활용하여 BDI를 예측하는 모델을 개발 이를 하이브리드형이라 함 - EMD-ANN모델의 예측 성능이 우수함을 보여줌
	- 과제명: Forecasting Dry Bulk Freight Index with Improved SVM - 연구자(년도): Qianqian Han 외 (2014) - 연구목적: 하이브리드기반 BDI예측 모델 개발	- 대상 :BDI - 하이브리드형 - 입력변수를 웨이블릿(Wavelet)으로 분해 - 분해한 값을 입력변수로 활용해 SVM기반 예측 모델을 개발함 - 비교모델로 VAR, ARMA모델을 제시함	- 데이터 전처리 과정에서 신호분석 기법 중 하나인 웨이블릿으로 BDI를 분해해 특징을 추출하고 이를 입력변수로 활용 - 하이브리드형 SVM모델이 다른 비교모델에 비해 우수함을 보여줌
폐기	복수의 모델 확보를 통해 예측 성능을 제고하고 예측 기간을 분기까지 확장해 정보의 활용성을 확장하고자 함	- 데이터 전처리 - 하이브리드형 인공지능 모델(지도학습 중심: 신경망,순환신경망 등)	- 하이브리드형 해상운임예측 모델 개발 - 단일모델대비 예측력 향상을 위한 다양한 모델 구축 - 예측 기간 확장을 위한 데이터 변수 조합 및 전처리 단계 추가

자료: 저자 작성

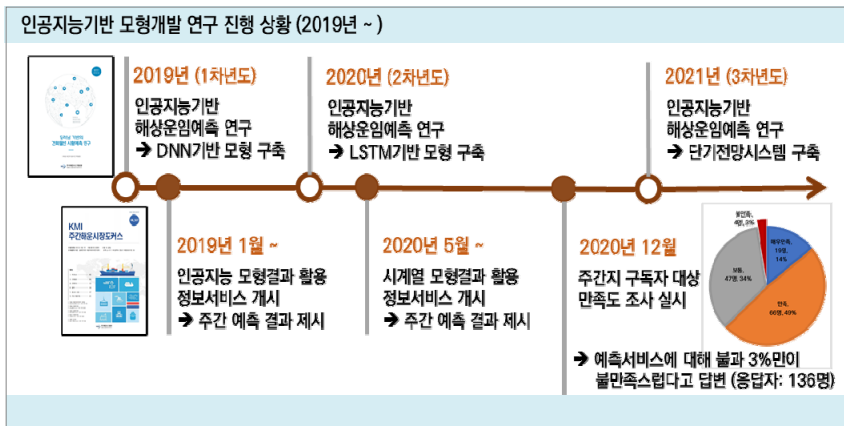
제3절 해상운임예측모델 연구진행 현황

1. 기존 보유 모델

이번 절에서는 한국해양수산개발원 해운빅데이터연구센터가 2018년 3월 설립된 이후 일반과제에서 수행된 연구 결과를 통해 보유하고 있는 모델에 관해 설명하고자 한다.

인공지능 기반 모델을 구축하는 연구는 2019년부터 연구를 수행하기 시작했으며 올해로 3년 차에 접어들었으며 지난 과거 2년 동안 총 두 가지 타입의 신경망 계열 알고리즘에 기반한 인공지능 모델을 개발하였다(그림 2-5) 참고).

〈그림 2-5〉 인공지능기반 모델 개발 연구 진행 상황



자료: 저자 작성

1차 연도 연구에서는 인공지능 알고리즘 중에 가장 널리 알려진 심층 신경망(DNN, Deep Neural Network) 모델을 구축하였고, 차기 연구에서

는 해운 데이터의 시계열적 특성을 활용하기 위해 이러한 특성을 활용할 수 있도록 구조화된 순환신경망 알고리즘 LSTM을 활용하여 모델을 개발하는 연구를 수행하였다. 각 모델에 대한 특성과 결과에 대해 자세히 알아 보도록 한다.

1) DNN에 기반한 해상운임 예측연구(1차 연도)⁴²⁾

2019년도 신경망 계열 인공지능 알고리즘인 심층 신경망(Deep Neural Network) 방식을 활용하여 진화물선 시장 운임을 예측하는 모델을 구축하였다. 이 모델 특징은 해운 데이터가 변동성이 크다는 점에 주목해 주식 투자자들이 주식의 방향성을 파악할 때 주로 사용하는 방법으로 알려진 기술분석을 이용했다는 점이다.

이 방법을 통해 운임 데이터 특징을 추출하고 모델 입력변수로 활용하여 모델의 예측성능을 향상하고자 하였다. 또 다른 하나는 시장의 시장심리가 반영된 정보로 간주하는 용선료와 선도거래 데이터를 모델의 입력변수로 활용했다는 점이다. 당시 연구를 통해 구축된 모델은 운임의 방향성을 예측하는 분류모델이며, 예측 정확도는 아래와 같이 나타났다.

구축한 모델은 차주 케이프선 운임 70.9%, 파나막스선 72.3%, 수프라막스선 79.7%로 방향성을 맞췄으며, 시장별로 그 결과가 다르게 나타나나 대체로 70% 이상 높은 예측력을 보여주었다(〈그림 2-6〉 참고), 이 연구를 통해 구축된 모델을 활용하여 2019년 1월부터 실제 예측 서비스를 시작하였다.

이 연구를 통해 동 모델이 데이터 전처리 과정의 기술분석을 추가하고 입력변수에 시장심리를 나타낼 수 있는 데이터를 확보하여 활용함으로써 모델의 해운시장에 대한 학습 능력을 향상하였으나, 입력변수로 활용한 운

42) 윤희성 외(2019), pp. 1~81.

임, 장단기용선료, FFA 가격 데이터 중에 실제 기술분석을 통해 성분을 변동성을 추출해 활용한 데이터는 운임에 한정했다.

2) LSTM에 기반한 해상운임 예측 연구(2차 연도)⁴³⁾

2차 연도에는 해상운임 데이터의 시계열적 특성을 활용하는데 중점을 둔 모델을 개발했다는 점에서 기존 연구와 차별화하였다. 건화물선 시장은 철광석, 원료탄, 연료탄, 곡물, 보크사이트, 마이너벌크 등 대표적인 화물 특성으로 인해 상반기 운임이 하반기보다 낮은 수준을 기록하는 계절적 패턴을 보인다.

인공지능 알고리즘 중에서도 시계열적 특징을 활용하기에 최적화된 구조인 순환신경망 기반의 모델을 구축해 데이터 특성을 최대한 활용함으로써 운임 예측의 정확도를 개선하고 더 나아가 새로운 방법론이 기존 시계열 모델의 한계를 보완할 수 있을지에 대한 가능성을 검토해 보았다.

순환신경망 알고리즘은 RNN(Recurrent Neural Network)과 LSTM(Long short term memory)을 이용해 차주 해상운임을 예측했으며, 동 모델이 건화물 시장의 변동성과 시계열적 특성을 학습할 수 있도록 기존의 기술분석 방법에 시계열 요소 분해법을 추가하여 데이터를 전처리했다.

기술분석 방법은 운임 데이터로부터 추세, 모멘텀과 같은 특징을 추출할 수 있게 했다면 추가된 전처리 방법은 운임 데이터로부터 추세와 계절, 순환, 불규칙성 등의 시계열적 특성을 추출 가능케 한다. 시장을 선박 크기별로 세분하였으며 이 연구를 통해 제안한 순환신경망 기반 해상운임 예측 모델은 DNN 모델과 성과를 비교하여 제시하였다.

43) 황수진 외(2020), pp. 1~65.

여기에 새롭게 추가한 데이터의 전처리 방법론에 대한 평가를 위해 모델의 입력변수로 사용할 데이터는 기술분석으로 운임을 분석한 데이터 그룹과 그 결과에 시계열 요소 분해법으로 운임을 분석한 데이터를 추가한 그룹으로 구분하고 개별 데이터 그룹은 순환신경망과 DNN의 학습 데이터로 사용해 예측 정확도 향상에 기여하는지를 확인하였다.

그 결과 이번 연구를 통해 새롭게 추가한 데이터 전처리 방법이 모델 성능을 개선하였으나 해운 데이터의 시계열적 특성을 활용할 수 있는 구조를 가진 순환신경망 알고리즘의 성능이 일반 신경망 알고리즘보다 뛰어난 성능을 보이지는 않았다.

2. 시사점

인공지능 알고리즘을 이용한 모델 구축 연구는 2019년부터 시작하였으며 2차까지 연구를 진행하던 시기에는 해상운임 예측 분야에는 인공지능을 적용한 경우는 그 수가 적거나 시작 단계에 불과했다. 이 부분에서 지난 2년 동안에 연구는 인공지능 기반 모델을 구축하고 그 결과를 실제 서비스했다는 점에서 의미가 있다. 인공지능 분야 연구의 초석을 다지는 계기가 되었다는 점에서도 의미가 깊다.

동 연구는 연구적 측면과 함께 실용적인 면도 고려했는데 대부분 건화물선 시장의 운임 예측 연구가 시장을 세분화하기보다는 건화물선 운임지수⁴⁴⁾를 예측하는 연구에 그쳐 해운 시장참여자가 실무에 활용하는 데에 한계가 있었다면 당시 연구는 이러한 시장의 특성을 고려하여 시장을 세분화

44) 건화물선운임지수(Baltic Dry Bulk Index)는 건화물선의 대표 선박인 케이프선, 파나막스선, 수프라막스선, 핸디선의 운임에 가중치를 부여해 하나의 지수로 나타냄으로써 전반적인 건화물선의 상황을 판단하는 근거로 활용한다. 2018년 3월 1일부터 핸디선을 제외한 케이프선, 파나막스선, 수프라막스선 시장의 운임만을 지수를 만드는데 사용하고 있다.

하고 각각의 개별 예측 결과를 제시해 기존 연구의 한계를 보완하였다.

건화물선 운임지수는 전체적인 상황을 판단하는 데에는 유용한 지표로 활용하나 실무 활용에는 한계가 있다. 실제 건화물선 시장은 선박의 크기 별로 케이프선, 파나막스선, 수프라막스선, 핸디선 시장으로 구분되어 있고 운송하는 화물에 따른 시장 특성의 차이가 상대적으로 크게 나타나기 때문이다.

예를 들어 건화물선은 대표적인 부정기선 시장으로 선박의 선적지와 목적지가 정해져 있지 않아 시장참여자가 시장에 노출되는 위험을 조정하는 데에 어려움을 겪는 것으로 알려져 있으나 케이프선은 상대적으로 주요 생산지와 소비지가 정해져 있는 화물인 철광석과 석탄을 운송하기 때문에 중소형 선박과 비교해 위험을 관리하기가 상대적으로 용이하다.

현재 해상운임 예측 서비스는 변동성이 큰 건화물선 시장을 중심으로 해운기업, 정부 등에 정보를 제공하고 있으며, 데이터 기반에 의사결정을 하고자 하는 구독자 수요에 맞게 자체 모델을 구축하고 매주 차주 운임의 방향성에 관한 정보를 제공하고 있다.

대부분 인공지능 알고리즘을 활용하여 구축한 해상운임 예측모델은 결측값 제거 또는 정규화 이외에 추가적인 데이터 전처리 작업 없이 입력값을 그대로 모델 구축에 활용한다. 반면 2019년부터 연구를 통해 해운빅데이터연구센터가 보유하고 있는 모델 특징은 변동성이 큰 해운시장의 운임에 대한 모델의 학습력과 예측 정확도를 높이하고자 기술분석을 추가하여 운임 데이터의 일일 변화 특징을 추출한다는 점이다. 다만 실제 시장의 미래 가치에 대한 심리를 보여주는 대표적 변수 중 하나인 FFA 가격 데이터에 대한 분석이 제대로 이루어지지 않아 이번 연구에서는 FFA 가격 데이터의 활용 범위의 확장을 시도하였다.

제4절 기존 연구와 차별성

1. 데이터 전처리 방식 개선

이번 연구에서는 새로운 데이터 전처리 방법을 도입하여 기존 모델이 기술분석을 사용함으로써 내포하고 있던 오류 가능성을 최소화하고자 하였다. 이를 위해 새로운 모델이 해운시장 변동성이 크다는 점에 주목하여 변동성 분석을 데이터 전처리 과정에 도입하고자 하였다.

기술분석은 분석자가 임의로 특정 구간을 설정하고 지표와의 관계를 이론적 근거 없이 경험에 의존해 판단하는 방식으로 같은 지표의 관계에도 분석자별로 그 해석의 차이가 발생한다. 이로 인해 모델이 기술분석 지표를 학습하는데 오류를 범할 가능성을 내포하게 된다.

2. 다양한 알고리즘을 활용한 다수의 예측 모델 구축

복수의 인공지능기반 알고리즘을 활용해 모델을 개발하고 확보함으로써 예측 서비스에 사용되는 모델의 단점을 보완하고 예측의 정확도를 높이고자 하였다. 이를 위해 인공지능 지도학습 알고리즘 중에서 신경망 계열 종류인 심층 신경망, 순환신경망을 활용해 단일 모델의 단점으로 지적되는 예측 결과에 편향과 분산 가능성 문제를 해결하는 데 초점을 두었다.

인공지능에는 다양한 알고리즘이 있으며 각 알고리즘은 모델을 구축하는 과정에서 입력변수의 선택, 데이터 처리방식 및 모델의 구성과 하이퍼파라미터 선택 등에 따라 그 성능이 달라진다. 다수의 모델을 구축하여 그 예측력을 비교함으로써 편향을 최소화하는 등 수시로 보유하고 있는 모델의 성

능을 검토하고 이후에도 최신 방법론을 도입하여 예측력을 보완해 나갈 수 있도록 환경을 조성할 필요가 있다.

3. 데이터의 활용성 확대를 통한 예측 정확도 향상

건화물선 시장참여자에 기대 심리변화를 모델의 학습에 활용하기 위해 선도거래 데이터 활용성을 확대했으며, 선도거래가 내포하고 있는 미래에 대한 시황정보를 활용하여 예측 기간의 정확도를 높여 기간을 확장하고자 하였다.

기존 모델도 운임과 함께 용선료⁴⁵⁾, 선도거래⁴⁶⁾ 데이터를 모델의 학습에 사용했으나 기술분석으로 변환하여 변동을 추출한 데이터는 운임에 한한다. 기존 모델이 시황의 변동 특징을 운임 변동에 집중해 활용했다면 본 연구에서는 미래 시장에 대한 참여자의 기대 심리가 반영된 선도거래 데이터의 활용성을 확대하여 해운 시황의 변화에 대한 모델의 학습력을 향상하고 예측의 정확도를 개선하고자 하였다. 또한 선도거래가 내포하고 있는 예측정보를 활용하여 새롭게 제안할 모델의 예측 기간을 확장하고자 시도하였다 세 가지 세부 목표를 달성하는 것에 초점을 두었다.

45) 클락슨은 6개월, 1년, 3년, 5년 단위 기준으로 용선료를 제공한다.

46) 시장참여자가 미래의 일정 시점에 미리 약정한 가격으로 물건을 매수하거나 매도하기로 계약하는 것을 말한다.

03

인공지능기반 해상운임 예측 모델 구축

제1절 개요

본 연구는 2장에서 설명한 KMI 해운 빅데이터연구센터 보유한 모델에
측력 향상과 예측 기간 확장을 통해 결과의 활용성을 높이 데에 집중하였
다. 모델 개선을 위해 변수선정과 변수 간의 관계를 재정의하고 데이터 전
처리에 변동성 분석을 새롭게 도입하였다. LSTM과 GRU 알고리즘에 기반
한 분류모델을 추가로 구축하였다. 제안하고자 하는 해상운임 예측 모델
구축과정을 아래 <표 3-1>과 같은 순서로 설명하고자 한다.

<표 3-1> 3장 구성

구분	세부 구성	목적
제1절 개요	• 개요 • 장 구성 및 설명	• 이번 장의 체계 설명
제2절 모델의 구축	• 모델의 구축과정 • 시사점	• 일반적인 인공지능 모델 구축과정 설명 • 제안 모델의 구축과정 설명 제시 • 구축과정 비교 제시를 통한 기존 모델 과 차별점 도출

구분	세부 구성	목적
제3절 데이터 및 전처리	• 사용 데이터 설명 • 데이터의 기초통계 • 데이터 전처리	• 모델에 사용한 데이터 설명 • 일반적인 데이터 전처리 방법설명 • 새로운 전처리 방법에 대한 상세한 설명 제시
제4절 인공지능 알고리즘	• 심층 신경망 알고리즘 • LSTM 알고리즘 • GRU 알고리즘	• 개별 알고리즘에 대한 설명

자료: 저자 작성

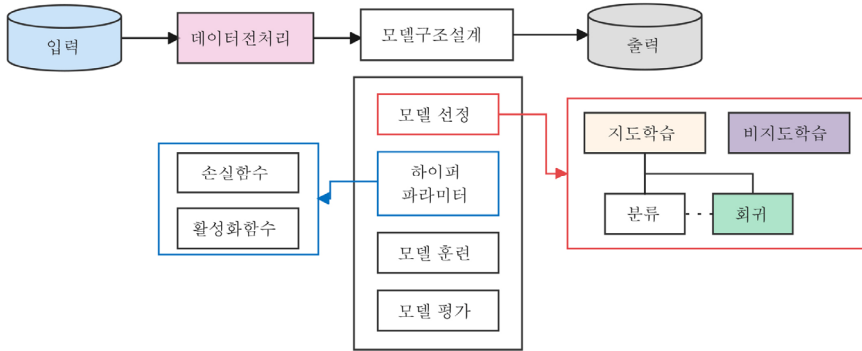
제2절 모델의 구축

일반적인 모델 구축과정을 상세히 설명하여, 본 연구를 통해 제안하는 모델이 기존 모델과 어떤 차별성을 갖는지 보여주고자 한다. <그림 3-1>에서 보는 바와 같이 인공지능 모델 구축과정은 데이터를 입력하고 나서부터, 전처리 및 모델 구조 설계, 훈련, 평가 및 출력과 같은 일련의 과정으로 나뉜다.

모델은 주어진 문제를 해결하기 위해 설립한 가정에 맞게 변수 및 인공지능 알고리즘 등을 선정하고, 입력된 데이터 값을 이용해 설계 의도에 따라 출력값을 산출하는지 검증을 거친다.

인공지능 모델은 데이터를 반복적으로 학습하면서 가중치(weight)와 편향(bias)을 갱신해 최적의 값을 정한다는 특징을 갖는다. 학습 과정에 따라 기계에 정답이 주어지는 지도학습 모델과 정답 없이 규칙을 생성해 나가는 비지도 학습 모델 두 종류로 구분할 수 있다. 또는 출력 형태에 따라 연속된 숫자를 답으로 산출하는 회귀모델과 결과를 범주형 데이터로 산출하는 분류모델로도 구분한다.

