AI/빅데이터 전략

리서치본부
2021 년 10 월 19 일

김규진(퀀트전략)
02)2229-6737
kyoujin.kim@nhqv.com

AI 와 대체데이터를 활용한 KOSPI 예측 모델

자율주행 등 제어시스템 분야와 음성, 영상 인식 등 신호처리 분야에서는 AI 가 이미 뛰어난 성능을 보이고 있습니다. 아직 더디게 진행되고 있는 금융 분야에서의 AI 도입 현황을 살펴보고, AI 기술을 활용한 KOSPI 예측을 시도해 보았습니다.

NH투자증권

CONTENTS

Ⅳ. AI 와 대체데이터를 활용한 KOSPI 예측 모델

V. Appendix

I. Summary

자율주행 등 제어시스템 분야와 음성, 영상 인식 등 신호처리 분야에서는 AI가 이미 뛰어난 성능을 보이고 있습니다. 아직 도입이 더디게 진행되고 있는 금융 분야에서의 AI 현황을 살펴보고, AI 기술을 도입한 KOSPI 예측을 시도해 보았습니다.

AI가 주가도 맞출 수 있을까

2010년 이후 AI는 컴퓨터의 비약적인 성능 향상에 힘입어 각종 신호처리 분야에서 가장 주목받는 기법이 되었다. 2017년부터는 금융학계에서도 AI에 대한 연구가 활발해졌으며, 금융산업에서도 응용을 시도하고 있다.

미국 ETF 시장에서도 AI 기법으로 운용되는 ETF들이 상장되고 있으나, 아직 금융 분야에서의 AI 적용 선례가 부족하여 운용 성과를 보장할 수는 없는 상황이다. 이에 본 보고서에서는 AI와 대체데이터를 활용한 KOSPI 예측 모델을 통하여 한국 주식시장에의 AI 적용 가능성을 점검해보았다.

AI와 대체데이터를 활용한 KOSPI 예측 모델

KOSPI 예측 모델의 학습에는 ①경제지표, ②펀더멘털지표와 대체데이터인 ③감성지표가 사용되었다. 결과적으로 조정 KOSPI를 포함하여 경제지표 19개, 펀더멘털지표 8개, 감성지표 10개 총 38개의 입력변수를 선정하였다.

수익률 측면에서 KOSPI 예측 모델은 KOSPI 대비 연평균 6%p의 초과수익률을 기록하였다. 샤프 지수는 KOSPI 대비 0.37p 개선되었으며, 최대낙폭은 11.8%p 개선되었다.

한편, 모델에서 감성지표를 제외할 경우, 수익률이 큰 폭으로 악화되었으며, 최대낙폭도 증가하였다. 감성지표의 KOSPI 예측 모델의 수익률 기여도가 높음을 확인할 수 있다.

표1. KOSPI 예측 모델 성과 개요

성과요약표	KOSPI 예측 모델	감성지표 제외 모델	KOSPI
연율화 수익률(%)	12.10	-3.16	6.09
샤프 지수(x)*	0.74	-0.19	0.37
소티노 지수(x)*	1.21	-0.25	0.46
최대낙폭(%)	-27.42	-65.72	-39.17

주 1: 2014 년 03 월 14 일부터 2021 년 10 월 15 일까지 백테스트 결과

주 2: 샤프 지수 = 연간수익률/연간변동성

주 2: 소티노 지수 = 연간수익률/연간하방변동성

자료: NH 투자증권 리서치본부

II. AI(인공지능)와 머신러닝

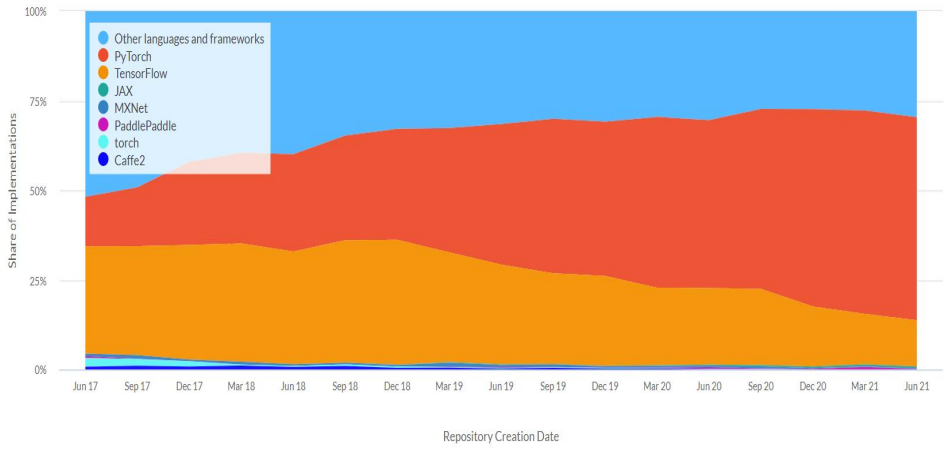
1. 꿈의 전령, AI

머신러닝 AI 개발 언어 시장은 현재 구글의 Tensorflow와 페이스북의 PyTorch가 양분하고 있다. 이 중 Tensorflow에서 제공하는 Keras라는 개발 언어는 사용자 친화성을 기본 원칙으로 한다. Keras를 사용하면 비전문가도 손쉽게 머신러닝 AI를 개발할 수 있다.