〈그림 3-1〉 일반적인 인공지능 모델 구축과정



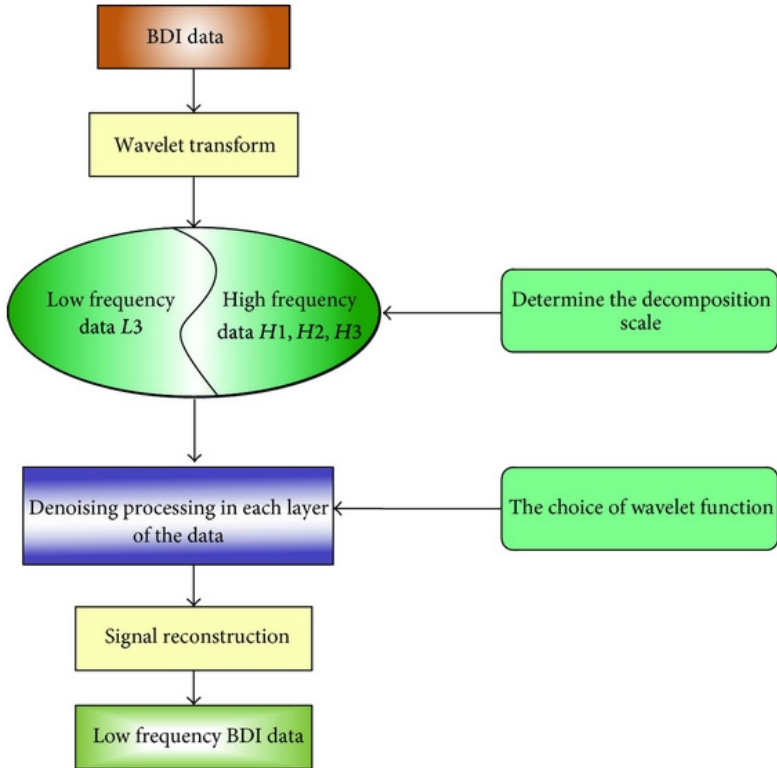
〈그림3-2〉는 데이터 전처리 과정을 상세하게 보여준다. 일반적으로 데이터 전처리는 데이터 결측값을 제거하거나 모델 구성 변수를 정규화하는 과정을 말한다. 전자는 오류 없이 모델 학습이 진행되게 하도록, 후자는 개별 데이터를 일괄적으로 동일 범위로 변환하여 입력변수 간 범위 차이로 인해 발생할 수 있는 모델 오류를 최소화하기 위해 수행된다.

최근에는 이러한 목적과 함께 모델이 스스로 입력데이터의 특징을 추출하고 학습하도록 설계하여 예측력을 향상하려는 시도가 이루어지고 있다. 대표적으로 활용하고 있는 신경망 계열 알고리즘은 주어진 입력변수를 행렬 또는 벡터, 차원 단위로 인식하여 패턴을 추출할 뿐, 스스로 데이터 특징을 추출하여 학습하도록 구조화되어 있지는 않기 때문이다.

Han *et al.*(2014)⁴⁷⁾은 신호 주파수를 분해할 때 사용하는 기법의 하나인 웨이블릿(Wavelet)을 전처리 과정에 추가하여 건화물선 운임지수를 분해해 특징을 추출하고 이를 입력변수로 활용하는 모델을 제안하였다(〈그림 3-2〉 참조).

47) Han *et al.*(2014), pp. 1-12.

〈그림 3-2〉 웨이블릿 분석을 활용한 데이터 특징 추출

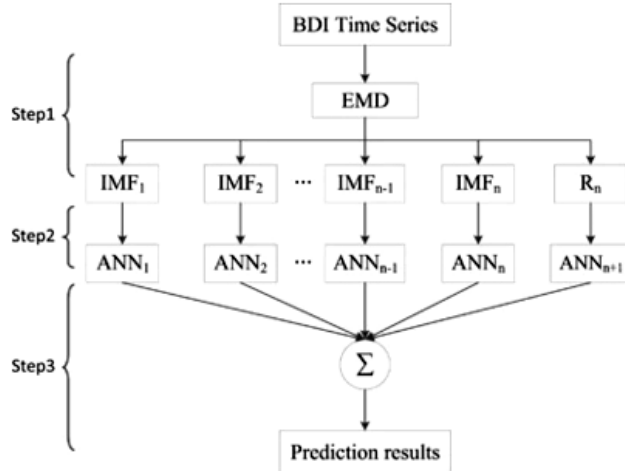


자료: Han *et al.*(2014), p.8

Zeng *et al.*(2016)⁴⁸⁾은 데이터 전처리 과정에 신호를 분석하는 경험적 모드 분해법(EMD, Empirical Mode Decomposition)을 추가하여 건화물선 운임지수를 예측하는 모델을 구축하였다. 경험적 모드 분해법을 이용해 건화물선 운임지수를 〈그림 3-3〉과 같이 내재 모드 함수(IMF, Intrinsic Mode Function) 집합과 잔차신호(Residual) 등으로 분해하여 이를 모델의 변수로 이용하여 향후 운임지수를 예측하는 모델을 제안하였다.

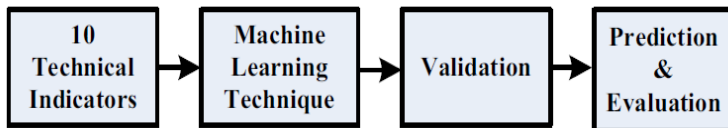
48) Zeng *et al.*(2016), pp. 192-210.

〈그림 3-3〉 EMD를 활용한 데이터 특징 추출

자료: Zeng *et al.*(2016), p.197

Shynkevich *et al.*(2017)⁴⁹⁾은 이동평균, 윌리엄스%R, 스토캐스틱 등 총 10가지 기술지표를 이용하여 주가 방향성을 예측하는 모델을 구축하여 예측성능을 높이하고자 시도하였다(〈그림 3-4〉 참조).

〈그림 3-4〉 기술분석을 활용한 데이터 특징 추출

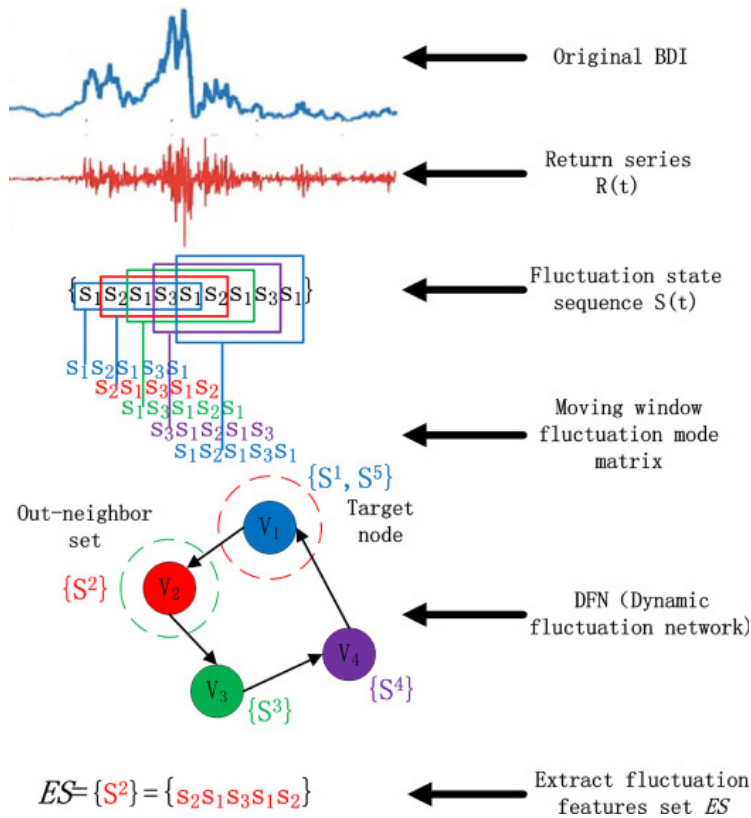
자료: Shynkevich *et al.*(2017), p.76

Zhang *et al.*(2019)⁵⁰⁾은 신호 데이터 분석 관점에서 운임 데이터가 예측 결과를 왜곡시키는 정보를 포함하고 있다고 가정하고 이를 제거하기 위

49) Shynkevich *et al.*(2017), pp. 71-88.50) Zhang *et al.*(2019), pp. 499-516.

해 주파수의 잡음을 제거하기 위해 사용하는 방법인 DFN(Dynamic Fluctuation Network)으로 데이터 추세, 주기, 오차항을 추출하고 이를 변수로 활용하는 모델을 구축하는 방법을 <그림 3-5>와 같이 제안하였다.

<그림 3-5> DFN를 활용한 데이터 특징 추출



자료: Zhang *et al.*(2019), p.503

이우식(2017)⁵¹⁾은 주가의 추세와 강도를 나타내는 상대강도지수, 모멘텀, 변화율 지표 등의 기술분석지표를 활용하여 한국 코스피 주가지수 예

51) 이우식(2017), pp. 287~295.

측 모델을 구축하여 데이터 전처리를 추가하는 방식이 예측성능 향상에 도움이 된다는 결론을 도출하였다.

〈표 3-2〉 데이터 전처리에 사용한 기술분석지표 상세

구분	내용
상대강도지수	• 주가 상승과 하락 간 이동을 비교하는 모멘텀 오실레이터(momentum oscillator)
모멘텀	• 주가 상승과 하락으로 움직일 때 속도 및 강도 파악에 사용하는 오실레이터
변화를 지표	• 시점 t 와 t_i 간 차이를 비교하여 주가 추세의 반전을 표시 (단, $i = 1, 2, 3, \dots, n$)
지수 이동평균	• 주식 추세(trend) 파악 및 기준 시점 t 에 가까운 데이터에 가중치 부여
장·단기 이동평균 교차	• 장기 이동평균과 단기 이동평균의 교차 시점에서의 주가 흐름 파악

자료: 이우식(2017), p.289의 내용을 바탕으로 저자 작성

김형준·류두진·조훈(2019)⁵²⁾은 시계열적 특징을 갖는 건화물선 운임지수 데이터의 특정 성분을 추출하기 위해 시계열 분해법을 데이터 전처리 과정에 추가하였다. 또한 입력 데이터에서 호드릭-프레스콧 필터(Hodrick- Prescott filter), 가법, 승법 모델 등으로 추세, 순환 등 성분을 추출해 모델의 입력변수로 활용하였으며, 이 과정을 통해 모델예측력이 향상했다는 결과를 제시하였다.

52) 김형준·류두진·조훈(2019), pp. 715~731.

〈그림 3-6〉 호드릭-프레스콧 필터를 활용한 데이터 특징 추출



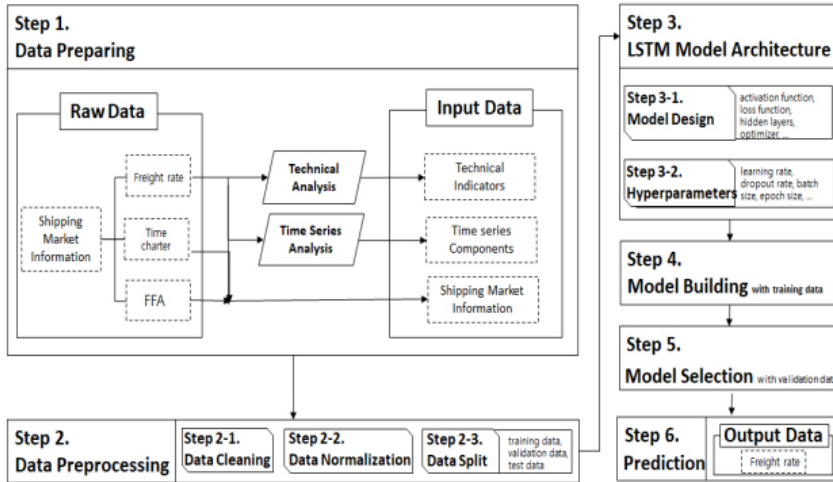
자료: 김형준 외(2019), p. 721

윤희성 외(2019)⁵³⁾는 기술분석을 데이터 전처리 과정에 추가하여 건화물 선박의 종류별 운임을 예측하는 모델을 구축하였으며, 황수진 외(2020)⁵⁴⁾는 기술분석과 시계열 분해법을 활용해 입력 데이터의 특징을 추출하여 건화물선 운임을 예측하는 모델을 구축하였다. 데이터 전처리 과정에 기술분석 또는 시계열 분해법을 추가해 변동성이 큰 건화물선 운임 데이터의 특징을 추출하는 것이 모델 예측성능 향상에 도움이 된다는 결론을 보여준다.

53) 윤희성 외(2019), pp. 1~81.

54) 황수진 외(2020), pp. 1~65.

〈그림 3-7〉 기술 분석 및 시계열 분해를 활용한 데이터 특징 추출



자료: 황수진 외(2020), p.39

이처럼 데이터 특징을 추출할 때는 DFN⁵⁵⁾, 웨이블릿⁵⁶⁾, EMD⁵⁷⁾ 등 주파수 분해법과 호드릭-프레스콧 필터, 가법 모델과 같은 시계열 분해법⁵⁸⁾, 주가 변화를 판단하는 기술분석⁵⁹⁾을 적용한다.

모델 변수 구성은 예측성능에 영향을 미치는 중요한 요소 중 하나이며, 최근 들어 데이터에서 추출한 특징을 모델 구성 변수로 활용하여 예측력을 향상하려는 연구가 늘고 있으나, 아직은 소수에 불과하다.

본 연구에서는 〈그림 3-8〉과 같이 모델의 데이터 전처리 과정에 변동성 분석을 추가하여 데이터 특징을 추출하고 이를 모델이 학습할 수 있도록

55) Zhang *et al.*(2019), pp. 499-516.

56) Han *et al.*(2014), pp. 1-12.

57) Zeng *et al.*(2016), pp. 192-210.

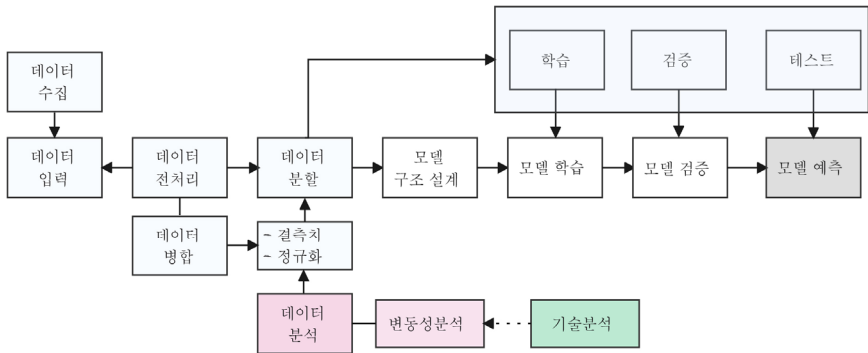
58) 김형준 외(2019), pp.715~731; 황수진 외(2020), pp. 1~65.

59) Shynkevich *et al.*(2017), pp. 71-88; 이우식(2017), pp. 287~295; 윤희성 외(2019), pp. 1~81; 황수진 외(2020), pp. 1~65.

알고리즘을 설계하였다. 변동성 분석⁶⁰⁾이란 금융 분야에서 시장을 분석할 때 일반적으로 사용하는 방법이며, 특정 기간 변화하는 상품 가격 간 차이를 정량화할 수 있다.

〈표 3-3〉에서 보는 바와 같이 기존 모델은 기술분석을 활용해 데이터 특징을 추출했다면⁶¹⁾, 본 연구를 통해 제안하는 모델은 확률밀도함수와 윌리엄스 % R과 함께 변동성 분석을 추가로 활용하여 시간 흐름에 따라 운임과 선물 가격 변화를 정량화하여 모델이 학습할 수 있도록 했다는 점에서 기존 모델과 차별화된다.

〈그림 3-8〉 모델 구축과정 중 데이터 전처리의 상세과정



자료: 저자 작성

〈표 3-3〉 데이터 전처리 비교

기존 모델 (기술적 변수)	제안 모델 (변동성 분석)
• (추세지표) 이동평균, MACD • (모멘텀지표) K모멘텀, 변화율, 상대강도지수, 윌리엄스 %R, 스토캐스틱	• 로그수익률, 역사적 변동성, 일간 변동성 • 변동폭, 변동비율, 변동범위 • 변동성 범위 내에서 움직일 확률

60) 이한득(2015), p. 3.

61) 윤희성 외(2019), pp. 1~81; 황수진 외(2020), pp. 1~65.

기존 모델 (기술적 변수)	제안 모델 (변동성 분석)
• (기타) 백분위, 회귀계수, 로그수익률, 표준편차, 장단기 용선료 비율	• 베이시스, 스프레드 • 확률밀도함수 • 윌리엄스 %R

자료: 황수진(2020), p.15; 이한득(2015), pp.1~14; 매일경제(2021.7.1.); 우준명(2013), p.1의 내용을 바탕으로 저자 작성

제3절 데이터 및 전처리

1. 데이터 개요

본 연구는 건화물선 선형별 해상운임을 예측하는 모델을 제시하기 위해 수행하였으며, <표 3-4>에서 보는 것처럼, 모델을 구축하기 위해 선형별 운임과 용선료, 운임선도거래(FFA, Forward Freight Agreement)를 데이터로 사용하였다. 각 변수는 입력변수로 활용하기 위해 두 단계의 데이터 전처리 과정을 거치게 된다.

모델을 구축하는 데 사용한 데이터의 기간은 2011년 5월 10일부터 2021년 6월 30일까지이며, 운임과 FFA 데이터는 2,532개(일 단위), 용선료는 2011년 5월 13일부터 2021년 7월 2일까지 530개(주 단위)를 활용하였다.

〈표 3-4〉 데이터 요약

변수	설명	출처	단위
선형별 운임	케이프선 5TC, 파나마스선 4TC	클락슨 (Clarkson)	\$/day
용선료	단기용선료(6개월) 케이프선 18만 dwt ^{주)} 장기용선료(5년) 파나마스선 7.5K dwt	클락슨	\$/day
운임선도거래	케이프선 당월물, +1월물, +2월물 파나마스선 당월물, +1월물, +2월물	볼틱해운거래소 (Baltic Exchange)	\$/day

주: 케이프선 장단기용선료는 2014년 7월 4일부터 18만 DWT를 사용하고 이전 데이터는 17만 DWT를 사용
자료: 볼틱, 클락슨(2021. 7. 1)의 자료를 바탕으로 저자 작성

사용한 데이터에 기초통계에 관한 내용은 〈표 3-5〉와 〈표 3-6〉에서 제시하였다.

〈표 3-5〉 케이프선 기초통계 요약

단위: 달러/일

항목	스팟운임 (5TC)	FFA			용선료	
		당월물	+1개월	+2개월	6개월	5년
최소값	1,985	1,748	2,382	3,432	4,250	9,875
하위 25%	6,801	7,119	7,984	8,800	11,625	14,500
중위수	12,120	12,259	12,865	12,550	16,000	15,750
상위 25%	18,540	18,646	18,063	17,417	21,750	17,375
최대값	44,817	45,750	44,646	42,286	42,000	25,000
평균	13,790	13,744	13,948	13,932	17,200	16,076
표준편차	8,487	7,838	7,082	6,559	7,122	3,279

자료: 볼틱, 클락슨(2021. 7. 1)의 자료를 바탕으로 저자 작성

〈표 3-6〉 파나막스선 기초통계 요약

단위: 달러/일

항목	스팟운임 (4TC)	FFA			용선료	
		당월물	+1개월	+2개월	6개월	5년
최소값	2,260	2,515	2,595	3,265	4,625	7,550
하위 25%	6,312	6,439	7,053	7,375	9,050	9,950
중위수	8,917	9,071	9,466	9,681	11,125	11,250
상위 25%	11,977	11,796	12,140	12,065	13,500	12,375
최대값	35,738	37,639	37,314	37,089	34,375	15,000
평균	9,513	9,556	9,917	10,016	11,528	11,170
표준편차	4,397	4,184	4,171	3,927	3,915	1,887

자료: 볼틱, 클락슨(2021. 7. 1)의 자료를 바탕으로 저자 작성

2. 데이터 전처리

본 연구에서 데이터 전처리 과정은 데이터 정규화와 데이터 특징 추출 두 단계로 나뉜다. 데이터 정규화는 서로 다른 데이터의 범위를 일괄적으로 통일하는 과정을 말하며, 데이터 특징 추출 단계에서는 변동성 분석을 이용하여 시간 흐름에 따른 개별 변수의 변화 정도를 정량화하여 모델이 학습할 수 있도록 한다.