Keras의 개발 초기 프로젝트 가칭은 ONEIROS(Open-ended Neuro-Electronic Intelligent Robot Operating System)이었다. Oneiros는 그리스 신화에 등장하는 꿈의 전령이다. 신화에 따르면 꿈의 전령은 뿔(Keras)로 만든 문을 통해서 미래를 보여주는데, 가끔은 거짓된 미래를 보여주기도 한다. 호메로스의 '일리아드'에서는 꿈의 전령이 아가멤논에게 보여준 거짓된 미래가 비극적인 트로이 전쟁을 불러온다.

AI는 꿈의 전령과 같이 미래를 보여주는 예측 도구가 될 수도 있으나, 기대대로 작동하지 않을 가능성도 있다. 특히 현상과 원인이 명확한 자연현상과 달리 금융은 인간의 기대와 행동이 현상을 결정하므로 AI를 통한 예측이 맞지 않을 가능성이 큰 분야에 해당한다.

그림1. AI 개발 언어 시장은 Tensorflow와 PyTorch가 양분



주: Github 에 등록된 머신러닝 관련 프로젝트 생성수 기준
자료: paperswithcode.com, NH 투자증권 리서치본부

1) AI와 머신러닝

AI는 인간의 지능을 모사하는 기계

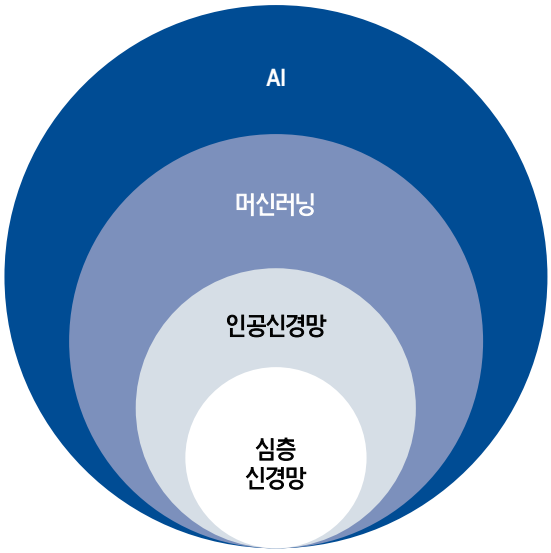
IBM에 따르면, **AI는 인간의 지능을 모사하는 기계**를 의미한다. 인간의 지능이란 외부의 정보를 분석하여 문제를 해결하는 능력이다. 그림을 보고 고양이 그림이 무엇인지 분류하거나, 자동차를 운전하여 목적지까지 도착하는 것도 AI를 통해 모방하고자 하는 인간의 지능에 포함된다.

머신러닝(Machine Learning, ML)은 이와 같은 AI를 만들기 위한 방법론 중 하나이다. 1959년 Arthus Samuel이 정의한 바에 의하면 **머신러닝은 외부의 프로그래밍 없이 스스로 학습할 수 있는 능력을 기계에게 부여하는 것**이다(Appendix A 참조). 기존에는 AI에게 인간의 행동 규칙을 일일이 가르쳐 주어야했으나 머신러닝의 발달로 이러한 과정을 생략할 수 있게 되었다.

최근 각광받고 있는 인공신경망(Aritificial Neural Network)은 머신러닝을 효과적으로 수행하는 알고리즘 중 하나이다. **인공신경망이라는 명칭은 입력값과 출력의 결합이 인간의 신경망과 비슷한 점에서 유래**되었다(Appendix B 참조).

인공신경망을 다중으로 연결하면 머신러닝이 더욱 효과적으로 이루어지는데, 이러한 **다중 인공신경망 구조를 심층신경망(Deep Neural Network)**이라고 하며, 심층신경망을 사용한 머신러닝을 딥러닝(Deep Learning)이라 부른다(Appendix C 참조).

그림2. AI > 머신러닝 > 인공신경망 > 심층신경망



자료: NH 투자증권 리서치본부

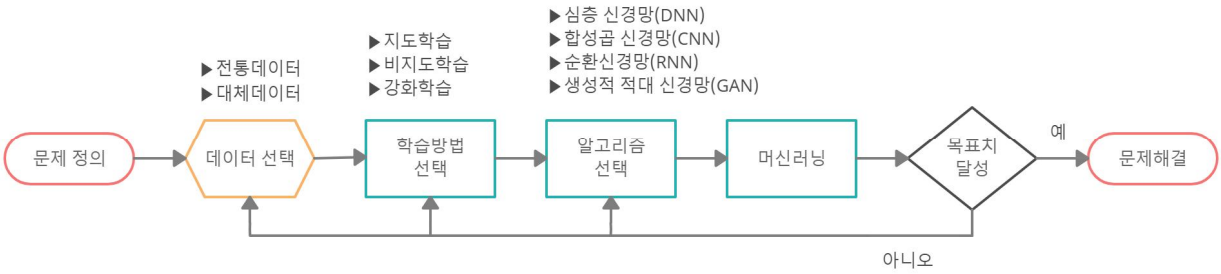
2) 머신러닝 학습방법의 종류

머신러닝 AI 모델 설계는 **'학습방법 선택 → 알고리즘 선택'** 과정을 통해 이루어지며, 알고리즘 외에 학습방법 역시 머신러닝에서 중요한 요소이다(그림3 참조).

학습방법은 기계가 무엇을 기준으로 학습할 것인지에 대한 정의이다. 기계가 주어진 정답과 최대한 일치하게 행동하도록 지도하는 방법을 ①지도학습(Supervised Learning)이라 한다. 정답이 정해져 있지 않으나 특정한 규칙에 따라 기계가 행동하여 수렴된 결과물을 만드는 방식은 ②비지도학습(Unsupervised Learning)으로 분류된다.

한편, 연속되는 시간상에서의 선택이 누적되어 결과가 결정되는 상황에서는 ③강화학습(Reinforcement Learning) 알고리즘을 사용한다(그림3 참조).

그림3. 머신러닝 문제해결 과정



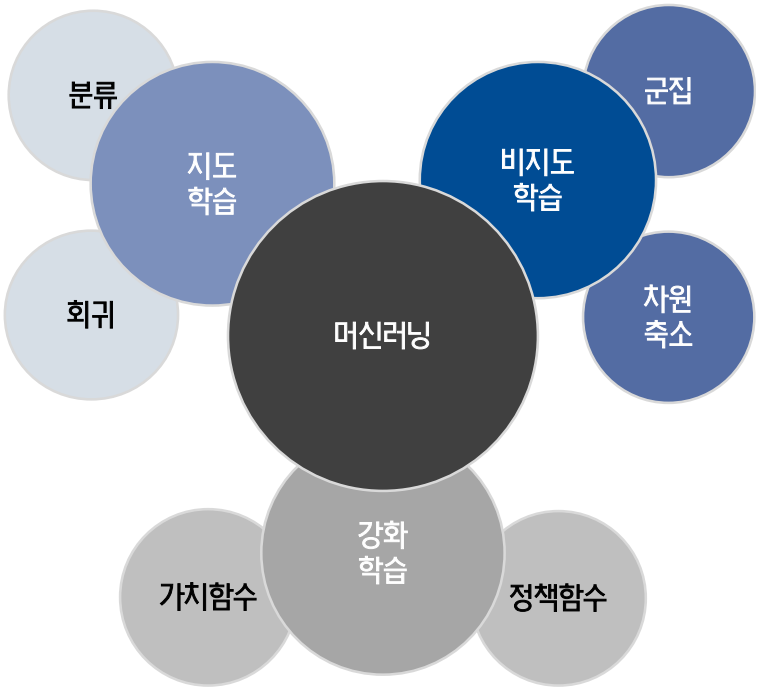
자료: NH 투자증권 리서치본부

세부 알고리즘으로는 지도학습 내 분류, 회귀 알고리즘이, 비지도학습 내 군집, 차원 축소 알고리즘이 대표적이다. 예를 들어, 개인 신용 위험을 평가하는 모델의 결과값으로 고객의 신용점수와 같은 연속된 값을 알고 싶은 경우에는 지도학습 중 회귀(Regression) 알고리즘, 고객의 신용등급이나 만기일 상환 여부 등을 분류하고 싶을 경우에는 분류(Classification) 알고리즘을 사용할 수 있다.

고객의 기존 신용점수나 신용등급 등 입력할 수 있는 정답이 없는 경우에는 비지도 학습의 군집(Clustering) 알고리즘을 사용해 고객의 특성별로 집단을 나누어 신용등급을 파악할 수 있다.

특성 데이터의 개수가 너무 많아 모델의 성능이 저하될 가능성이 있거나 입력값에 노이즈가 너무 많을 때에는 비지도 학습의 차원축소(Dimensionality Reduction) 알고리즘을 사용하게 된다.