1) 데이터 전처리 1단계: 정규화

(1) 데이터 정규화의 개념과 목적

데이터 정규화 과정은 서로 다른 데이터 값의 범위를 일괄적으로 통일하여 모델을 구축하는 과정에서 오류를 방지하고 학습 속도를 향상하기 위해 수행하는 과정이다.

인공지능 모델은 최적의 가중치와 편향 값을 정하기 위해 주어진 입력값

을 반복적으로 학습하며, 모델 구성 변수 간의 관계를 정의하고 상대적인 가중치를 부여한다. 이 과정은 예측하고자 하는 값을 좌우하는 모델 구성 변수의 영향력 크기를 결정한다는 점에서 매우 중요한 단계이다.

이 과정에서 입력값 간의 범위 차이가 크면 모델은 변수의 중요도를 결정하는 과정에서 크기가 큰 변수에 가중치를 더 크게 부여할 가능성이 커지며 모델의 예측력이 현저히 떨어지거나 왜곡된 예측값을 산출할 수 있으므로 각 변수 범위를 일괄적으로 통일하는 과정이 필요하다.

데이터를 정규화하는 대표적인 세 가지 방법은 최소-최대 정규화(Min-Max Normalization), Z-점수 정규화(Z-Score Normalization)와 최소-최대 정규화 변형이 있으며, 본 연구에서는 데이터의 범위를 [-1, 1]로 변환하는 최소-최대 정규화 변형 방법을 사용하였다.

(2) 데이터 정규화의 종류

가. 최소-최대 정규화(Min-Max Normalization)

최소-최대 정규화는 데이터를 변화하는 가장 대표적인 방법으로 개별 데이터 범위를 아래 식을 이용하여 [0, 1]로 변환한다.

$$X'_t = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}}$$

여기서,

X_t 는 시점 t에서 데이터

$X_{\min}$ 데이터 최소값

$X_{\max}$ 는 데이터의 최대값

X'_t 는 변환된 t 시점의 데이터이다.

나. Z-점수 정규화(Z-Score Normalization)

Z-점수 정규화는 개별 데이터를 아래의 식을 이용하여 평균이 0이고 분산이 1인 표준정규분포로 변환하는 방법을 말한다.

$$X'_t = \frac{X_t - \mu}{\sigma}$$

여기서,

X_t 는 시점 t에서 데이터

μ 는 데이터 평균

σ 은 데이터 표준편차

X'_t 는 변환된 t 시점에 데이터를 말한다.

다. 최소-최대 정규화 변형

최소-최대 정규화의 변형은 데이터의 범위를 [-1, 1]로 변환하는 방식을 말하며 아래 식을 이용하였다.

$$X'_t = \frac{2X_t - (X_{\max} - X_{\min})}{(X_{\max} - X_{\min})}$$

여기서,

X_t 는 시점 t에서 데이터

$X_{\min}$ 는 데이터 최소값

$X_{\max}$ 는 데이터 최대값

X'_t 는 변환된 t 시점 데이터를 말한다.

2) 데이터 전처리 2단계: 데이터의 특징 추출

입력 데이터인 운임, FFA, 용선료의 특징을 추출하기 위해 변동성 분석을 데이터 전처리 과정에 추가하였다.

변동성(Volatility)은 일정 기간에 가격 변화의 정도를 수치화한 것을 말하며, 데이터 변동 분석은 제안하는 모델이 건화물선 시황의 변화를 학습할 수 있도록 개별 데이터의 시간 흐름에 따른 변화를 다양한 방식을 이용해 수치화하여 나타내는 과정을 말한다.

시간의 흐름에 따른 변화를 수치화하여 나타내는 방법은 주로 금융시장에서 경기상황을 판단할 때 사용하며 이를 변동성 분석이라고 한다. 변동성은 시장 참가자의 기대 심리가 반영되어 서로 다른 양상으로 나타나기 때문에 경기상황을 판단하기 위한 정보로 활용된다.⁶²⁾

이 점에 착안하여 본 연구에서는 전처리 과정에서 해운 시황을 판단하는 중요한 정보로 운임과 FFA 변동성을 추출하는 방법을 활용하며, 사용한 식은 <표 3-7>과 같다.

<표 3-7> 변동성 분석

구분		수식
1	로그수익률	$\ln(X_t) - \ln(X_{t-1})$
2	기간 일종변동성	$\frac{1}{(n-1)} \sum_{t=1}^n (\ln(X_t/X_{t-1}))$
3	역사적 변동성 (기간별 변동성)	$\left(\sqrt{\frac{1}{(n-1)} \sum_{t=1}^n (\ln(X_t/X_{t-1}) - \mu)^2} \right)$
4	일간 변동성 (변동성 연율화)	$\left(\sqrt{\frac{1}{(n-1)} \sum_{t=1}^n (\ln(X_t/X_{t-1}) - \mu)^2} \right) \times \sqrt{m}$

62) 우준명(2013), 「금융시장 변동성의 경기 예측력」, BOK 경제연구 제2013-19, p. 1.

구분		수식
5	변동폭	$\frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1})$
6	변동비율	$(\frac{1}{n} \sum_{t=1}^n X_t) / (\frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1}))$
7	변동성 범위 내에서 움직일 확률	$\frac{(n-1) - \alpha}{(n-1)}$ $\alpha = (n-1) - \sum_{t=1}^n \gamma_n$ $\gamma = \begin{cases} 1, & \text{if } X_t > \frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1}) \\ 0, & \text{if } X_t \leq \frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1}) \end{cases}$
8	변동범위	$\ln(V_t) - \ln(V_{t-1})$ $V_t = H_t - L_t$
9	윌리엄스 %R	$\left(\frac{H_t - X_t}{V_t} \right) \times 100$
10	확률밀도함수	$\frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{X_t - \mu}{2\sigma^2}\right)$
11	베이지스	$X_{i,t} - X_{j,t}$
12	스프레드	$X_{i,t+k} - X_{i,t}$

자료: 매일경제(2021.7.1.)의 내용을 바탕으로 저자 작성

3. 변동성 분석 적용 결과

변동성을 확인하는 것이 주목적이기 때문에 케이프선 운임 데이터(일 단위)를 활용한 실증분석 결과를 제시하여 변동성 분석의 필요성을 확인하였다.

1) 일별 변화율

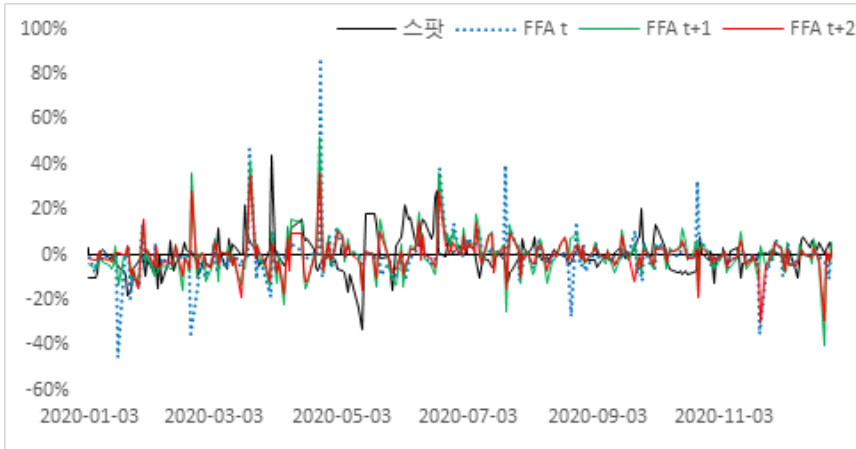
일별 변화율은 로그 수익률을 이용하였다. 로그 수익률은 금융 분야에서 수수와 로그의 성질을 활용하여 수익률을 계산할 때 주로 사용하는 방법이며, 식으로 표현하면 아래와 같다.

$$X_{i,t} = \ln(X_t) - \ln(X_{t-1})$$

시점 t 기준 X 값의 시간에 따른 변화 정도를 정량화하기 위해 사용하며, 시점별 변수 X 에 자연로그를 취한 후 그 차이를 이용하여 계산한다. 로그 수익률은 일중변동성과 일간 변동성 등 변동성 분석을 위한 데이터 전처리 과정에 사용되는 12가지 분석 방법에 가장 기본이 되는 식이자 자료이다.

〈그림 3-9〉는 케이프선 스팟운임과 FFA 가격에 로그 수익률을 보여준다. 여기서 스팟은 케이프선의 5TC 운임을 의미하며, FFA_t 는 t 시점의 FFA 가격을, FFA_{t+1} 은 t 시점을 기준으로 1개월 후 FFA 가격, FFA_{t+2} 는 t 시점을 기준으로 2개월 뒤의 FFA 가격을 말한다. 2020년 한 해 동안에 로그 수익률을 살펴보면 스팟운임의 t 시점을 기준으로 FFA 가격의 시점의 차이가 작을 때 로그 수익률의 크기가 크게 나타나는 경향이 있음을 알 수 있다.

〈그림 3-9〉 케이프선 변수 일별 변화율



자료: 저자 작성

2) 기간 일중변동성

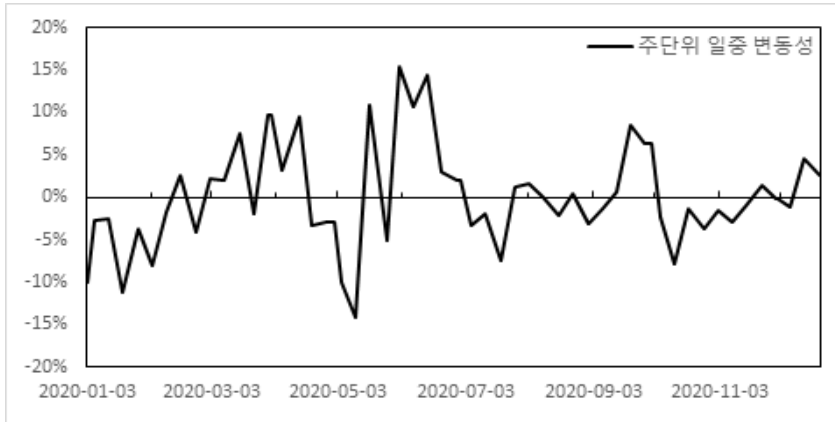
기간 일중변동성은 설정한 t 기간에 데이터 로그수익률에 합을 평균화한 수치로 아래의 식으로 정리할 수 있다.

$$\frac{1}{(n-1)} \sum_{t=1}^n (\ln(X_t/X_{t-1}))$$

기간 일중변동성을 데이터 전처리 과정에 추가함으로써 설정 기간 동안 운임과 FFA 가격 변화 정보를 모델에 학습시켜 모델이 운임을 예측할 때 더 많은 정보를 활용할 수 있도록 하여 예측력을 높이하고자 하였다.

〈그림 3-10〉은 2020년 1년 동안 케이프선 스팟운임에 5일 기간 일중변동성을 나타내며, 변동성이 가장 크게 나타나는 시기가 5월과 6월인 것으로 파악되었다.

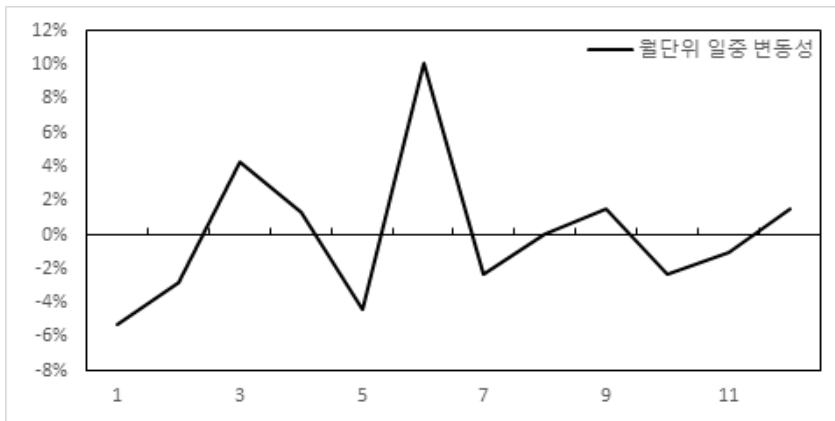
〈그림 3-10〉 케이프선 운임 5일간 일중변동성($n=5$)



자료: 저자 작성

〈그림 3-11〉은 동일 선종 운임에 30일 동안 일중변동성을 보여준다. 이처럼 기간 일중변동성 그래프를 통해 2020년 1년 동안 케이프선 운임의 변동성이 가장 큰 시기가 6월인 것으로 확인하였다.

〈그림 3- 11〉 케이프선 운임 30일간 일중변동성($n=30$)



자료: 저자 작성

〈그림 3-10〉과 〈그림 3-11〉을 비교했을 때, 가장 변동이 큰 6월에도 주별 변동성의 크기가 다름을 확인할 수 있다. 이처럼 기간 일중변동성은 분석자가 의미 있다고 판단하는 기간을 임의로 설정할 수 있어 단기간과 장기간에 상대적 가격 변화 차이를 모델이 학습할 수 있다는 점을 확인할 수 있다.

3) 역사적 변동성(기간별 변동성)

역사적 변동성은 일정 기간 대상 가격 수익률에 표준편차를 이용하며,⁶³⁾ 이를 아래의 식과 같이 정리할 수 있다.

$$\left(\sqrt{\frac{1}{(n-1)} \sum_{t=1}^n (\ln(X_t/X_{t-1}) - \mu)^2} \right)$$

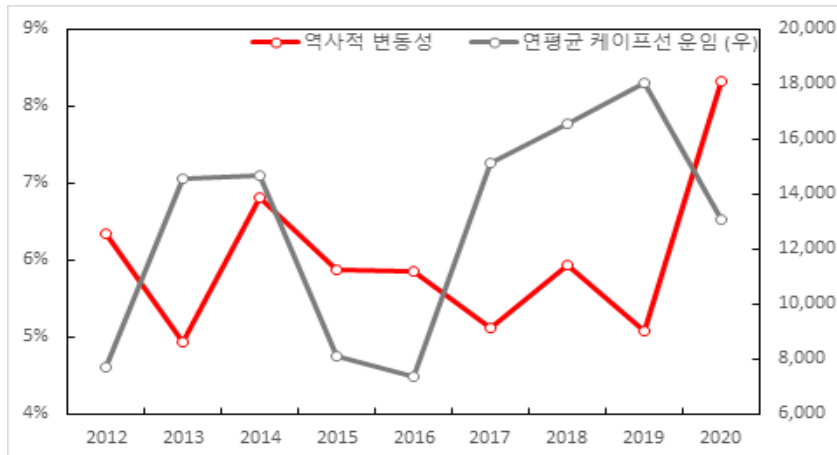
역사적 변동성은 표준편차의 특성을 그대로 적용하며 특정 기간의 평균 값에서 얼마만큼의 편차가 있는지를 백분율로 표시하여 나타냄으로써 대상 변수의 기간별 역사적 변동성을 비교할 수 있다. 또한 기간별 시장의 위험을 비교하고 모델의 구축과정에서 상대적인 입력 데이터의 특정 기간별 상대적 변화의 정도를 학습할 수 있도록 활용하였다.

〈그림 3-12〉는 2012년부터 2020년까지 케이프선 연평균 운임과 역사적 변동성을 비교한 결과이다. 2020년 코로나19 확산과 기상이변 등으로 인해 불확실성이 확산하면서 시장에 투자심리가 위축되던 시기의 변동성이 그 어느 때보다 높았던 점을 확인할 수 있다. 또한 연평균 운임의 큰 폭으로 상승하거나 하락하기 직전의 전년도 변동성이 높거나 낮았다는 점도 직관적으로 파악가능하다.

63) 김상국·박현우·성태웅(2013), p. 153.

이와 같은 점에 착안하여 역사적 변동성을 활용해 모델이 시장의 변화를 학습할 수 있으며 특히 운임의 예측 기간을 늘리는 데에 유용한 입력변수로 활용 가능함을 확인하였다.

〈그림 3-12〉 케이프선 스팟운임의 역사적 변동성(2012년-2020년)



자료: 저자 작성

4) 일간 변동성

앞서 역사적 변동성은 수익률의 표준편차의 개념을 적용했다고 설명하였다. 즉 수익률이 정규분포의 형태를 따른다고 가정한 것이며, 서로 다른 기간에 누적 수익률을 비교하는 방법으로 변동성 연율화 방식을 활용한다. 변동성을 연율화로 표시하기 위해서는 아래와 같은 식을 이용한다.

$$\left(\sqrt{\frac{1}{(n-1)} \sum_{t=1}^n (\ln(X_t/X_{t-1}) - \mu)^2} \right) \times \sqrt{m}$$

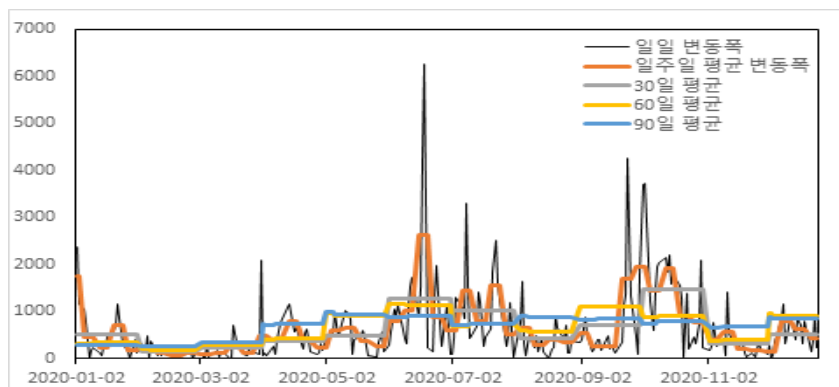
5) 변동폭

변동폭은 주, 월, 연간 데이터에 평균 변화 폭의 크기를 알려주기 위해 활용되며 이를 식으로 나타내면 아래와 같다.

$$\frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1})$$

변동폭은 특정 분포를 가정하지 않고 있다. 따라서 비모수적 척도로서 변동성의 추정치를 얻을 수 있다는 장점이 있어 변동성 분석에 활용된다.⁶⁴⁾ <그림 3-13>은 2020년 1년 동안 케이프선 스팟운임의 기간별 변동폭을 비교한 결과이다. 1년 동안의 일간, 일주일, 30일, 50일, 60일, 90일간 변동폭 평균을 비교하였다. 예를 들어 2020년 6월에는 코로나19로 인해 인도 등의 항만 봉쇄와 검역 및 방역 강화로 시장에 불확실성이 확대된 시점임을 확인 가능하였다.

<그림 3-13> 케이프선 스팟운임의 기간별 변동폭 비교(2020년)



자료: 저자 작성

64) 조담(2011), p. 1572.

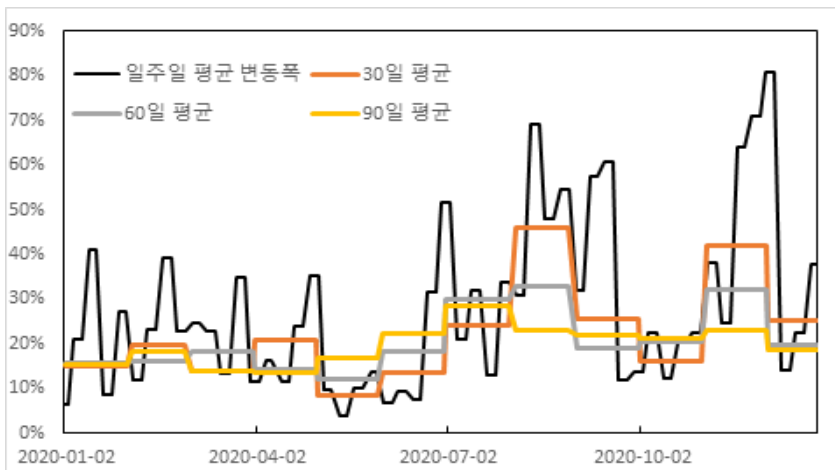
6) 변동비율

변동비율을 나타내는 식은 아래와 같다.

$$\left(\frac{1}{n} \sum_{t=1}^n X_t\right) / \left(\frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1})\right)$$

2020년 1년 동안 케이프션 스팟운임 변동비율은 <그림 3-14>와 같다. 일주일, 30일, 60일, 90일의 변동폭을 비교하여 나타내었으며 기간을 90일로 지정한 이유는 분기 예측을 고려하여 3개월까지의 장기 정보를 모델이 학습할 수 있도록 하기 위함이다.