그림4. 머신러닝 알고리즘의 종류



자료: Ullah, Zaib, et al.(2020), “Applications of Artificial Intelligence and Machine Learning in Smart Cities.”, NH 투자증권 리서치본부

III. 금융 분야의 머신러닝 활용 현황

1. 금융 분야의 머신러닝 연구

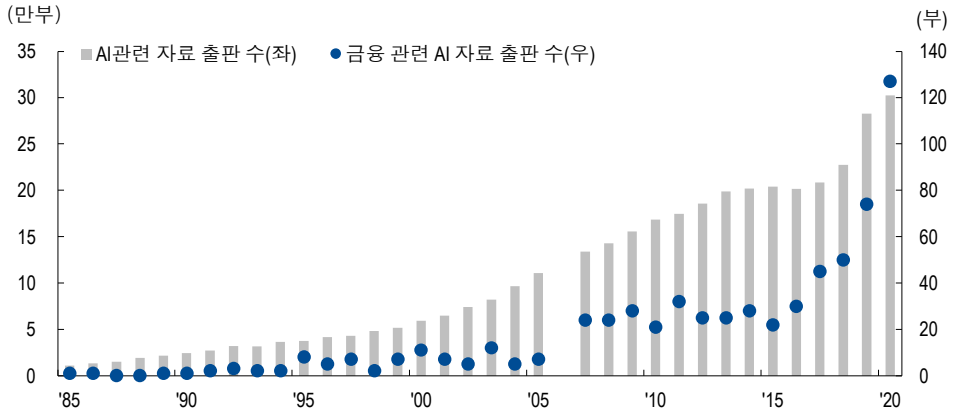
금융분야에서의 머신러닝 연구는 2017년부터 본격화

Microsoft Academic 기준으로 금융 분야에서 AI 관련 논문이 등재되기 시작한 것은 1986년부터이며 **2000년대 중반부터 머신러닝 관련 금융학계 논문의 개수가 확연하게 증가**하였다. 2000년대까지는 전통적인 경제, 금융 모델을 발전시킬 목적으로 연구가 진행되었다면, 2010년대부터는 컴퓨터 성능의 비약적인 개선에 힘입어 머신러닝 모델 자체(딥러닝 기반 모델)를 통한 금융 데이터 분석이 시도되고 있다.

2017년부터는 장단기메모리(LSTM)[주1] 시계열 처리 딥러닝 알고리즘의 개발에 힘입어 금융 분야의 자료 출판 건수가 급격히 증가한다(다음 페이지 그림5, 6 참조). 장단기메모리 알고리즘은 자연어 처리 분야에서 특히 뛰어난 성능을 보여주며, 이를 활용하여 만든 모델은 텍스트에서 작성자의 감정과 태도를 추출하는 작업에서도 좋은 성능을 보였다. 이에 따라 금융학계에서도 정성데이터를 자연어처리 기술을 사용하여 정량데이터로 변환하고 금융모델에 적용하려는 시도가 확산되고 있다.

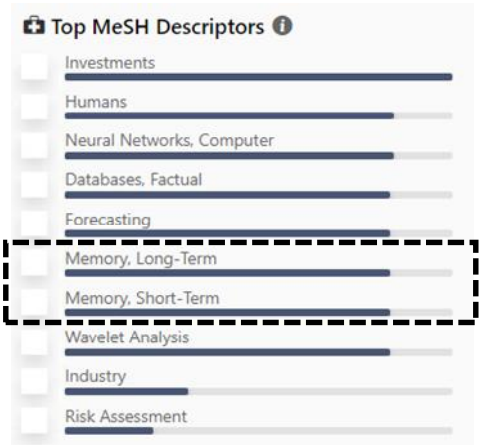
주1) 장단기메모리(LSTM: Long Short-Term Memory): RNN의 종류 중 하나이다. 순환 신경망(RNN)[주2]은 과거의 데이터를 참조하여 출력값을 결정하는데, 더 먼 과거를 참조해야할수록 기존의 순환 신경망은 잘 작동하지 않는다. 따라서 LSTM은 이전값을 다음으로 전달하는 통로를 따로 만들어 과거의 값을 전달하고, 해당 통로의 문을 열지 닫을지를 학습시킨다. 먼 과거 데이터를 참조해야하는 경우에 LSTM은 기존 순환 신경망보다 좋은 성능을 보인다.
주2) 순환 신경망(RNN: Recurrent Neural Network): 시퀀스(순서가 존재하는 데이터: 시계열 데이터 또는 한글 문장 등)를 처리할 때, t번째의 입력이 t-1번째의 출력을 참고하여 데이터를 처리하는 기법

그림5. 연도별 AI 관련 자료(논문, 도서 등) 출판 수



주: AI 관련 출판 자료 중 'Finance' 관련 주제 포함 여부로 스크리닝
자료: Microsoft Academic, NH 투자증권 리서치본부

그림6. 금융학계 머신러닝 관련 논문에서 LSTM 관련 Keyword가 높은 비중을 차지



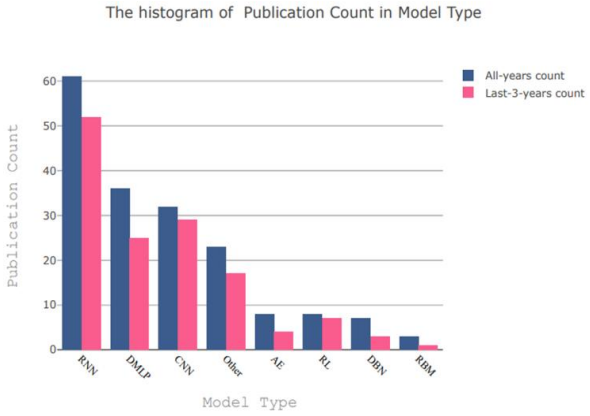
자료: Microsoft Academic, NH 투자증권 리서치본부

구글의 딥마인드가 알파고에서 사용한 강화학습 알고리즘 역시 금융학계에 도입되기 시작하였다. Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017)는 금융 시계열에 강화학습을 적용하여 유의미한 마켓 타이밍 시그널을 얻을 수 있음을 보였다.

현재는 **정성데이터의 정량화(금융 텍스트 마이닝, 금융 감성 분석), 금융데이터 분류(리스크 가능성 분석, 금융사기 감지) 및 예측(알고리즘 트레이딩, 포트폴리오 운용) 분야에서 주로 연구**가 이루어지고 있으며, 순환 신경망(RNN), 장단기메모리, 강화학습은 물론 이미지 처리에 좋은 성능을 보이는 합성곱 신경망(CNN)[주1]과 같은 알고리즘의 적용도 시도되고 있다(그림7, 8 참조).

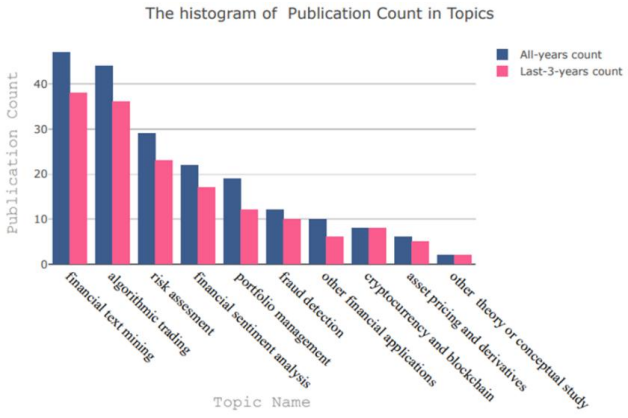
주1) 합성곱 신경망(CNN: Convolution Neural Network): 종류가 복잡한 데이터일수록 머신러닝 모델은 더 많은 비중값을 학습해야한다. 합성곱 신경망은 데이터를 한번에 보는 대신에 여러 개의 돋보기로 일부분을 조금씩 보는 방식으로 한번에 학습해야할 비중값을 감소시키는 방식이다. 학습해야할 비중값을 획기적으로 감소시켜 학습 시간을 단축함과 동시에 데이터의 특징을 효과적으로 추출하여 이미지 처리 분야에서 큰 성과를 거둔 알고리즘이다.

그림7. 전체 연구에서 금융 시계열 처리 모델이 가장 높은 비중을 차지



자료: Ozbayoglu, A. M et al.(2020). Deep Learning for Financial Applications: A Survey., NH 투자증권 리서치본부

그림8. 가장 많이 다뤄진 주제는 금융 텍스트 분석과 알고리즘 트레이딩



자료: Ozbayoglu, A. M et al.(2020). Deep Learning for Financial Applications: A Survey., NH 투자증권 리서치본부

2. 금융 분야의 머신러닝 활용

1) 로보 어드바이저(Robo Advisor)

로보 어드바이저는 금융의 AI 접목 가능성을 가장 먼저 보여주었다. 로보 어드바이저는 인터넷으로 가입 가능하며, AI를 활용한 투자성향 및 투자목표 상담으로 시간과 공간에 제약이 없다. 이와 같은 특성은 특히 젊은 층에게 어필하는 양상을 보이는데, CNBC에 따르면 2020년 10월 기준 Vanguard Digital Advisor 서비스 가입자 중 20~30대가 66%로 높은 비중을 차지하였다.

로보 어드바이저는 수익률의 제고가 아닌 자산배분의 자동화를 통한 비용 최소화를 목적으로 AI 기술을 사용한다. 실제 미국에서 기존 자문업체들이 1~2%의 수수료를 수취하는 반면, 로보 어드바이저의 수수료는 0.5% 수준에 불과하다(표2 참조).