<그림 3-14> 케이프션 스팟운임의 기간별 변동비율 비교(2020년)



자료: 저자 작성

7) 변동성 범위 내에서 움직임 확률

변동성 범위 내에서 움직임 확률은 아래의 식과 같이 나타낸다.

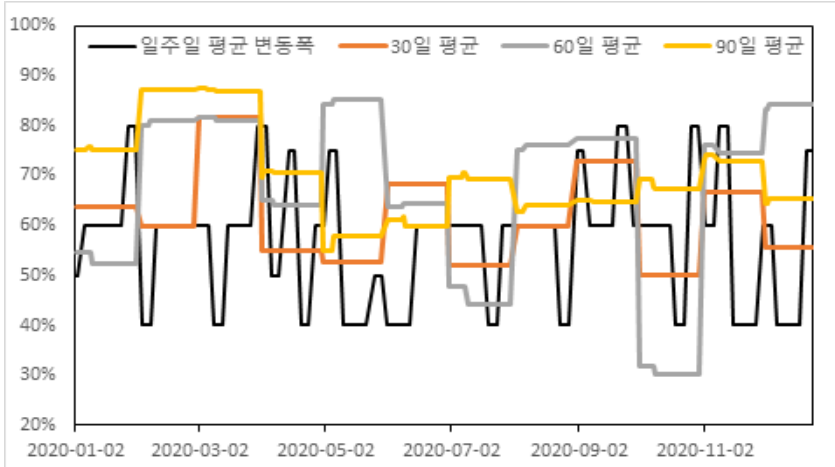
$$\frac{(n-1)-\alpha}{(n-1)}$$

$$\alpha = (n-1) - \sum_{t=1}^n \gamma_n$$

$$\gamma = \begin{cases} 1, & \text{if } X_t > \frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1}) \\ 0, & \text{if } X_t \leq \frac{1}{(n-1)} \sum_{t=1}^n (X_t - X_{t-1}) \end{cases}$$

케이프선 스팟운임이 2020년 1년 동안 변동성 범위 내에서 움직일 확률을 계산한 결과이다(〈그림 3-15〉 참고). 변동성은 일주일, 30일, 60일, 90일을 기준으로 계산하였으며 기간별 평균을 구한 뒤 비교하여 나타냈다.

〈그림 3-15〉 케이프선 스팟운임의 변동성이 범위 내에서 움직일 확률(2020년)



자료: 저자 작성

8) 변동범위

변동범위는 특정 기간 t 동안에 고점과 저점 차이를 자연로그를 이용하여 표현하며, 이를 식으로 나타내면 아래와 같다.

$$\ln(V_t) - \ln(V_{t-1})$$

$$V_t = H_t - L_t$$

변동범위는 데이터 변화 정도를 비교하는데 용이하여 기술분석에서도 활용되는 방식으로 본 연구에도 데이터 전처리에 활용하였다.

9) 윌리엄스 %R

윌리엄스 %R은 일정 기간 동안 주가 변화 대비 고가와 현재 주가의 차이를 이용하여 나타낸다. 동지표의 특징 중 하나는 주가가 고점일 때를 기준으로 계산된다는 것과 주가의 변화를 -100에서 0으로 표현한다는 점이다. 이를 수식으로 표현하면 아래와 같다.

$$\left(\frac{H_t - X_t}{V_t} \right) \times 100$$

10) 확률밀도함수

확률밀도함수는 확률론에서 확률 변수 분포를 나타내는 함수로 아래의 식과 같이 표현할 수 있다.

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{X_t-\mu}{2\sigma^2}\right)$$

11) 베이스스⁶⁵⁾

베이스스(Basis)는 현재 시세로 계약을 체결하는 현물과 미래 시점에 가치에 맞는 시세로 계약을 체결하는 선물 상품 간 가격 차이를 말하며, 이를 정리하면 아래 식과 같이 표현할 수 있다.

$$Basis = X_{i,t} - X_{j,t}$$

여기서,

$Basis$ 는 베이스스

$X_{i,t}$ 는 시점 t 에서 현물 가격 X_i

$X_{j,t}$ 는 시점 t 에서 선물 가격 X_j 을 말한다.

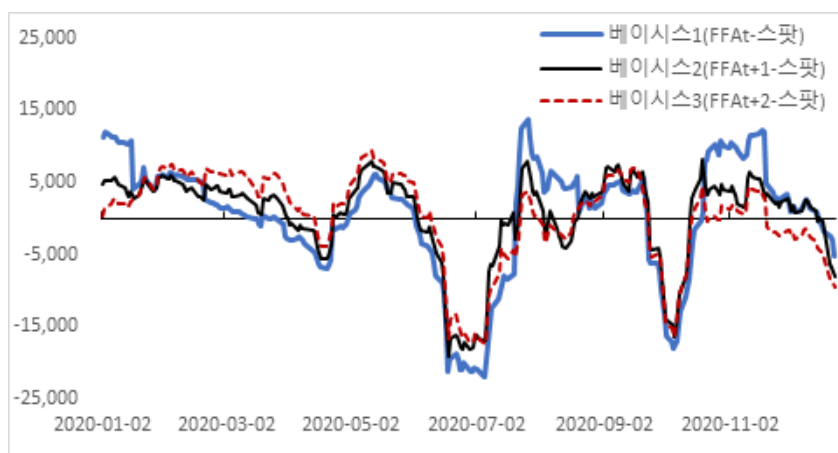
선물은 현물보다 고평가되는 것이 일반적이는데 이는 선물 가격에 현물 가격에 대한 미래 시점의 불확실성에 대한 보상이 포함된 개념이기 때문이다. 이러한 시장 상태를 콘탱고(Contango)라고 말하며, 이와 반대로 현물이 선물 가격 더 높게 형성된 비정상 시장을 백워드이션(Backwardation)이라고 한다.

개념적으로 단순하지만 실제로는 차익거래의 방향을 결정짓는 중요한 변수이며, 일반적으로 콘탱고 상태는 투자자들이 심리가 긍정적인 때, 백워드이션은 비정상 시장 또는 역조 시장 상태로 투자자들 심리가 위축된 경우로 판단한다.

65) 매일경제(2021. 7. 1)의 내용을 바탕으로 저작 작성 <https://www.mk.co.kr/dic/m/view.php?idx=6362>

〈그림 3-16〉은 총 세 가지 형태의 베이스스를 보여주고 있는데 t 시점 케이프션 스왑운임을 기준으로 동일 t 시점의 FFA 가격과의 베이스스 1과 1개월 뒤 시점인 $t+1$ 시점의 FFA와의 베이스스2, t 시점에서 2개월 뒤의 $t+2$ 시점의 FFA 가격과의 베이스스 3이다. 근월물에 해당할수록 베이스스 차이가 큰 정상적인 시장과 역조 시장이 나타나 의미 있는 정보로 활용 가능함을 보여주고 있다.

〈그림 3-16〉 케이프션 스왑운임과 FFA가격간의 베이스스



자료: 클락슨, 볼틱(2021. 7. 1)의 자료를 바탕으로 저자 작성

12) 스프레드⁶⁶⁾

스프레드(Spread)란 서로 다른 선물 가격의 차이를 말하며, 이를 식으로 표현하면 아래의 식과 같다. 즉 원월물 가격에서 근월물 가격을 뺀 수치를 말하며 이는 일반적으로 원월물 가격이 근월물의 가격보다 이론적으로 높게 형성되기 때문이다.

66) 신한금융투자(2021. 7. 1)의 내용을 바탕으로 저자 작성 http://m.shinhaninvest.com/edu/invest/option/strategy_03.html

$$Spread = X_{i,t+k} - X_{i,t}$$

여기서,

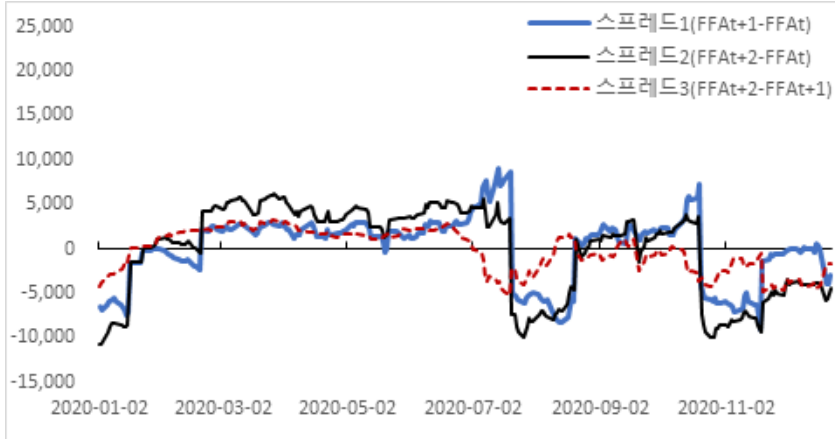
$Spread$ 는 스프레드

$X_{i,t}$ 는 시점 t 에서 현물 가격 X_i

$X_{i,t+k}$ 는 시점 t 에서 k 주 후 현물 가격 X_i 를 말한다.

〈그림 3-17〉은 총 세 가지 케이프선 FFA 가격 스프레드를 보여주고 있다. 베이스와 마찬가지로 스프레드도 현재 시황이 정상인지 역조인지를 판단할 수 있는 기준으로 활용될 수 있다는 점에서 모델 학습에 의미 있는 정보로 판단된다.

〈그림 3-17〉 케이프선 FFA가격의 스프레드



자료: 클락슨, 볼틱(2021. 7. 1)의 자료를 바탕으로 저자 작성

2) 시사점

변동성 분석 결과를 통해 시장을 쉽게 직관적으로 이해하고 운임 방향성의 크기를 수치화하기도 하지만 특정 시점에서의 상대적인 운임의 변화를 정량화하여 잘 나타내 주고 있음을 확인하였다. 같은 기간에도 서로 다른 크기로 정량화된 변동성 크기를 모델이 학습할 수 있기 때문에 해운 데이터 특징을 추출하는데 적합한 방법으로 판단한다.

제4절 인공지능 알고리즘

제4절에서는 다양한 알고리즘에 기반한 모델을 구축하기에 앞서 연구에서 사용하는 알고리즘에 관해 자세히 설명하고자 한다.

모델 구축에 사용하는 알고리즘은 심층 신경망(DNN, Deep Neural Network), 장단기메모리 신경망(LSTM, Long Short Term Memory), 게이트 순환신경망(GRU, Gated Recurrent Unit) 총 세 가지이다. 여기서 DNN과 LSTM은 신경망 계열에 대표 알고리즘이며, 특히 LSTM은 시계열 데이터를 분석하는 데 활용하며, GRU는 기존 LSTM과 비교하여 향상된 계산 성능으로 주목받고 있는 알고리즘이다.

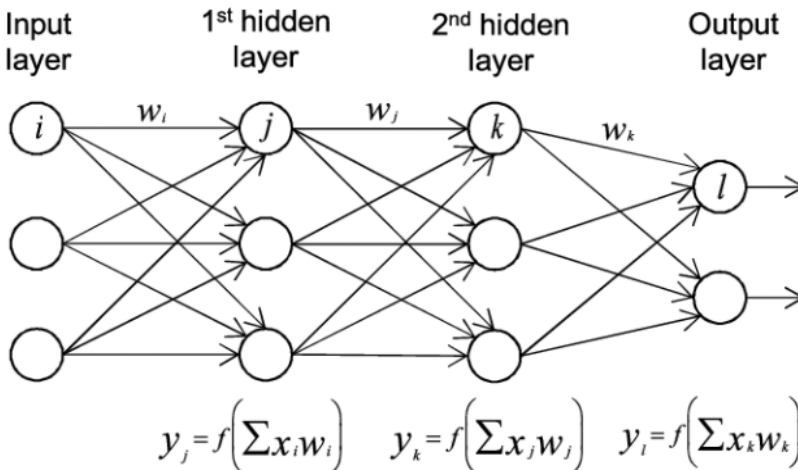
1. DNN

DNN은 <그림 3-18>과 같이 입력층, 출력층과 두 개 이상에 은닉층으로 이루어진, 즉 다층구조를 가진 인공지능 알고리즘이다. 한 개의 입력층

(Input layer)과 출력층(Output layer) 그리고 두 개의 은닉층(hidden layer)으로 구성된 단순한 구조의 DNN 모델이다. 입력층에서 세 개의 입력값(i)을 받아 최종적으로 2개의 출력값(l)을 계산하도록 설계되어 있으며, 층별 입력되는 값과 출력되는 값의 개수를 노드라고 한다. 일반적으로 y 값을 산출하는 과정에서 입력값에 가중치를 곱한 후 편향 값을 더하는데 <그림 3-18>에는 편향을 표현하지 않았다.

DNN 알고리즘은 입력된 데이터를 컴퓨터가 반복적으로 학습하여 가중치와 편향을 최적화한다. 즉 컴퓨터는 학습할 때마다 다른 가중치와 편향치를 사용하여 값을 예측하고 그 결과를 주어진 정답과 비교하는 과정을 반복한다. 이 과정에서 예측 결과와 주어진 정답의 차이가 가장 작을 때를 최종모델로 판단하고 학습을 멈춘다. DNN의 구조와 이를 식으로 표현하면 아래와 같다.

<그림 3- 18> DNN 모델의 구조



자료: Sandra Vieira.,etc(2017), p. 61

각 층은 입력값을 받아 최종값을 산출하기 위해 아래의 식을 이용한다.⁶⁷⁾

$$y_j = f\left(\sum x_i w_i\right)$$

$$y_k = f\left(\sum x_j w_j\right)$$

$$y_l = f\left(\sum x_k w_k\right)$$

여기서,

y_l 는 최종 출력값

y_j, y_k , 는 은닉층의 입력값에 가중치를 곱한 값들에 합(weighted sum)

을 활성화 함수(f)를 이용해 다음 층 노드에 전달되는 값

i 는 입력값

x_i, x_j, x_k 는 각 층의 입력값에 가중치를 곱한 값

w_i, w_j, w_k 는 가중치를 말한다.

이를 하나의 식으로 정리하면 아래와 같다.

$$\mathbf{a} = \mathbf{W}^T \mathbf{x}$$

여기서,

$\mathbf{a}$ 는 출력값

$\mathbf{W}$ 는 가중치

$\mathbf{x}$ 는 입력값을 말한다.

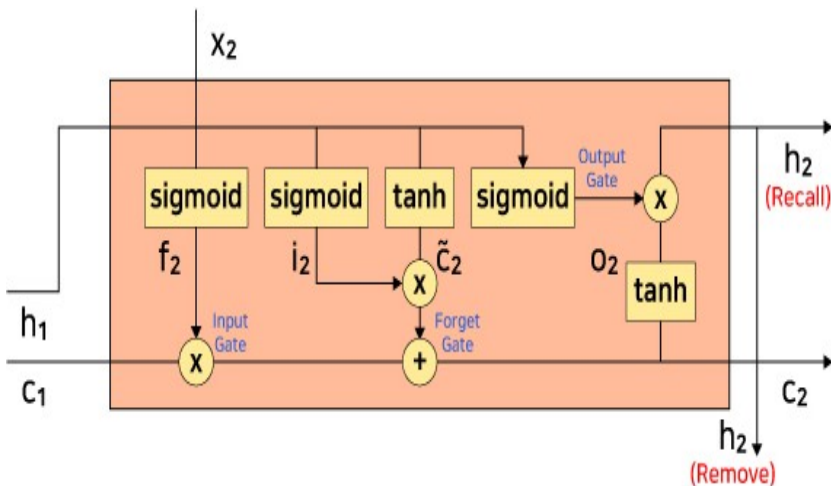
67) Sandra Vieira., etc(2017), p. 61.

2. LSTM

LSTM은 시계열 데이터를 처리하기 위해 구조화된 인공지능 알고리즘이며, 다층구조로 이루어져 있다는 점에서 DNN과 같다. 반면 <그림 3-19>와 같이 은닉층에 셀이라는 가상의 정보 기억 공간이 있다는 점에 차이가 있으며, 이를 셀스테이트(Cell state)라고 부른다.

LSTM 셀스테이트는 입력(Input Gate), 망각(Forget Gate), 출력게이트(Output Gate) 세 개의 요소로 구성되어 있으며, 과거와 현재의 입력 정보 중에 중요한 정보를 판단하여 그 크기를 결정하며, 크기는 시그모이드와 쌍곡탄젠트 함수를 사용한다.

<그림 3-19> LSTM 셀스테이트(Cell state)의 구조



자료: 이우섭·김형규(2021), p. 97

입력게이트는 현재 시점 t 에서 입력된 정보(x_t)를 시그모이드 함수에 넣은 결과값(i_t)과 이전 시점 은닉층에 저장된 정보(h_{t-1})를 쌍곡탄젠트 함수

에 넣은 결과값($\tilde{c}_t$)을 곱하는 역할을 하며, 현재 입력된 정보에 가중치를 결정한다. 이를 식으로 표현하면 아래와 같다.⁶⁸⁾

$$g_i = \sigma(\omega_i[h_{t-1}, x_t])$$

여기서,

g_i 는 입력게이트

σ 는 시그모이드 함수

ω_i 는 가중치

h_{t-1} 는 $t-1$ 시점 출력된 정보

x_t 는 t 시점에 입력된 정보를 말한다.

망각게이트는 이전 시점 은닉층에 저장되어 있던 정보에 가중치를 결정하는 역할을 하며, 시그모이드 함수를 사용한다. 식은 아래와 같이 표현된다.⁶⁹⁾

$$g_f = \sigma(\omega_f[h_{t-1}, x_t])$$

여기서,

g_f 는 망각게이트;

ω_f 는 가중치를 말한다.

출력게이트는 표현 그대로 출력값(h_t)을 결정하는 단계이며, 시그모이드 함수를 이용하며, 식으로 표현하면 아래와 같다.⁷⁰⁾

68) 미디엄(검색일: 2021. 12. 25.)

69) 미디엄(검색일: 2021. 12. 25.)

$$g_o = \sigma(\omega_o[h_{t-1}, x_t])$$

여기서,

g_o 는 출력게이트

ω_o 는 가중치를 말한다.

이외에도 이전 단계 정보를 장기적으로 기억하고 셀의 정보를 갱신하며, 이를 통해 기울기가 소멸하는 문제를 해결한다. 이를 식으로 표현하면 아래와 같다.⁷¹⁾

$$c_t = g_{f,t}c_{t-1} + g_{i,t}\tilde{c}_t$$

$$\tilde{c}_t = \tanh(\omega_c[h_{t-1}, x_t])$$

$$h_t = g_{o,t}\tanh(c_t)$$

여기서,

c_t 는 특정 시점 t 에서 셀스테이트 값

$\tilde{c}_t$ 는 c_t 이전 셀스테이트(cell state)를 말한다.

3. GRU

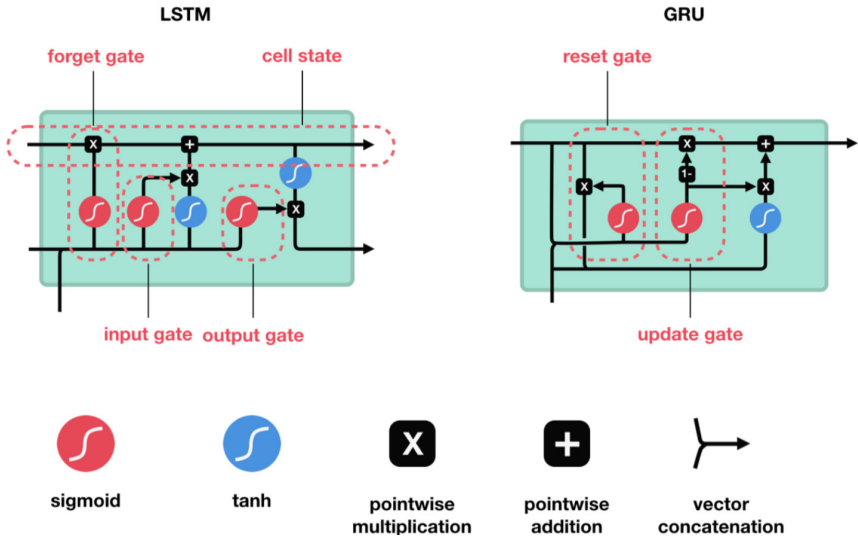
게이트 순환신경망은 순환신경망 계열 인공지능 알고리즘 중에 최근 주목받기 시작한 기존의 RNN과 LSTM의 단점을 보완하여 제안되었다. RNN

70) 미디엄(검색일: 2021. 12. 25.)

71) 미디엄(검색일: 2021. 12. 25.)

이 장기의존성 문제에 대한 해결책으로 제시된 LSTM은 복잡한 구조로 출력, 입력, 망각이라는 세 가지 게이트로 이루어져 있다면, GRU는 업데이트(update) 게이트와 리셋(Reset) 게이트와 같이 구조가 단순해 학습할 파라미터가 줄어들고 학습 속도를 향상하면서 LSTM과 유사한 수준의 성능을 보인다. GRU의 구조는 <그림 3-20>과 같다.

<그림 3-20> LSTM과 GRU 모델 구조 비교



자료: 미디어엄(검색일: 2021. 9.15.)

여기서 리셋 게이트 역할은 이전 시점의 과거 정보와 현재 시점에 입력된 정보를 시그모이드 함수를 이용하여 결합하는데 이때 과거 시점 데이터 값 활용 정도를 결정한다. 업데이트 게이트는 과거와 현재 정보 비율을 구하는 것이 핵심이다.

4. 시사점

모델에 사용하는 딥러닝 알고리즘 종류를 <표 3-8>과 같이 정리하였다.

<표 3-8> 제안 모델이 사용한 딥러닝 알고리즘 종류

항목	개념		특징
	공통점	개념	
DNN	- 대표적인 딥러닝 알고리즘 계열임 - 알고리즘 구조가 입력층과 출력층, 그리고 다수의 은닉층으로 구성됨	다수의 층 구조로 이루어져 있으며, 순환하는 개념은 포함하지 않음	가장 먼저 접하게 되는 인공지능 알고리즘임
LSTM		순환신경망 형태로 이전 시점의 정보 활용	시계열 데이터 학습을 위해 구조 설계
GRU		LSTM과 유사하나, 계산 성능 향상으로 실제 사용에 적합함	

자료: 저자 작성

04

모델별 예측 결과 및 평가

제1절 개요

1. 모델 체계

각 모델은 다층 구조의 DNN, LSTM, GRU 알고리즘을 바탕으로 구축했으며, 향후 운임의 상승과 하락을 예측하는 이진 분류모델로 설계했다. 모델은 크게 두 종류로 구분할 수 있는데, 시점 t 기준으로 한 주 후를 예측하는 주간 운임 예측 모델과 12주 후를 예측하는 단기 운임 예측 모델이다.

모든 모델은 하이퍼파라미터, 학습 크기 및 횟수를 동일 조건으로 설정하여 수행했으며, 입력층과 출력층을 제외하고 각 2개와 3개의 은닉층을 갖는다. 이때 은닉층에서는 렐루(ReLU)를 활성화 함수로 사용했으며, 출력층은 시그모이드(sigmoid)를 이용해 모델 결과가 운임의 '상승'과 '하락' 중 하나의 값을 산출하도록 했다. 이러한 과정을 정리하여 [식 4-1]과 같이 나타내었다.

$$model_{ii}(X_t) \rightarrow Y_{t+i}$$

$$i = 1, 3, 7, 12$$

$$\text{if } i = 1,$$

$$j = d, l, g,$$

$$\text{if } \begin{cases} Y_{t+i} \geq 0.5, & up \\ Y_{t+i} < 0.5, & down \end{cases}$$

여기서,

*model*은 인공지능 알고리즘 *j*에 기반하여 구축한 모델

*i*는 시차

*t*는 시간 단위(주)

*j*는 인공지능 알고리즘 종류

*d*는 *i*가 1일 때 DNN 알고리즘

*l*은 *i*가 1일 때 LSTM 알고리즘

*g*는 *i*가 1일 때 GRU 알고리즘.

Y_{t+i} 는 시차 *i* 이후 운임에 방향을 말한다.

2. 모델 검증 방법

모델을 검증하는 방법은 학습과 데이터 전처리 단계에서 분류한 테스트 데이터를 이용해 예측한 결과를 특정 기준에 따라 평가하여 성능이 좋다고 판단된 모델을 최종 선정하는 것이 일반적이다.

모델 구축에 사용한 데이터는 2011년 5월 10일부터 2021년 6월 30일까지이며, 학습과 테스트 데이터 비율은 각 90-10%, 80-20%, 70-30%로 구성하였으며, 검증 데이터는 학습데이터의 20%로 정해 모델을 구축하였

다. 이번 연구에서는 기존 테스트 데이터 이외에 외부에서 새로운 데이터를 (이하 "테스트2")를 이용해 모델 성능을 추가로 평가하고 그 결과를 함께 제시하였다. 테스트 2는 2021년 2월 19일부터 2021년 6월 30일까지 총 90개(일 단위), 19주 동안 데이터 셋을 활용해 모델을 검증하고 그 결과를 이번 장에 함께 제시하였다.

1) 혼동행렬

분류모델의 예측성능은 <표 4-1>과 같이 모델 예측 결과와 실제 정답을 혼동행렬(Confusion Matrix)을 사용해 정리한 후, 이들 간의 비율을 이용해 평가한다. 분류모델에서 예측값과 실제 정답은 이진수로 표현하며, 긍정(Positive)은 1, 부정(Negative)은 0이다

<표 4-1> 혼동행렬 구성

		실제 정답	
		Positive (1)	Negative (0)
예측 결과	Positive (1)	TP	FP
	Negative (0)	FN	TN

자료: 저자 작성

여기서,

TP(True Positive)는 예측 결과와 실제 정답이 긍정으로 일치

TN(True Negative)은 예측 결과와 실제 정답이 부정으로 일치

FP(False Positive)는 예측 결과는 긍정이나 실제 정답은 부정

FN(False Negative)은 예측 결과는 부정이나 실제 정답은 긍정으로 일치하지 않는 경우를 말한다.

3. 분류모델 성능 평가지표

대표적인 분류모델 성능 평가지표는 정확도(accuracy), 재현율(recall), 정밀도(Precision)와 F1-스코어(F1-Score)가 있다.

1) 정확도

정확도는 모델과 실제값이 일치하는 정도에 비율을 이용하여 수치화한 지표이며, 이를 식으로 표현하면 다음과 같다.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

정확도는 직관적으로 모델 예측치와 실제 데이터 일치 여부를 확인할 수 있다는 장점이 있지만, 실제 데이터의 값이 0 또는 1로 편향된 경우를 확인할 수 없다는 한계가 있다.

2) 재현율

재현율은 모델 평가를 위해 사용한 실제 데이터 값이 부정을 나타낼 때 모델 역시 부정을 예측할 경우를 비율로 표현한 것이다.

$$Recall = \frac{TP}{TP + FN}$$

3) 정밀도

정밀도는 모델 예측값이 긍정일 때와 실제 값이 긍정인 경우를 수치화한 지표이며, 식으로 표현하면 아래와 같다.

$$Precision = \frac{TP}{TP + FP}$$

4) F1-스코어

F1-스코어는 조화평균을 이용해 정확도의 한계를 개선한 지표이다. F1-스코어를 사용하면 모델이 부정 또는 긍정으로 모델이 편향된 값을 산출하고 있는지를 판단할 수 있다.

$$\begin{aligned} F1-Score &= 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \\ &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned}$$

제2절 주간 예측 모델의 결과

1. DNN에 기반한 해상운임예측모델 결과

1) 케이프션

케이프션의 운임 예측 결과는 학습과 테스트 비율에 따라 정확도에 차이가 나타났다. 은닉층 2개에서는 90대 10에 비율일 때 테스트 1의 예측 정확도가 91.02%로 가장 높게 나타났으며, 테스트 2의 예측 정확도는 76.67%였다.

은닉층이 3개인 경우에는 학습과 테스트 비율이 80대 20일 때 가장 높은 예측 정확도를 보였으며 테스트 1의 예측 정확도가 92.84%, 테스트 2의 예측 정확도는 80.00%로 은닉층이 2개일 때보다 은닉층이 3개일 때 운임의 예측 정확도가 높은 것으로 나타났다(〈표 4-2〉 참고).

〈표 4-2〉 DNN기반 모델을 활용한 케이프션 차주 운임의 방향성 예측 결과

단위: %

은닉층 학습-테스트 비율	2개			3개		
	학습	테스트 1	테스트 2	학습	테스트 1	테스트 2
70-30	97.33	88.98	71.11	98.63	89.80	75.56
80-20	99.04	90.80	71.11	98.75	92.84	80.00
90-10	98.58	91.02	76.67	99.19	92.65	65.56

주: 테스트 1은 프로그램상에서 모델을 구축하기 위해 불러온 데이터 중에 모델을 검증하기 위해 학습 과정에서 제외된 데이터를 말하며, 테스트 2는 구축된 모델을 다시 한번 최종 평가하기 위해 외부에서 새로운 데이터를 불러와 테스트 1의 평가 단계를 거쳐 최종 선택된 모델을 활용해 예측검증을 하기 위해 사용한다.
자료: 저자 작성

〈표 4-3〉은 정확도가 높은 것으로 평가되는 모델을 선정하고 그 결과를 분류하여 혼동행렬로 나타난 결과이다. 분류 결과를 이용하여 모델을 다양한 관점에서 평가하였다.

〈표 4-3〉 케이프션 모델의 혼동행렬

		실제		
		상승	하락	
분류	상승	28	9	37
	하락	12	41	53
		40	50	

		실제		
		상승	하락	
분류	상승	27	10	37
	하락	8	45	53
		35	55	

(a) 은닉층 2개, 비율 90:10 테스트 2결과

(b) 은닉층 3개, 비율 80:20의 테스트2 결과

자료: 저자 작성

모델을 다양한 지표로 평가한 결과를 〈표 4-4〉에 정리하였다. 정확도에 서는 (b) 모델이 80.00%로 (a) 모델(76.67%)과 비교하였을 때 성능이 좋은 것으로 나타났으나, 다른 성능지표에서는 (a) 모델의 결과가 높은 것으로 나타나 (b) 모델보다 안정적인 결과를 산출하였다.

〈표 4-4〉 성능지표 비교

성능지표	(a)	(b)
정확도(accuracy)	76.67	80.00
재현율(recall)	75.68	72.97
정밀도(Precision)	40.58	37.50
F1-스코어(F1-Score)	52.83	49.54

단위: %

자료: 저자 작성

2) 파나마스선

파나마스선 운임 예측 결과는 학습과 테스트 비율에 따라 정확도에 차이가 나타났다. 은닉층 2개에서는 70대 30 비율일 때 테스트 1에 예측 정확도가 89.50%로 가장 높았으며, 테스트 2의 예측 정확도는 77.78%였다.

은닉층이 3개인 경우에는 학습과 테스트 비율이 90대 10일 때 가장 높

은 예측 정확도를 보였으며 테스트1의 예측 정확도가 90.20%, 테스트2의 예측 정확도는 86.67%로 은닉층이 2개일 때보다 은닉층이 3개일 때 은임의 예측 정확도가 높았다(〈표 4-5〉 참고).

〈표 4-5〉 DNN기반 모델을 활용한 파나막스선 차주 운임의 방향성 예측 결과

단위: %

은닉층	2개			3개		
학습-테스트 비율	학습	테스트1	테스트2	학습	테스트1	테스트2
70-30	97.85	89.50	77.78	98.24	88.95	81.11
80-20	97.67	89.16	76.67	96.98	87.73	84.44
90-10	97.82	91.02	73.33	98.53	90.20	86.67

주: 테스트 1은 프로그램상에서 모델을 구축하기 위해 불러온 데이터 중에 모델을 검증하기 위해 학습 과정에서 제외한 데이터를 말하며, 테스트 2는 구축된 모델을 다시 한번 최종 평가하기 위해 외부에서 새로운 데이터를 불러와 테스트1의 평가 단계를 거쳐 최종 선택된 모델을 활용해 예측검증을 하기 위해 사용한다.

자료: 저자 작성

앞서 정확도를 통해 예측성능이 좋은 모델을 선정하고 각 예측값의 결과를 〈표 4-6〉과 같이 혼동행렬로 정리하였다. 두 모델의 분류 결과를 사용하여 다양한 지표로 변환하여 비교하고자 하였다.

〈표 4-6〉 파나막스선 모델의 혼동행렬

		test 2		
		상승	하락	
분류	상승	13	20	33
	하락	0	57	57
		13	77	

		test 2		
		상승	하락	
분류	상승	24	9	33
	하락	3	54	57
		27	63	

(a) 은닉층 2개, 비율 70:30 테스트 2결과

(b) 은닉층 3개, 비율 90:10의 테스트2 결과

자료: 저자 작성

비교한 결과는 <표 4-7>에서 제시하였다. 정확도를 비롯한 전체적인 성능지표를 통해 (b) 모델이 (a) 모델보다 좋은 성능을 보이는 모델인 것으로 나타났다.

<표 4-7> 성능지표 비교

단위: %

성능지표	(a)	(b)
정확도(accuracy)	77.78	86.67
재현율(recall)	39.39	72.73
정밀도(Precision)	18.57	30.77
F1-스코어(F1-Score)	25.24	43.24

자료: 저자 작성

2. LSTM에 기반한 해상운임예측모델 결과

1) 케이프선

케이프선 운임 예측 결과는 학습과 테스트 비율에 따라 정확도에 차이가 나타났으며, 은닉층 2개에서는 90대 10의 비율일 때 테스트 1의 예측 정확도가 72.65%로 가장 높게 나타났으며, 테스트 2의 예측 정확도는 64.20%였다.

은닉층이 3개인 경우에는 학습과 테스트 비율이 90대 10일 때 가장 높은 예측 정확도를 보였으며 테스트1의 예측 정확도가 69.39%, 테스트2의 예측 정확도는 70.00%로 은닉층이 2개일 때보다 은닉층이 3개일 때 운임의 예측 정확도가 높은 것으로 나타났다.

〈표 4-8〉 LSTM기반 모델을 활용한 케이프션 차주 운임의 방향성 예측 결과

단위: %

은닉층	2개			3개		
학습-테스트 비율	학습	테스트1	테스트2	학습	테스트1	테스트2
70-30	77.76	64.85	51.11	79.7	66.76	64.44
80-20	77.89	67.48	55.56	79.63	70.96	62.22
90-10	77.98	72.65	64.20	77.98	69.39	70.00

주: 테스트 1은 프로그램상에서 모델을 구축하기 위해 불러온 데이터 중에 모델을 검증하기 위해 학습 과정에
서 제외한 데이터를 말하며, 테스트 2는 구축된 모델을 다시 한번 최종 평가하기 위해 외부에서 새로운 데
이터를 불러와 테스트1의 평가 단계를 거쳐 최종 선택된 모델을 활용해 예측검증을 하기 위해 사용한다.
자료: 저자 작성

〈표 4-9〉는 정확도가 높은 것으로 평가되는 모델을 선정하고 그 결과를
분류하여 혼동행렬로 나타낸 결과이다.

〈표 4-9〉 케이프션 모델의 혼동행렬

		test 2		
		상승	하락	
분류	상승	18	26	44
	하락	12	50	62
		30	76	

		test 2		
		상승	하락	
분류	상승	31	6	37
	하락	21	32	53
		52	38	

(a) 은닉층 2개, 비율 90:10 테스트 2결과

(b) 은닉층 3개, 비율 90:10의 테스트2 결과

자료: 저자 작성

〈표 4-10〉을 통해 선정된 두 모델의 성능을 정확도 이외에 재현율, 정밀
도, F1-스코어를 이용해 비교한 결과를 제시하였다.

〈표 4-10〉 성능지표 비교

단위: %

성능지표	(a)	(b)
정확도(accuracy)	64.2	70.00
재현율(recall)	40.9	83.78
정밀도(Precision)	26.5	49.21
F1-스코어(F1-Score)	32.1	62.00

자료: 저자 작성

눈여겨볼 부분은 은닉층이 2개 일 때 학습데이터를 통한 예측 정확도가 동일하게 나온 모델일지라도 테스트의 예측 정확도가 다르게 나왔는데 이는 학습 과정에서 과적합이 발생했기 때문이다(〈표 4-6〉 참고).

2) 파나막스선

파나막스선 운임 예측 결과는 학습과 테스트 비율에 따라 정확도에 차이가 나타났으며, 은닉층 2개에서는 90대 10의 비율일 때 테스트 1에 예측 정확도가 74.29%로 가장 높게 나타났으며, 테스트 2의 예측 정확도는 70.00%였다. 〈표 4-11〉은 정확도가 높은 것으로 평가되는 모델을 선정하고 그 결과를 분류하여 혼동행렬로 나타낸 결과이다.

〈표 4-11〉 LSTM기반 모델을 활용한 파나막스선 차주 운임의 방향성 예측 결과

단위: %

은닉층 학습-테스트 비율	2개			3개		
	학습	테스트1	테스트2	학습	테스트1	테스트2
70-30	88.06	62.26	61.11	97.33	88.98	71.11
80-20	82.70	65.44	66.67	99.04	90.8	71.11
90-10	87.22	74.29	70.00	98.58	91.02	76.67

주: 테스트 1은 프로그램상에서 모델을 구축하기 위해 불러온 데이터 중에 모델을 검증하기 위해 학습 과정에서 제외한 데이터를 말하며, 테스트 2는 구축된 모델을 다시 한번 최종 평가하기 위해 외부에서 새로운 데이터를 불러와 테스트1의 평가 단계를 거쳐 최종 선택된 모델을 활용해 예측검증을 하기 위해 사용한다.

자료: 저자 작성

은닉층이 3개인 경우에는 학습과 테스트 비율이 90대 10일 때가 가장 높은 예측 정확도를 보였다. 테스트 1의 예측 정확도가 91.02%, 테스트 2의 예측 정확도는 76.67%로 은닉층이 3개일 때 운임의 예측 정확도가 높았다(〈표 4-11〉 참고).

〈표 4-12〉는 정확도가 높은 것으로 평가되는 모델을 선정하고 그 결과를 분류하여 혼동행렬로 나타낸 결과이다.

〈표 4-12〉 파나막스선 모델의 혼동행렬

		test 2		
		상승	하락	
분류	상승	15	22	37
	하락	12	41	53
		27	63	

		test 2		
		상승	하락	
분류	상승	31	6	37
	하락	21	32	53
		52	38	

(a) 은닉층 2개, 비율 90:10 테스트 2결과 (b) 은닉층 3개, 비율 90:10의 테스트2 결과

자료: 저자 작성

〈표 4-13〉 성능지표 비교

성능지표	(a)	(b)
정확도(accuracy)	62.2	70.00
재현율(recall)	40.5	83.78
정밀도(Precision)	26.8	49.21
F1-스코어(F1-Score)	32.3	62.00

자료: 저자 작성

3. GRU에 기반한 해상운임예측모델 결과

1) 케이프선

케이프선 운임 예측 결과는 학습과 테스트 비율에 따라 정확도에 차이가 나타났으며, 은닉층 2개에서는 70대 30의 비율일 때 테스트 1에 예측 정확도가 65.4%로 가장 높게 나타났으며, 테스트 2의 예측 정확도는 70.00%였다.