표2 2020년 12월 31일 기준 미국 로보 어드바이저 회사의 평균 수수료

회사명	총 운용자산 (백만달러)	운용수수료
Fidelity Go	-	1 만달러 이하: 무료, 5 만달러 이하 월간 3 달러, 5 만달러 이상 연간 0.35%
Morgan Stanley	-	연간 0.35%
Wells Fargo	-	연간 0.35%
Vanguard	212,000	연간 0.3%. 500 만달러 이상부터 감소
Schwab	57,900	일반: 무료, 프리미엄: 최초 300 달러 + 월간 30 달러
Betterment	18,065	Digital: 0.25%, Premium: 0.4%
Personal Capital	16,000	연간 0.89%
Wealthfront	15,847	연간 0.25%
Wealthsimple	8,400	연간 0.5%, 10 만달러 이상부터 감소
United Income	1,199	연간 0.99%, 100 만달러 이상부터 감소
FutureAdvisor	1,135	연간 0.5%
SigFig	835	1 만달러 이하 무료, 1 만달러 이상 연간 0.25%
Ellevest	634	서비스별 월간 1, 5, 9 달러
SoFi	203	무료

주: 연간 수수료 %는 계좌별 운용자산 기준
자료: Bankend Benchmark, NH 투자증권 리서치본부

다수의 미국 로보 어드바이저사는 AI를 통해 **TLH 전략(Tax-Loss Harvesting)을 고객에게 제공**하는 서비스를 강점으로 내세우고 있다. TLH 전략은 기말 실현이익 이익을 포트폴리오 내 미실현손실을 통해 제한하여 양도세 부과 기준이 되는 과세액을 감소시키는 절세 전략이다(Appendix D 참조). 사람이 모든 계좌에 대해 이와 같은 서비스를 제공하는 것은 불가능하지만 알고리즘에 따라 작동하는 로보 어드바이저는 충분히 가능하다.

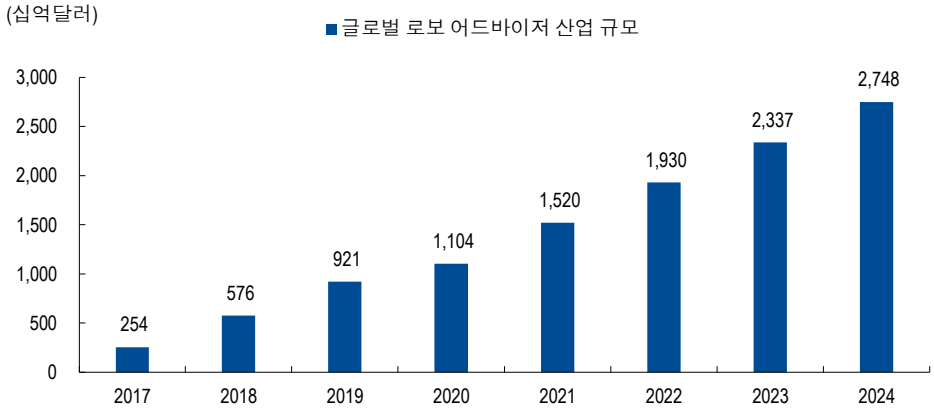
로보 어드바이저 서비스를 제공하고 있는 qplum에 따르면 잘 작동한 절세 전략은 2014년부터 2017년까지 세후 연간수익률을 6.08%에서 7.72%로 1.64%p 높였다(그림9 참조).

이러한 장점을 기반으로 미국의 로보 어드바이저 시장은 2018년부터 2020년까지 총 운용자산이 80% 증가하는 등 가파른 성장세를 보여주고 있다. Statista는 글로벌 로보 어드바이저 산업의 총 운용자산이 2024년에는 2.7조달러까지 증가할 것으로 전망했다(그림10 참조).

그림9. 60/40 포트폴리오에 절세전략을 시행할 경우의 장기 누적성과(파란색선)

자료: qplum(2018), Tax Optimization in Portfolio Management using A.I

그림10. 글로벌 로보 어드바이저 산업 규모 전망



자료: Statista.com

국내 로보 어드바이저 시장 역시 2025년까지 30조원대로 성장할 것으로 전망된다(그림11 참조). 파운트, 에임, 핀트 등 핀테크 로보 어드바이저가 국내 시장을 주도하고 있으며, 특히 파운트와 핀트는 포트폴리오 종목 선정 및 투자 비중 결정에 머신러닝 모델을 사용한다고 밝히며 머신러닝 기반 AI 도입에 적극적인 모습을 보이고 있다(표3 참조).