〈표 4-14〉 GRU기반 모델을 활용한 케이프선 차주 운임의 방향성 예측 결과

단위: %

은닉층	2개			3개		
학습-테스트 비율	학습	테스트1	테스트2	학습	테스트1	테스트2
70-30	89.53	65.4	70.00	91.05	67.44	48.89
80-20	86.39	64.62	64.44	84.54	62.37	57.78
90-10	87.76	76.33	61.11	97.72	61.63	62.22

주: 테스트 1은 프로그램상에서 모델을 구축하기 위해 불러온 데이터 중에 모델을 검증하기 위해 학습 과정에서 제외된 데이터를 말하며, 테스트 2는 구축된 모델을 다시 한번 최종 평가하기 위해 외부에서 새로운 데이터를 불러와 테스트1의 평가 단계를 거쳐 최종 선택된 모델을 활용해 예측검증을 하기 위해 사용한다.

자료: 저자 작성

은닉층이 3개인 경우, 학습과 테스트 비율이 90대 10일 때 가장 높은 예측 정확도를 보였으며 테스트 1 예측 정확도가 61.63%, 테스트 2 예측 정확도는 62.22%로 은닉층이 3개일 때보다 은닉층이 2개일 때 운임의 예측 정확도가 높은 것으로 나타났다(〈표 4-10〉 참고).

〈표 4-15〉는 정확도가 높은 것으로 평가되는 모델을 선정하고 그 결과를 분류하여 혼동행렬로 나타낸 결과이다.

〈표 4-15〉 케이프션 모델의 혼동행렬

		tes 2		
		상승	하락	
분류	상승	21	16	37
	하락	11	42	53
		32	58	

		test 2		
		상승	하락	
분류	상승	16	21	37
	하락	13	40	53
		29	61	

(a) 은닉층 2개, 비율 70:30 테스트 2결과 (b) 은닉층 3개, 비율 90:10의 테스트2 결과

자료: 저자 작성

〈표 4-16〉은 정확도 이외에도 모든 지표가 (a) 모델이 더 높은 성능을 보인다고 나타났다.

〈표 4-16〉 성능지표 비교

성능지표	단위: %	
	(a)	(b)
정확도(accuracy)	70.00	62.22
재현율(recall)	56.76	43.24
정밀도(Precision)	33.33	28.57
F1-스코어(F1-Score)	42.00	34.41

자료: 저자 작성

2) 파나막스선

파나막스선 운임 예측 결과는 학습과 테스트 비율에 따라 정확도에 차이가 나타났으며, 은닉층 2개에서 90대 10 비율일 때 테스트 1의 예측 정확도가 78.37%로 가장 높게 나타났다.

〈표 4-17〉 GRU기반 모델을 활용한 파나막스선 차주 운임의 방향성 예측 결과

단위: %

은닉층	2개			3개		
학습-테스트 비율	학습	테스트1	테스트2	학습	테스트1	테스트2
70-30	93.45	62.26	48.89	91.05	67.44	48.89
80-20	63.00	54.60	42.22	84.54	62.37	57.78
90-10	96.22	78.37	72.22	92.72	61.63	62.22

주: 테스트 1은 프로그램상에서 모델을 구축하기 위해 불러온 데이터 중에 모델을 검증하기 위해 학습 과정에서 제외한 데이터를 말하며, 테스트 2는 구축된 모델을 다시 한번 최종 평가하기 위해 외부에서 새로운 데이터를 불러와 테스트1의 평가 단계를 거쳐 최종 선택된 모델을 활용해 예측검증을 하기 위해 사용한다.

자료: 저자 작성

테스트 2의 예측 정확도는 72.22%였으며, 은닉층이 3개인 경우에는 학습과 테스트 비율이 90대 10일 때 가장 높은 예측 정확도를 보였다.

테스트1의 예측 정확도가 61.63%, 테스트 2의 예측 정확도는 62.22%로 은닉층이 2개일 때보다 은닉층이 3개일 때 운임 예측 정확도가 높았다(〈표 4-8〉 참고).

〈표 4-18〉은 정확도가 높은 것으로 평가되는 모델을 선정하고 그 결과를 분류하여 혼동행렬로 나타낸 결과이다.

〈표 4-18〉 파나막스선 모델의 혼동행렬

		tes t2		
		상승	하락	
분류	상승	8	25	33
	하락	0	57	57
		8	82	

(a) 은닉층 2개, 비율 90:10 테스트 2결과

		test 2		
		상승	하락	
분류	상승	16	21	37
	하락	13	40	53
		29	61	

(b) 은닉층 3개, 비율 80:20의 테스트2 결과

자료: 저자 작성

〈표 4-19〉는 모델 간 지표를 비교한 결과를 보여주고 있다. (a) 모델의 정확도가 (b) 모델보다 높은 것으로 나온 반면 이외의 모든 지표가 상대적으로 모델의 결과값이 안정적이지 못한 것을 보여주고 있다.

〈표 4-19〉 성능지표 비교

단위: %

성능지표	(a)	(b)
정확도(accuracy)	72.22	62.22
재현율(recall)	24.24	43.24
정밀도(Precision)	12.31	28.57
F1-스코어(F1-Score)	16.33	34.41

자료: 저자 작성

4. 시사점

모델 구축을 통해 DNN 모델이 다른 인공지능 알고리즘보다 현재의 입력변수를 잘 학습하고 있음을 확인하였다. 이러한 결과는 DNN 모델이 파라미터 갱신 과정을 통해 변수를 선택하는 방식을 채택해 다수의 변수를 학습하는데 적합한 알고리즘이기 때문으로 판단된다. 즉, 모델이 주어진 데이터에 반복적 학습을 통해 변수별 가중치의 크기를 조정하는 과정에서 변수의 중요도가 바뀌게 되어 입력값이 다변량일 때 사용하기에 적합한 모델인 것을 확인하였다.

제3절 단기 예측 모델의 결과

DNN 알고리즘에 기반하여 운임 예측 기간을 1주에서 12주까지 확장해 보았다. 모델 학습은 테스트 비율과 은닉층의 개수에 따라 구분되며, 그 결과를 이번 절에서 제시하고자 한다. <표 4-20>에서 제시한 바와 같이 모델을 명칭 하도록 한다.

<표 4-20> 모델의 명칭

단위: %

학습-테스트 비율	은닉층 개수	
	2층	3층
70-30	nm_ij	nm_ij
80-20	nm_ij	nm_ij

자료: 저자 작성

여기서, n 은 선형을 의미하며 n 이 c 인 경우는 케이프선, n 이 p 인 경우를 파나막스선으로 정의한다. m 은 은닉층 개수를 나타내며, i 와 j 는 각각 학습과 테스트 비율을 말한다.

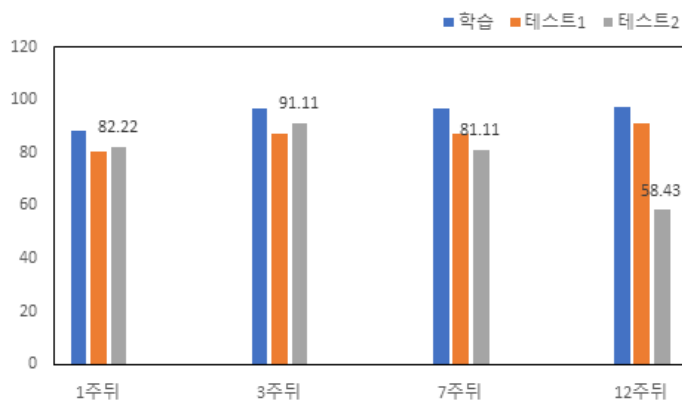
1. 케이프선

1) cn_7030 모델

<그림 4-1>은 모델 예측 정확도를 나타낸다. 7주 뒤에 81.11%의 예측 정확도를 보여주고 있으나, 12주 후에는 58.43%를 성능이 떨어지는 것을 확인할 수 있다.

〈그림 4-1〉 케이프선 운임 방향성 예측 결과(은닉층 2개, 학습 비율 70:30)

단위: %

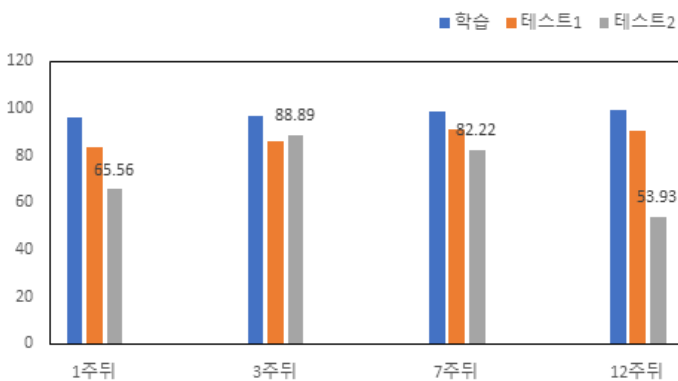


자료: 저자 작성

〈그림 4-2〉는 모델이 같은 학습과 테스트 비율로 학습을 하였으나, 은닉층의 개수 변화에 따라 정확도 역시 함께 변화한다는 사실을 확인할 수 있었다. 또한 은닉층의 수가 증가한다고 해서 예측성능이 높아지는 것은 아니라는 것을 보여준다.

〈그림 4-2〉 케이프선 운임 방향성 예측 결과(은닉층 3개, 학습 비율 70:30)

단위: %



자료: 저자 작성

〈표 4-21〉은 모델 성능을 정확도, 재현율, 정밀도와 F-스코어로 평가한 결과를 보여준다. 이를 통해 차주 예측에는 은닉층 2개 모델이 상대적으로 좋은 성능과 안정적인 결과를 도출하는 것을 알 수 있다.

〈표 4-21〉 성능지표 비교

단위: %

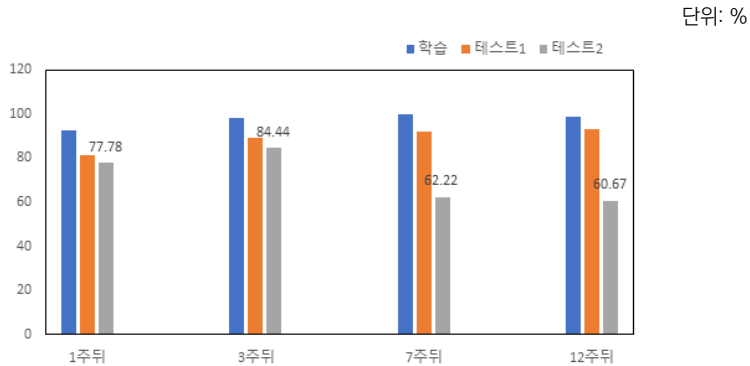
학습-테스트 비율 (70-30)	은닉층 개수							
	2층				3층			
	정확도	재현율	정밀도	F1	정확도	재현율	정밀도	F1
1주 후	82.22	80.56	39.19	52.73	65.56	47.22	28.81	35.79
3주 후	91.11	94.12	39.02	55.17	88.89	85.29	36.25	50.88
7주 후	81.11	100.00	43.84	60.95	82.22	96.88	41.89	58.49
12주 후	58.43	76.19	30.77	43.84	53.93	100.00	43.75	60.87

주: F1은 F1-스코어이며, 테스트 2의 결과를 사용하여 평가함
자료: 저자 작성

2) cn_8020 모델

〈그림 4-3〉은 모델 학습 비중이 늘어났음에도 불구하고 예측 기간이 길어질 경우 정확도가 낮아지는 것을 확인하였다.

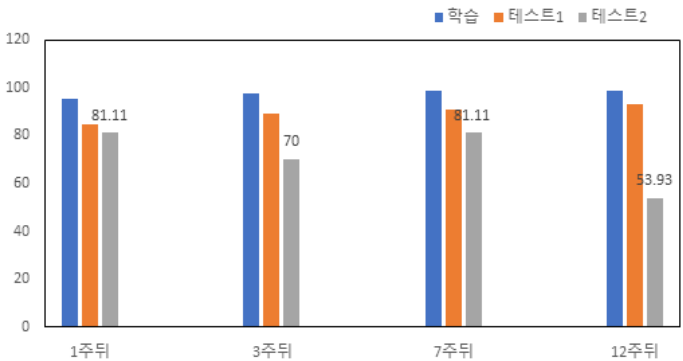
〈그림 4-3〉 케이프션 운임 방향성 예측 결과(은닉층 2개, 학습 비율 80:20)



자료: 저자 작성

〈그림 4-4〉는 3주 뒤 운임 예측력이 7주 뒤보다 낮은 정확도를 보였으며, 12주에서는 역시 53.93%로 낮은 예측성능을 보이고 있음을 확인하였다.

〈그림 4-4〉 케이프선 운임 방향성 예측 결과(은닉층 3개, 학습 비율 80:20)
단위: %



자료: 저자 작성

두 모델을 성능지표를 통해 평가한 결과를 〈표 4-22〉와 같이 정리하였다. 은닉층 개수가 2개인 모델에 비해 3층 모델이 전체적으로 F1-스코어가 높아 안정적인 결과를 도출하는 것을 확인하였다.

〈표 4-22〉 성능지표 비교

단위: %

학습-테스트 비율 (80-20)	은닉층 개수							
	2층				3층			
	정확도	재현율	정밀도	F1	정확도	재현율	정밀도	F1
1주 후	77.78	61.11	31.43	41.51	81.11	91.67	45.21	60.55
3주 후	84.44	91.18	40.79	56.36	70.00	29.41	15.87	20.62
7주 후	62.22	34.38	19.64	25.00	81.11	100.00	43.84	60.95
12주 후	60.67	100.00	38.89	56.00	53.93	95.24	41.67	57.97

주: F1은 F1-스코어이며, 테스트 2의 결과를 사용하여 평가함
자료: 저자 작성

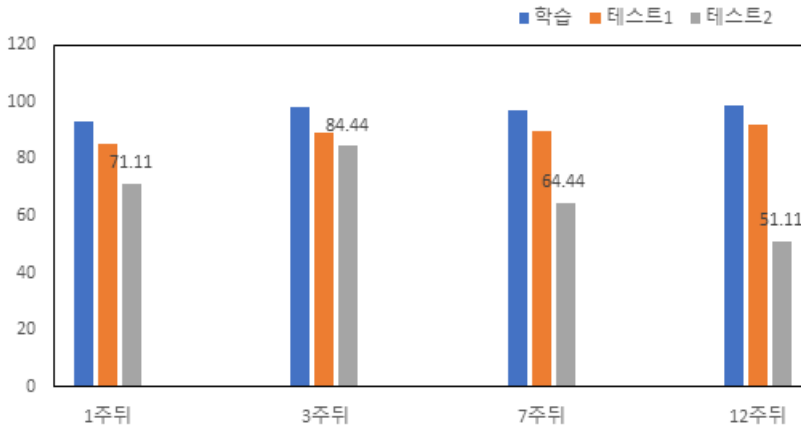
2. 파나막스선

1) pn_7030 모델

〈그림 4-5〉를 통해 은닉층 2개 모델이 12주 뒤 운임 예측력이 상당히 높은 반면 3개의 은닉층 정확도는 대체로 낮은 수준인 것으로 나타났다.

〈그림 4-5〉 파나막스선 운임 방향성 예측 결과(은닉층 2개, 학습 비율 70:30)

단위: %

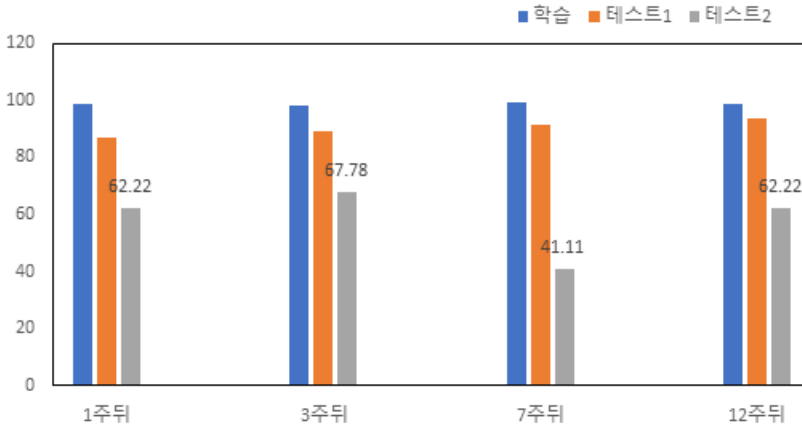


자료: 저자 작성

〈그림 4-6〉은 은닉층이 2개인 모델에 비해 전반적으로 낮은 운임 예측 성능을 보여준다.

〈그림 4-6〉 파나막스선 운임 방향성 예측 결과(은닉층 3개, 학습 비율 70:30)

단위: %



자료: 저자 작성

〈표 4-23〉은 모델 성능지표를 보여주고 있다. 대부분 모델이 정확도로 놓고 평가했을 때 좋은 성능을 보이나, 다른 지표들과 종합적으로 모델을 평가했을 때는, 불안정한 모델인 것으로 확인하였다.

〈표 4-23〉 성능지표 비교

단위: %

학습-테스트 비율 (70-30)	은닉층 개수							
	2층				3층			
	정확도	재현율	정밀도	F1	정확도	재현율	정밀도	F1
1주 후	71.11	25.00	12.50	16.67	62.22	6.25	3.57	4.55
3주 후	84.44	91.18	40.79	56.36	67.78	68.00	27.87	39.53
7주 후	64.44	92.86	22.41	36.11	41.11	92.86	35.14	50.98
12주 후	51.11	87.50	15.22	25.93	62.22	100.00	14.29	25.00

주: F1은 F1-스코어이며, 테스트2의 결과를 사용하여 평가함

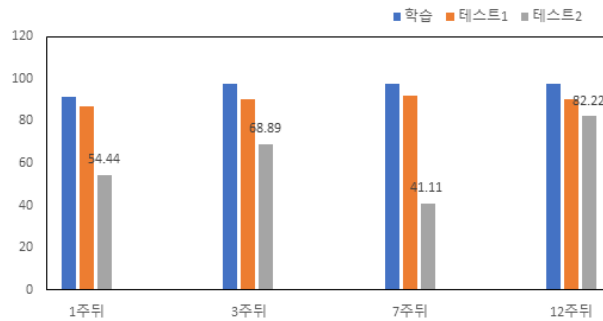
자료: 저자 작성

2) pn_8020 모델

〈그림 4-7〉에서 보는 바와 같이 p2_8020 모델은 12주 후, 운임 방향성을 예측하는데 성능이 뛰어난 것으로 나타났다.

〈그림 4-7〉 파나막스선 운임 방향성 예측 결과(은닉층 2개, 학습 비율 80:20)

단위: %

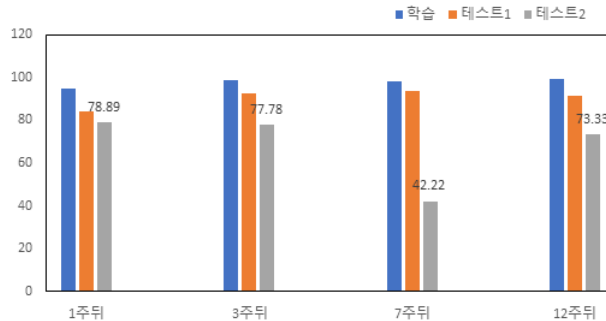


자료: 저자 작성

〈그림 4-8〉도 유사한 결과를 보여주고 있다. 7주 뒤의 정확도가 12주보다 낮게 나타났다.

〈그림 4-8〉 파나막스선 운임 방향성 예측 결과(은닉층 3개, 학습 비율 80:20)

단위: %



자료: 저자 작성

〈표 4-24〉는 각 모델의 성능지표를 보여주고 있다. 모델 정확도는 은닉층 개수가 3층인 경우가 더 높은 것으로 나타났지만, 다른 지표상으로 확인했을 때 모델이 안정적인 결과를 도출하지는 못하는 것으로 확인하였다.

〈표 4-24〉 성능지표 비교

단위: %

학습-테스트 비율 (80-20)	은닉층 개수							
	2층				3층			
	정확도	재현율	정밀도	F1	정확도	재현율	정밀도	F1
1주 후	54.44	31.25	20.41	24.69	78.89	68.75	30.99	42.72
3주 후	68.89	56.00	22.58	32.18	77.78	44.00	15.71	23.16
7주 후	41.11	92.86	35.14	50.98	42.22	42.86	15.79	23.08
12주 후	82.22	75.00	8.11	14.63	73.33	87.50	10.61	18.92

주: F1은 F1-스코어이며, 테스트 2의 결과를 사용하여 평가함
자료: 저자 작성

3. 시사점

이번 연구를 통해 전처리 과정에 변동성 분석을 추가한 결과 운임의 예측성능을 향상하였고 12주 후 운임의 방향성 예측이 가능함을 확인하였다.

05

결론 및 정책제언

제1절 결론

본 연구는 다양한 인공지능 알고리즘 기반 모델을 확보하고, 예측 기간 확장 가능성을 검토하는 데 집중하여 진행하였다. 변동성이 큰 건화물선 시장에 특화된 해상운임 예측 모델을 구축한다는 일은 쉽지 않은 일이며, 본 연구에서는 현재 보유하고 있는 모델의 구축과정을 개선하고 새로운 알고리즘을 활용해 모델을 추가하여 복수 모델을 확보하였다. 또한 예측 기간을 최대 12주까지 확장하여 향후 분기 전망에 활용 가능성을 제시하였다.

1. 하이브리드 단기 예측 모델 구축

본 연구는 기존 모델의 한계와 문제점을 개선하고 활용성을 높이려는 방안을 모색하였다. 해운시장 데이터에 내포된 가치와 가능성에 대해 주목하고 이를 활용할 수 있도록 데이터 전처리 과정에 변동성 분석을 추가하였다. 또한 운임, 용선료, FFA 가격을 변수로 활용하여 해운 시황과 미래 시장에 대한 시장참여자의 기대 심리 변화를 모델에 반영하였다. 즉, 데이터

변수 전처리에 새로운 방법론 도입한 새로운 하이브리드형 모델을 제안하였다.

해운빅데이터연구센터의 기존 DNN을 활용한 모델과 예측 성능을 비교해보면 이번 연구에서 제안한 모델의 예측력이 더 높은 것을 확인할 수 있었다. 기존 케이프션 운임 예측 모델의 성능(70.9%)⁷²⁾에 비해 제안한 모델의 예측성능이 10%p 향상된 결과를 제시하였으며, 파나마스선 모델은 기존 (72.3%)⁷³⁾보다 86.67%로 향상된 결과를 보여주었다.

2. 다양한 데이터 조합 및 알고리즘 기반 복수 모델의 확보

주간 단위 예측에 GRU와 같은 새로운 인공지능 알고리즘 모델을 추가함으로써 다수의 모델을 구축하였고, 각 알고리즘 특성을 활용하여 시점별 운임의 예측력 정도에 따라 모델 성능을 향상할 수 있는 분석 기반을 마련하였다.

다수의 모델 확보는 단일 모델에 대한 의존도를 낮추고 모델 상호평가를 통해 예측 정확도를 향상할 수 있다는 점에 의의가 있다. 다수 모델의 예측력을 평가해 본 결과 DNN 모델이 기존 모델보다 예측력이 향상된 결과를 보였지만 LSTM과 GRU와 같은 시계열에 특화된 알고리즘 기반 모델이 월등히 나은 결과를 보이지는 않았다. 다만, 동일 알고리즘에서도 학습의 비율이 다른 데이터 간 결과치를 비교하고 그 결과를 선택하여 제시함으로써 향상된 예측성능을 제시할 수 있다는 가능성을 확인하였다.

72) 윤희성 외(2019), pp. 1~81.

73) 윤희성 외(2019), pp. 1~81.

제2절 시사점

1. 다양한 방법론을 도입한 예측 모델 필요

모델을 활용한 예측연구는 단순히 데이터를 방법론에 적용하고 그 결과가 나온다고 해서 해결되는 문제가 아니다. 실제 모델을 예측에 적용하여 활용하면 실험과 달리 예측성능이 크게 떨어지는 경우가 비일비재하다. 따라서 전통적인 방법론이나 최신의 방법론을 기반으로 한 다수의 모델을 확보하여 구축함으로써 예측 성능을 향상하고 활용도를 높이는 부분에 집중하는 등 모델 개발과 확보를 위한 지속적인 노력을 기울여야 한다.

2. 전망 시점에 따른 방법론 적용의 차별화 필요

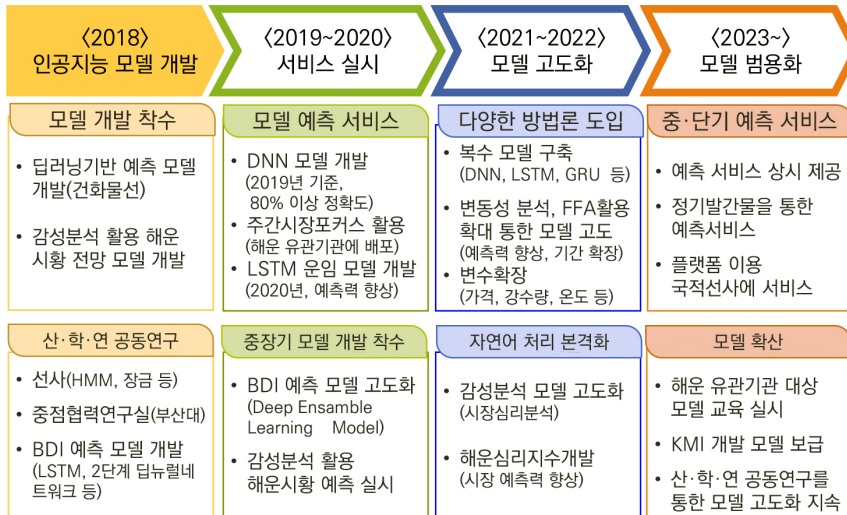
인공지능을 활용한 방법론 중 AIS, 텍스트 마이닝을 활용한 연구는 자료의 생성 주기가 빠르므로 계량적인 분석에 비해 현실 반영도를 높일 수 있어 예측 시점별 방법론 적용을 달리할 수 있다. 단기적으로는 AIS, 기사 등 빅데이터 기반 연구를 통해 시황을 전망하고 중기적으로는 인공지능 방법론을 적용해 변동성이 큰 시장을 예측하며 장기적으로는 전통적인 계량분석을 통해 추세를 분석하는 방법론을 적용할 수 있다.

3. 産-學-研 공동연구 필요

현재 본원에서 제공하고 있는 주 단위 건화물선 운임 예측은 2018년부터 진행된 연구 산물이다. 하지만 2018년부터 진행된 산-학-연 공동연구

는 2020년 이후 중단된 상황이나 선사의 수요(needs) 파악과 함께 공동연구의 필요성은 증가하고 있다. 선사는 환경 규제 등의 해운 시황에 영향을 미치는 변화에 대응하기 위해 자체적으로 자료를 축적하거나 관련 연구를 진행하고 있지만 그렇지 못한 선사가 대다수이다. 특히 중소형 선사 중심으로 시황뿐만 아니라 항로 최적화, 환경규제 대응을 위한 공동 협의체가 필요하다. 이를 위해 관련 예산을 공동으로 편당하거나 정부 지원을 받는 방식을 모색해야 한다.

〈그림 5-1〉 해운시황 예측모델 개발 및 서비스



자료: 저자 작성

4. 뉴노멀에 따른 새로운 변수 필요

최근 우리나라를 포함한 세계 모든 경제는 2008년과 같이 금융자산 위기가 실물자산으로 이전되는 상황이 아닌 전염병(코로나19)으로 인한 불확실성을 겪고 있다. 기존에 연구되어 해운에 영향을 미치는 거시경제변수와 의 관계가 팬데믹 이전과 이후로 크게 바뀔 수 있다. 예로 과거 건화물선 시황에 항만 정체가 영향은 미쳤으나 지금처럼 크게 영향을 미친 예는 없었다. 중국을 중심으로 검·방역 서비스의 강화로 2021년 10월 기준 총 건화물 선대 5% 이상인 600척 이상 선박이 중국항만에 대기하고 있다. 이는 자연스럽게 선박 공급 회전율을 하락하게 되었으며 2008년 이래 최대 운임실적을 기록하였다. 이러한 추세가 지속되면 항만 혼잡은 매우 일상적인 형태로 자리 잡을 수 있으며 시장참여자에게는 뉴노멀로 인식될 수 있다. 이에 항만 혼잡을 지수화하여 새로운 변수로 활용하는 연구도 진행되어야 한다. 인공지능 모델은 학습을 기반으로 예측되어 새로운 시대의 도래에 따른 영향을 추정할 수 있는 새로운 변수선택이 필요하다.

참고문헌

국내 문헌

- 김경환(2021), 「순환신경망 모델을 활용한 선박연료유 가격예측에 관한 연구」, 한국해양대학교 대학원 경영학박사 학위논문. pp. 1~120.
- 김도희·김혜미·심성현·최유림·배혜림·윤희성(2019), 「딥러닝을 활용한 건화물선 운임 지수 예측」, 『대한산업공학회지』, 45(2), pp. 111~116.
- 김형준·류두진·조훈(2019), 「시계열 요소분해법을 활용한 발틱건화물운임지수(BDI) 예측」, 『경영학연구』, 48(3), pp. 715~731.
- 마틴 스토포드, 「해운경제학」, 양창호 외 역(2015), 박영사, pp. 1~955.
- 모수원(2010), 「2010 년 BDI 의 예측: ARIMA 모델과 HP 기법을 이용하여」, 『한국항만경제학회지』, 26(1), pp. 222~233.
- 모수원(2013), 「해운경기의 예측: 2013 년」, 『한국항만경제학회지』, 29(1), pp. 67~76.
- 민인식·최필선(2009), 「STATA 기초통계와 회귀분석」, 한국STATA학회, pp. 1~291.
- 손상기(1995), 「국제운임 시장에서의 ARIMA 모델을 이용한 볼틱(Baltic)운임지수의 시계열 구조분석 및 예측에 관한 실증 연구」, 해운물류연구, 21, pp. 197~212.
- 야마시타 타카요시, 「그림과 수식으로 배우는 통통 딥러닝」, 심효섭 역(2017), 제이펍, pp. 1~224.
- 유동희·황소영(2008), 「AIS(자동식별시스템) 기술 개요」, 『한국정보통신학회지』, 12(2), pp. 38~45.
- 윤희성·최건우·황수진·박동원(2019), 「딥러닝 기반의 건화물선 시황예측 연구」, 한

-
- 국해양수산물개발원, pp. 1~81.
- 이석용(2020), 「컨테이너선 해운시장의 추세와 순환주기 분석을 통한 시황 예측에 관한 연구」, *해운물류연구*, 36(4), pp. 597~618.
- 이성윤·안기명(2018), 「VECM에 의한 BDI 예측과 영향요인에 관한 실증연구」, 『한국항해항만학회지』, 42(6), pp. 546~554.
- 이승윤·이은경·한진현(2017), 「기계학습과 혼합주기 모델을 활용한 단기전망시스템」, 『BOK 조사통계월보』, 71(5), pp.16~41.
- 이우식(2017), 「딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측」, 『한국데이터정보과학회지』, 28(2), pp. 287~295.
- 이재길(2017), 「R프로그램에 기반한 시계열 자료 분석」, 황소걸음 아카데미, pp. 1~472.
- 이현동·하현정·허수진(2019), 「수산부문 전망모델 [KMI-FOSiM] 구축 연구 (2차년도)」, 한국해양수산개발원, pp. 1~255.
- 정상국·김성기(2011), 「국제유가의 변화가 건화물선 운임에 미치는 영향과 건화물선 운임간의 상관관계에 관한 연구」, 『한국항만경제학회지』, 27(2), pp. 217~240.
- 조담(2011), 「코스피 지수 변동폭의 실증적 분석」, 『산업경제연구』, 24(3), pp. 1571~1593.
- 한민수·유성진(2017), 「Recurrent Neural Networks를 활용한 Baltic Dry Index(BDI) 예측」, 『한국항해항만학회 학술대회논문집』, pp. 50~53.
- 한민수·유성진(2019), 「Long Short-Term Memory를 활용한 건화물운임지수 예측」, 『품질경영학회지』, 47(3), pp. 497~508.
- 황수진·고병욱·최건우·안영균·김병주(2020), 「인공지능기반 해상운임예측 연구」, 한국해양수산개발원, pp. 1~65.

국외 문헌

- Allocation Models using Marine Logistics Data and its Application to Bulk Carrier Demand Forecasting and Basic Planning Support, *Journal of the Japan Society of Naval Architects and Ocean Engineers*, 27, pp. 139-148.

- Ascarya, A. (2017), "The Real Determinants of Financial Crisis and How to Resolve it in Islamic Economics Perspective", *International Journal of Economic Research*, 14(3), pp. 501-531.
- Bao, J., L. Pan, & Y. Xie(2016), "A new BDI forecasting model based on support vector machine", *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*, pp. 65-69.
- Beenstock, M. & A. Vergottis(1989), "An econometric model of the world tanker market", *Journal of Transport Economics and Policy*, 23(3), pp. 263-280.
- Brynjolfsson, E., L. M. Hitt & H. Kim(2011), "Strength in numbers: How does data-driven decision-making affect firm performance?", *International Conference on Information Systems 2011, ICIS 2011*, 1, pp. 263-280.
- Chen, S., H. Meersman & E. Van de Voorde(2012), "Forecasting spot rates at main routes in the dry bulk market", *Maritime Economics & Logistics*, 14(4), pp. 498-537.
- Dai, L., H. Hu, & D. Zhang(2015), "An empirical analysis of freight rate and vessel price volatility transmission in global dry bulk shipping market", *Journal of Traffic and Transportation Engineering (English Edition)*, 2(5), pp. 353-361.
- Eslami, P., et al(2017), "Predicting tanker freight rates using parsimonious variables and a hybrid artificial neural network with an adaptive genetic algorithm", *Maritime Economics & Logistics*, 19(3), pp. 538-550.
- Foster, P. & F. Tom(2013), "Data science and its relationship to big data and data-driven decision making", *Big Data*, 1(1), pp. 51-59.
- Guan, F., et al(2016), "Multi-step hybrid prediction model of baltic supermax index based on support vector machine", *International journal on neural and mass-parallel computing and information systems. Neural Network World*, 26(3), pp. 219-232.
- Han, Q., B. Yan, G. Ning & B. Yu(2014), "Forecasting dry bulk freight index with improved SVM", *Mathematical Problems in Engineering*, 2014.
- Iftikhar, R., & M. S. Khan(2020), Social Media Big Data Analytics for Demand Forecasting, *Journal of Global Information Management*, Volume 28(1), pp. 103-120.

-
- Kamal, I. M., et al (2019). Forecasting high-dimensional multivariate regression of Baltic Dry Index (BDI) using Deep Neural Networks (DNN). *ICIC Express Letters*, 13(5), pp. 427-434.
- Kavussanos, M. G. & A. H. Alizadeh-M(2001), "Seasonality patterns in dry bulk shipping spot and time charter freight rates", *Transportation Research Part E: Logistics and Transportation Review*, 37(6), pp. 443-467.
- Kim, H., R. Jeong, & H. Bae(2019), "Hyper Parameter Tuning Method based on Sampling for Optimal LSTM Model", *Journal of The Korea Society of Computer and Information*, 24(1), pp. 137-143.
- Lechtenberg., et al(2019), Automatic identification system (AIS) data based ship-supply forecasting, *Hamburg University of Technology (TUHH)*, pp. 1-24.
- Leonov, Y. & V. Nikolov(2012), "A wavelet and neural network model for the prediction of dry bulk shipping indices", *Maritime Economics & Logistics*, 14(3), pp. 319-333.
- Li, J. & M. G. Parsons(1997), "Forecasting tanker freight rate using neural networks", *Maritime Policy & Management*, 24(1), pp. 9-30.
- Lim, S. & H. Yun(2018), "Supramax Bulk Carrier Market Forecasting with Technical Indicators and Neural Networks", *Journal of Navigation and Port Research*, 42(5), pp. 341-346.
- Luo, M., L. Fan, & L. Liu(2009), "An econometric analysis for container shipping market", *Maritime Policy & Management*, 36(6), pp. 507-523.
- Lyridis, D., et al(2004), "Forecasting tanker market using artificial neural networks", *Maritime Economics & Logistics*, 6(2), pp. 93-108.
- Martin stopford(2009), "Maritime Economics" 3rd ed., *Routledge*, pp. 1-815.
- Şahin, B., S. Gürgen, B. Ünver & İ. Altin(2018), "Forecasting the Baltic Dry Index by using an artificial neural network approach", *Turkish Journal of Electrical Engineering & Computer Sciences*, 26(3), pp. 1673-1684.
- Shynkevich, Y., T. M. McGinnity, S. A. Coleman, A. Belatreche, & Y. Li(2017), "Forecasting price movements using technical indicators: Investigating the impact of varying input window length", *Neurocomputing*, 264, pp. 71-88.

- Uyara, K., Ü. İlhan & A. İlhan(2016), "Long term dry cargo freight rates forecasting by using recurrent fuzzy neural networks", *Procedia Computer Science*, 102, pp. 642-647.
- Yang, Z. & E. E. Mehmed(2019), "Artificial neural networks in freight rate forecasting", *Maritime Economics & Logistics*, 21(3), pp. 390-414.
- Zeng, Q., C. Qu, A.K.Y. Ng & X. Zhao(2016), "A new approach for Baltic Dry Index forecasting based on empirical mode decomposition and neural networks", *Maritime Economics & Logistics*, 18(2), pp. 192-210.
- Zhang, J., Q. Zeng, & X. Zhao(2014), "Forecasting spot freight rates based on forward freight agreement and time charter contract", *Applied Economics*, 46(29), pp. 3639-3648.
- Zhang, X., M.Y. Chen, Y.E. Ge, H.E. Stanley & M.G. Wang(2019), "A novel hybrid approach to Baltic Dry Index forecasting based on a combined dynamic fluctuation network and artificial intelligence method", *Applied Mathematics and Computation*, 361, pp. 499-516.
- Zhongzhen, Y., J. Lianjie, & W. Minghua(2011), "Forecasting baltic panamax index with support vector machine", *Journal of Transportation Systems Engineering and Information Technology*, 11(3), pp. 50-57.

인터넷 자료

- 클락슨, <https://sin.clarksons.net/>(검색일: 2021.12.31.)
- 볼틱, <https://www.balticexchange.com/content/balticexchange/consumer/en/my-baltic/dashboard.html>(검색일: 2021.12.31.)
- 미디어, <https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af>(검색일: 2021.12.31)
- 미디어, <https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af>(검색일: 2021.12. 25.)
- IMF, <https://www.imf.org/en/Home>(검색일: 2021.12.31.)
- 야후, <https://finance.yahoo.com/>(검색일: 2021.12.31.)