그림11. 한국 로보 어드바이저 시장 규모 전망치

자료: KEB 하나은행 하이로보센터, NH 투자증권 리서치본부

표3 국내 주요 로보 어드바이저 핀테크 기업

기업명	운용자산 (억원)	운용수수료	비용 (%)
파운트	8,074	연간 수익의 15%	0.00
에임	3,159	연간 5 만원, 계약금액의 1.0% 중 높은 금액	0.08
디셈버앤컴퍼니(핀트)	307	연간 수익의 9.5%	0.07

주: 2020 년 말 기준
자료: 파운트, 에임, 디셈버앤컴퍼니, NH 투자증권 리서치본부

2) 대체데이터(Alternative Data) 활용

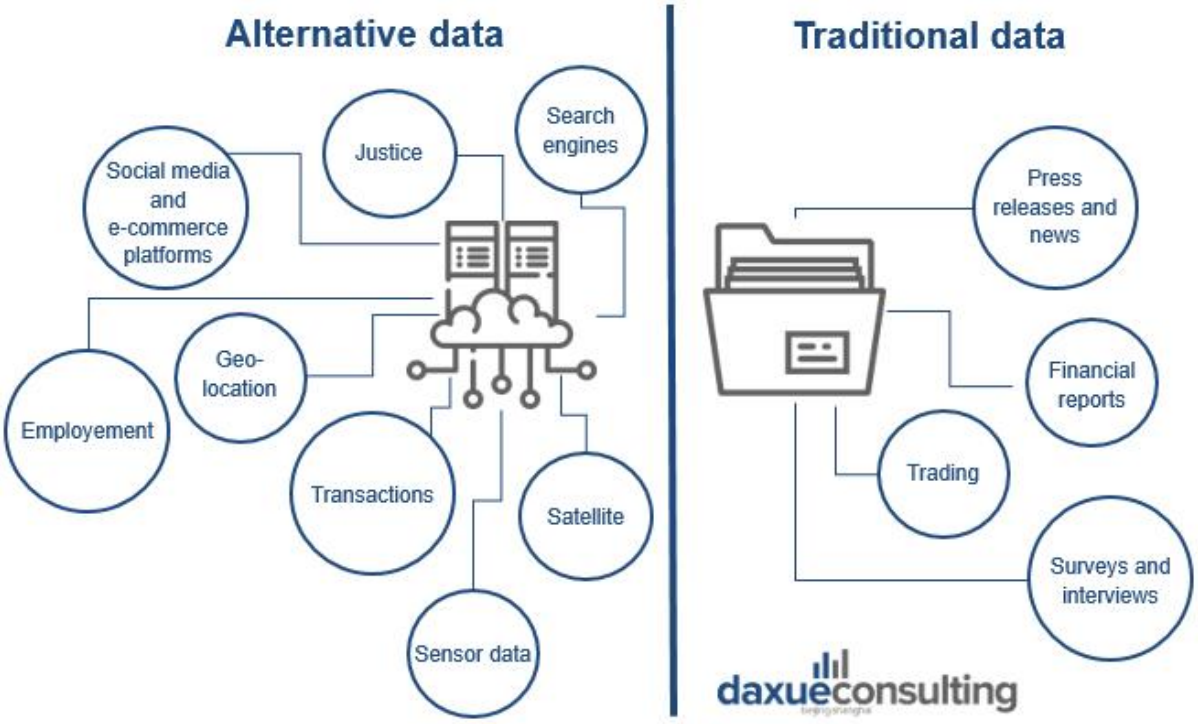
대체데이터를 활용해 알파를 찾을 수 있음

머신러닝 기술의 발달에 따라 기존에는 분석하기 어려웠던 사진, 영상, 음성 등을 처리하는 작업이 가능해졌다. 이에 따라 금융시장에서도 위치 정보, 신용카드 이용 데이터, 웹 트래픽, 위성사진, 소셜 미디어 등의 대체데이터를 가공하여 제공하는 서비스가 빠른 속도로 발전하고 있다.

대체데이터는 전통적인 데이터와는 달리 사용방법이 난해하며 접근성이 제한적이다. 이런 특징으로 **데이터의 수익률 알파 훼손이 상대적으로 적으며, 차별화된 투자가 가능**하다.

예를 들어, 2010년부터 위성사진을 수집 및 분석하여 각종 원자재의 생산과 저장량 추정에 활용하고 있는 RS Metrics은 2015년 2분기 JC Penney의 물동량이 급증했음을 파악하였고, 실제로 JC Penney는 2015년 8월, 2분기 어닝 서프라이즈를 발표하였다. RS Metrics의 대체데이터를 사용하여 JC Penny에 투자했다면 한달 동안 10%의 수익률을 얻을 수 있었을 것이다. 또한 CNBC에 따르면 데이터기업 ThinkNum은 Reddit 데이터를 통해 GameStop 광풍 하루 전에 관련 언급이 급증했음을 확인하였다.