부록

빅데이터 예측 방법

최근 해운과 관련하여 빅데이터를 이용한 연구가 진행되고 있는데 크게 두 가지로 구분된다. 첫째, 기존에 접근하기 어려울 뿐만 아니라 자료의 양이 큰 많은 AIS와 같은 자료를 이용한 연구이다. 비단 AIS 자료 외에도 최근 IoT의 발달로 인해 다양한 자료가 생성되고 있어 이를 활용한 연구들이 진행되고 있다.

둘째, 정형 자료에서 벗어난 비정형자료 사용되지 않은 자료의 확장이다. 시장을 예측하기 위해서는 정량적인 지표들을 활용한 연구들이 주를 이루었으나 최근에는 기사, SNS 등 비정형 자료에 대한 분석 방법이 발달해 관련 연구들이 진행되고 있다.

빅데이터를 이용한 대부분에 연구는 인공지능 방법론을 활용해 분석하며 함의를 도출하기 때문에 독자적인 연구로 보기 어려울 수 있다. 하지만 과거에 사용하지 않은 자료의 확장 개념으로 시장을 분석하고 예측하는 관점에서 인공지능 기반 예측과 구별된다. 빅데이터 기반 예측 모델은 대부분 인공지능을 활용한 분석 방법이 활용되므로 본고에서는 자료의 개념과 연구 목적을 소개하는 데 집중한다.

■ 선박자동식별장치자료(AIS)

선박자동식별장치(AIS: Automatic Identification System)는 항해 사고방지 및 항만관제를 위해 선박의 제원, 속도와 같은 정보를 자동 송수신할 수 있는 장치이다. 국제해사기구(IMO)는 1998년 해사안전위원회에서 AIS 관련 표준을 채택하여 총톤수 300톤 이상 국제항해 선박에 대해 2002년 7월부터 장착을 의무화했다. AIS 자료는 연안에서 선박 통항이 잦기 때문에 선박 관제, 충돌 방지 예방을 위해 사용되며 원양에서는 위성 안테나를 이용해 선속 또는 위치를 확인할 수 있다.

AIS 자료는 정적(static), 동적(dynamic), 항해(voyage) 정보를 담고 있는데 정적정보는 선박 제원에 관한 정보로 호출부호 및 선종 등이 입력되어 있다. 동적정보는 선박의 위치, 속도, 항해상태 등이 기록되어 있으며 자동으로 입력된다. 항해정보는 선장이 수동으로 기입하는 정보로 목적항만, 선박의 흘수 등을 입력한다. 이외에 문자통신은 기상정보나 항해시 발생하는 중요한 사안에 대한 정보이다.

〈표 1〉 AIS 정보 개요(계속)

구분	정보	비고
정적정보	• IMO번호(MMSI 번호) • 호출부호 및 선명 • 선박의 종류, 길이, 폭, 너비 • 안테나의 위치(선미/선수/중선선의 좌우)	변경사항 발생 시 수시 수정 입력
항해정보	• 선박의 흘수(draft, sea gauge) • 위험화물 • 목적지 및 도착예정시간 • 항로계획	항해 전 및 항해 중 주기적 수동입력

자료: 유동희 외(2008)⁷⁴⁾, p.39

74) 유동희·황소영(2008), AIS(자동식별시스템) 기술 개요, 한국정보통신학회지 제12권 2호, pp.38-45

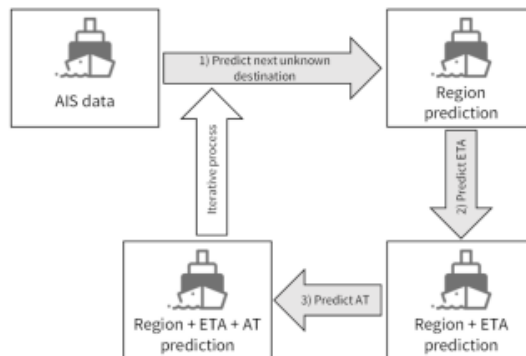
〈표 2〉 AIS 정보 개요

구분	정보	비고
동적정보	• 선박의 위치 • UTC 시간 • 대지침로(Course over Ground) • 대지속력(Speed over Ground) • 선수방위 • 항해상태(항해, 정박 등) • 선회율(Rate of Turn), 경사각도	선박의 항해상태에 따라 자동 입력 (수동입력 가능)
문자통신	• 중요한 항해 또는 기상정보 등	서비스

자료: 유동희 외(2008), p.39

Lechtenberg, Sandra 외(2019)의 연구에서는 용선료 지불을 위해 선박자동식별장치 자료를 이용하여 3단계 분석을 실시하여 선박 공급을 예측하였다. 1단계에서는 목적지를 예상하고 2단계에서 도착 시간, 3단계에서 정박시간을 예측하였다. 이를 위해 1단계에서 마코프(Markov) 의사결정 모델을 사용했고 2, 3단계에서 그래디언트 부스팅(gradient boosting)을 이용해 분석하였다.

〈그림 1〉 VAR/VECM 프로세스



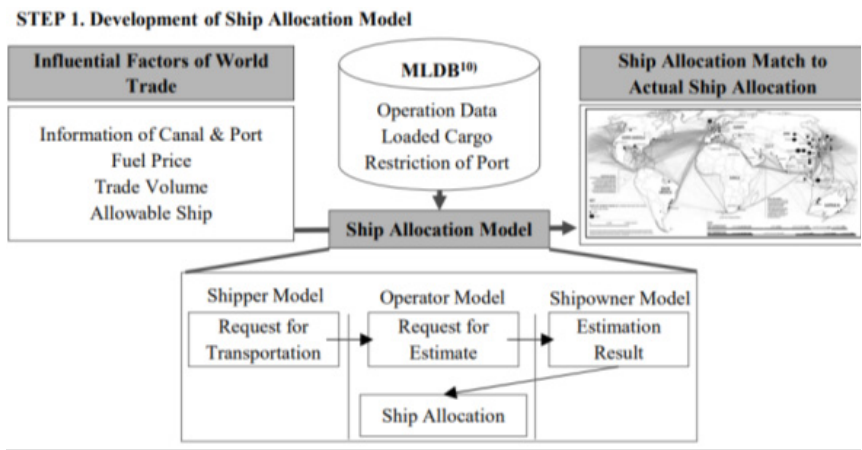
자료: Lechtenberg, Sandra 외(2019), p.10

Mohammad Danil Arifin(2018)⁷⁵⁾은 다양한 자료를 통합하여 선박 배정 모델을 개발했다. 동 연구는 AIS, 선박, 항만, 항로, 무역 정보를 통합한 해양물류데이터베이스(MLDB: Marine Logistics Database)를 이용해 호주-일본 간 벌크 선박 배정에 관한 연구를 실시했다. 동 연구는 AIS와 항만/선박 자료를 이용해 품목과 정시성을 확인했으며 화물량까지 추정하였다. 화물량을 흘수 데이터를 기반으로 추정하였는데 다음의 수식을 이용했다.

$$V_i(\text{ton}) = DWT_i \times \frac{(d_i - 0.2)}{(1 - 0.2)}$$

여기서, V는 화물량, DTW는 재화중량톤수, d_i 는 흘수이다.

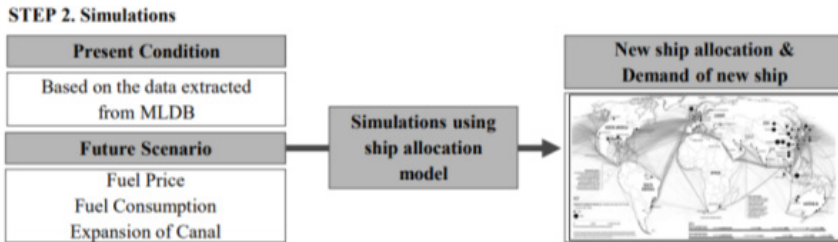
〈그림 2〉 선박 할당모델(계속)



자료: Mohammad Danil Arifin 외(2018), p.141

⁷⁵⁾ Mohammad Danil Arifin, Kunihiro Hamada, Noritaka Hirata, Kai Ihara, Yuki Koide(2018), pp.139-148

〈그림 2〉 선박 할당모델



자료: Mohammad Danil Arifin 외(2018), p.141

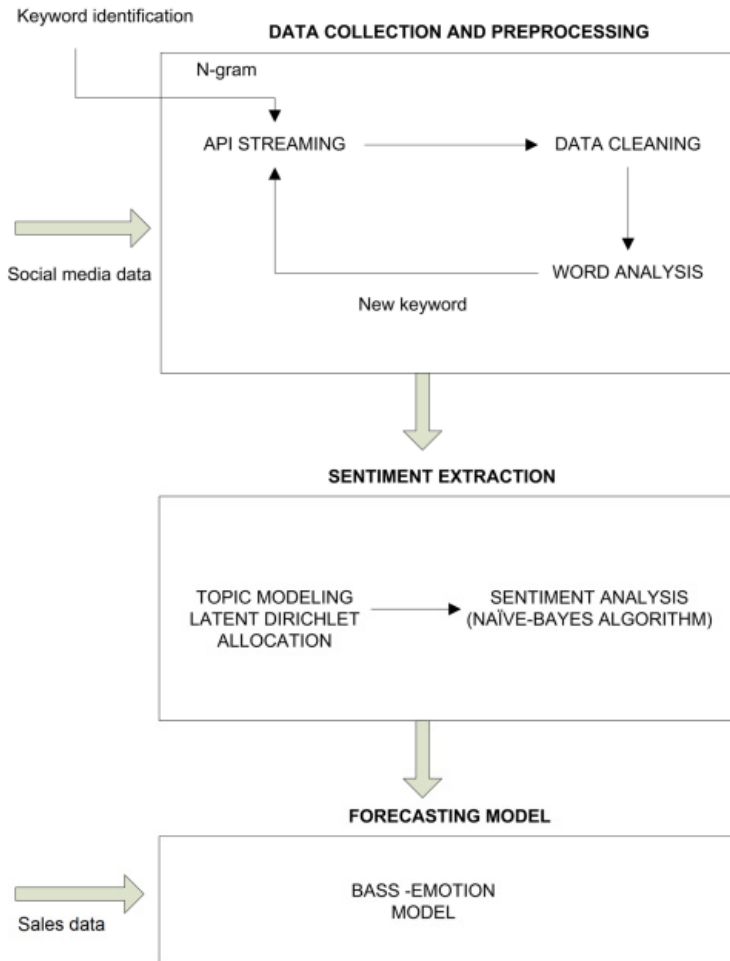
■ 비정형자료

Rehan Iftikhar 외(2020)⁷⁶⁾은 소셜미디어 자료를 수집하여 감성분석을 이용하여 수요를 전망하였다. 자료수집은 페이스북, 트위터 API를 이용하였으며 전처리를 통해 분석이 용이하게 자료를 가공하였다.

가공된 자료를 바탕으로 텍스트 분석 방법의 하나인 토픽모델링(LDA)을 실시하였으며 나이브-베이즈(Naive-Bayes) 알고리즘 감성분석을 실시하였다.

76) Rehan Iftikhar, Mohammad Saud Khan(2020), pp.103-120

〈그림 3〉 소셜미디어를 이용한 수요 예측 프로세스



자료: Mohammad Danil Arifin 외(2018), p.141

기본연구보고서 발간목록

Ⅰ 2021년

01	시민참여형 해안돌봄 정책 도입방안 연구	정지호
02	해양 플라스틱 쓰레기가 선박 운항에 미치는 영향분석 연구	김보람
03	여성어업인 노동정책 방향 연구	홍혜수
04	순환경제 시스템을 활용한 어업폐기물의 자원화 방안 연구	고동훈
05	1conomy 시대, HMR 시장 확대에 따른 수산부문 대응전략 수립 연구	마창모
06	AIS 데이터 기반 해상교통 안전 평가모델 개발 연구	황선일
07	항만 에너지 관리시스템 도입을 위한 로드맵 구축 연구 - 부산항 신항 전기에너지 소비를 중심으로 -	김근섭
08	해양분야 리빙랩 활성화 방안 연구	좌미라
09	빅데이터의 연계·활용을 통한 선박의 배출량 산정체계 고도화 방안 연구	안용성
10	해양공간계획 집행체계 고도화 방안 연구	최희정
11	수산식품산업 푸드테크 환경분석 및 적용전략 연구	이상건
12	AIS데이터기반 해상물동량 추정 연구	황수진
13	항만부문 정책의사결정 지원 빅데이터 플랫폼 구축 방안 연구	이기열
14	항만 컨테이너 물류 프로세스 디지털화 모형 연구	서정용
15	해양수산 분야 국제산업연관표 구축기반 연구	정수빈
16	해양수산 과학기술 정책평가모형 연구 - 해양수산 R&D를 중심으로 -	전형모
17	글로벌 경제위기와 해운산업 대응체계 연구	박성화
18	항만산업 경기진단체계 구축 연구 - 경기동행지수 개발을 중심으로 -	김성아
19	전국 무역항 부두 분류 체계 개선방안 연구	이수영
20	IMO 규제기반 해사산업의 글로벌 지속발전방안 연구 - 新해사산업의 육성 및 지원을 위한 법제화 연구(5차년도) -	박한선
21	수소에너지 거점 구축을 위한 항만의 대응방안 연구	신수용

2020년

01	해양수산업의 지역 간 연관구조 분석	장정인
02	해양수산 분야 기술 대외의존도 분석연구 - 스마트항만을 중심으로 -	전형모
03	지역 해양수산 오픈 플랫폼 구축방안 연구	최지연
04	갯벌 거버넌스 개선방안에 관한 연구	육근형
05	해양환경정책의 능동적 추진을 위한 재원체계 개선 방안 연구	박수진
06	항만 대기환경 관리 표준 및 평가모형 연구	안용성
07	해양수산분야 사회문제해결형 R&D 기반 구축 연구	좌미라
08	해양 유입 하천쓰레기 관리체계 개선방안	이윤정
09	수산식품 품질·안전관리 제도 개선방안 연구	이헌동
10	국제법 변화에 대응한 어선원 안전 및 권리 제고방안 연구	한덕훈
11	스마트 양식 클러스터 추진 방안 수립 연구	이상철
12	해양포유류 보호에 관한 수산업 대응 방안 연구	정명화
13	수산물 수급통계 개선 방안 연구	김수현
14	IMO 온실가스 규제 대응 정책방향 연구	박한선
15	퇴직전문가 활용을 통한 해운업 경쟁력 강화방안 연구	안영균
16	글로벌 선사들의 물류통합화 전략에 대한 국적선사의 대응방안	전형진
17	내항여객운송항로 정책 발전방안 연구	김태일
18	블록체인 기술기반 식품콜드체인 체계 구축 연구	조지성
19	항만자동화 도입 관련 노무 갈등 해소 방안 연구	김찬호
20	스마트항만과 스마트도시 연계 발전 방안 연구	이연경
21	항만의 회복탄력성 측정 모형 구축에 관한 연구	김성기
22	IMO 규제기반 해수산업의 글로벌 지속발전방안 연구 - 新해수산업의경제적파급효과분석연구(4차년도) -	박한선
23	국내 항만연계 산업의 가치사슬 및 공급사슬 연계성 강화방안 - 자동차 산업을 중심으로	신수용

수시연구보고서 발간목록

Ⅰ 2021년

01	해양환경산업 육성 방안 연구	한기원
02	선원의 인권 및 근로여건 향상을 위한 선원근로감독관 제도 개선 방안 연구	허성례
03	포스트 코로나 시대를 대비하는 지역 해양축제 활성화 연구	최일선
04	회복탄력성 개념을 적용한 항만 위기관리 시스템 구축 연구	김성기
05	중대재해처벌법 시행에 따른 항만에서의 대응방안 연구	최상균
06	크루즈산업 COVID-19 방역체계 구축방안 연구	황진희
07	한국형 선주사의 최적 운영방안 연구	고병욱
08	어촌형 생활서비스 전달체계 개선 연구	이호림
09	원양산업의 ESG 도입 기초 연구	윤미경
10	수상레저활동 관리체계 개선방안 연구	홍장원
11	연안지역 자연성 회복을 위한 정책 방향_간척지·담수호 중심으로	정지호
12	스마트항만 구축에 따른 탄소저감 효과 연구 - 동력전환을 중심으로 -	김가현

Ⅰ 2020년

01	포스트 코로나 19 해양수산 분야 정책방안	박광서
02	생분해성 어구 사용 활성화 방안 연구	심성현
03	해양법 전문인력 양성 방안 연구	박영길
04	무인도서 해양주권 강화와 이용 활성화를 위한 제도 개선방안	정지호
05	지역 해양수산 재정분권 대응방향	황재희
06	데이터 3법 개정에 따른 항만·물류 데이터 활용도 제고 방안	이기열
07	양식 활어 유통 효율화 방안 연구	마창모
08	해양레저관광사업 추진을 위한 제도정비 방안	홍장원
09	해양바이오기업의 규제 정비 방안 연구	최석문
10	비상체제 시 선원의 안전을 위한 선박-항만-항공 이동경로 구축방안	이혜진

일반연구보고서 발간목록

Ⅰ 2021년

01	해양수산 정책영향평가를 위한 기초 연구	김주현
02	해양교육의 사회·경제적 가치 평가 연구	이슬기
03	해양수산업 경기진단체계 기초연구 - 해운업을 중심으로	권장한
04	선박 기술진보를 고려한 탄소 배출량 추정 연구	최건우
05	서포트벡터머신(SVM) 기법을 활용한 해운시황 예측 연구	김병주
06	크루즈 여객 수요 전망 모형 구축 연구	안승현
07	디지털 공급사슬 물류정보통합 구축전략 연구(Ⅲ)	서정용
08	수출용 전복의 유통경로 분석 연구	이정필
09	인공지능기반 해상운임 예측 연구(3차년도)	황수진
10	인공신경망모형을 이용한 양식수산물 단수 전망에 관한 연구 - 김 양식을 중심으로 -	천성훈
11	해양수산분야 미래 리스크 발굴 및 파급효과 분석 연구	박광서

Ⅰ 2020년

01	지역 해양관광 경쟁력 지수 체계화 연구	최일선
02	AIS 기반 글로벌 선박 배기가스 배출량 분석 연구	강무홍
03	김 중기 수급전망모형 「Gim-MFoS」 구축 연구	허수진
04	중앙 북극 공해 비규제어업방지협정 이행방안 연구	김민수
05	동해 평화관광구역 조성 방안 연구	윤인주
06	디지털 공급사슬 물류정보통합 구축전략 연구(Ⅱ) - 일반 수출입 컨테이너 정보교환방식 중심 -	이연경
07	스트레스 테스트를 통한 우리나라 해운·조선 기업의 안정성 분석	박성화
08	인공지능기반 해상운임예측 연구	황수진
09	국제물류주산업 실태분석 및 경쟁력 제고 방향 연구	최나영환
10	디지털화에 따른 개별 직업의 대체 가능성 추정 연구	박희대
11	시계열 분석을 통한 해운시장 분석 및 예측 연구	고병욱
12	해양수산분야 글로벌 경제효과 분석모형(KMI-GEM) 시범 구축 연구	임병호
13	해운항만물류 인력양성사업 관리운영 제도 개선 방안 연구	이자연

일반연구 2021-09

인공지능기반 해상운임 예측 연구 (3차년도)

인쇄 2022년 1월 29일

발행 2022년 1월 31일

발행인 김 종 덕

발행처 한국해양수산개발원

주소 49111 부산시 영도구 해양로 301번길 26(동삼동)

연락처 051-797-4800 (FAX 051-797-4810)

등록 1984년 8월 6일 제313-1984-1호

조판·인쇄 크리커뮤니케이션 (02-2273-1775)

판매 및 보급: 정부간행물판매센터 Tel: 02-394-0337

정가 6,000원