그림12. 전통적 데이터와 대체데이터의 구분 및 종류

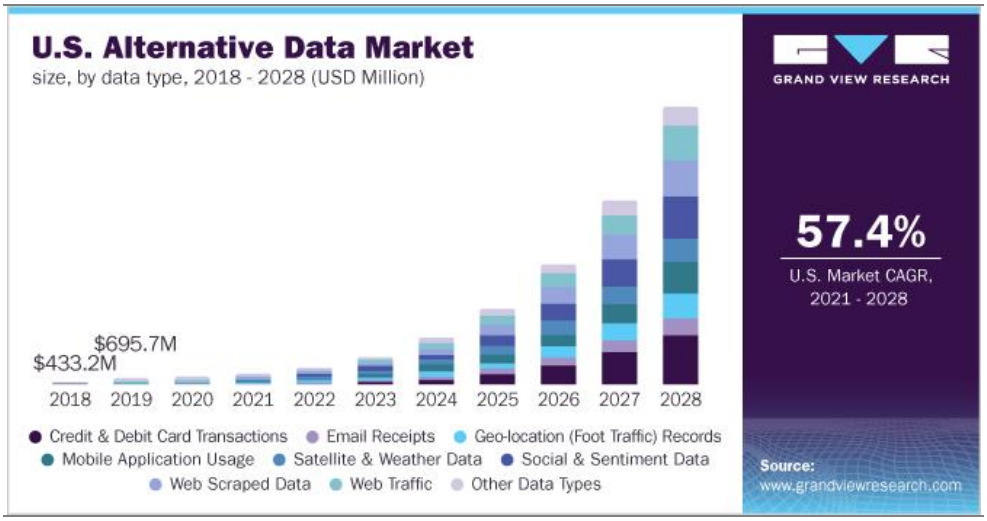


자료: daxue Consulting, NH 투자증권 리서치본부

대체데이터 시장의 규모 역시 가파르게 확대되고 있다. Grand View Research에 따르면 2020년 글로벌 대체데이터 시장 규모은 17.2억달러였으며, 2021년부터 2028년까지 연평균 58.5%의 성장세를 기록할 것으로 전망된다(그림13 참조).

현재 제공되는 대체데이터는 감성 데이터, 위성 데이터, 드론 데이터, 센서 데이터 등이 있다. 감성 데이터는 소비자, 투자자, 소셜 미디어, 뉴스 등의 텍스트를 분석하여 긍정/부정 또는 객관성/주관성 등을 평가한 점수이다. 해당 텍스트와 기업간의 관계를 분석하는 것도 가능하다. 위성, 드론 데이터는 인공위성 또는 드론이 촬영한 사진을 분석한다. 소매점에 주차되어 있는 자동차 대수, 내방객 숫자, 정유소 가동 현황 등을 통하여 기업의 영업활동 현황을 추정하는 것이 가능하다. 센서 데이터는 산업현장에 배치된 센서를 통해 수집된 데이터를 의미한다. 예를 들면, 원유 파이프 내부에 설치된 자기장 센서를 통해 파악한 원유 수송 현황 등이다(표4, 표5 참조).

그림13. 미국 대체데이터 시장 성장 전망



자료: grandviewresearch.com

표4. S&P가 제공 중인 대체데이터 종류

데이터명	세부내용
Panjiva Supply Chain Intelligence	10 억개 이상의 출하 기록을 통해 1300 만 개의 기업간 관계 분석
Machine Readable Transcripts	컨퍼런스, 어닝콜 등 비정형화 기업 활동을 기계 학습이 가능한 형태로 제공
Preqin Alternative Asset Fund Performance	사모펀드, 헤지펀드 등의 자금 유입, 투자자관계, 자산배분 등의 데이터 제공
Real Estate Foot Traffic	REITs 가 보유한 부동산의 시간당 이용현황(소득, 성별, 인종 등)
Trucost Environmental	탄소배출, 대기/수질 오염, 석탄 사용 등 데이터 제공
S&P Global ESG Scores	2000 개 기업의 ESG 점수 제공(+기업의 미디어에서의 ESG 이슈에 대한 반응 분석)
BuildFax Permit Data	리모델링, 태양광패널 설치, 재건축, 건물 개보수 내역 등 제공
Ipqwery Patent & Trademark Data	특허, 특허 소유권 변동, 특허 소유 기업(스타트업 등) 정보 제공
FiscalNote US Legislation & Regulation	연방정부 법안과 규제를 분석하여 산업 또는 포트폴리오의 리스크 분석
Weather Source	우편주소를 기준으로 글로벌 기상 정보와 기상 예보 데이터 제공
Global Machine Readable Filings	글로벌 사업보고서 분석(이미지, 표 등은 제외)
Yewno Judicial Analytics	법원 문서에서 글로벌 기업과 관련된 정보 추출

자료: S&P, NH 투자증권 리서치본부

표5. Refinitiv가 제공 중인 대체데이터 종류

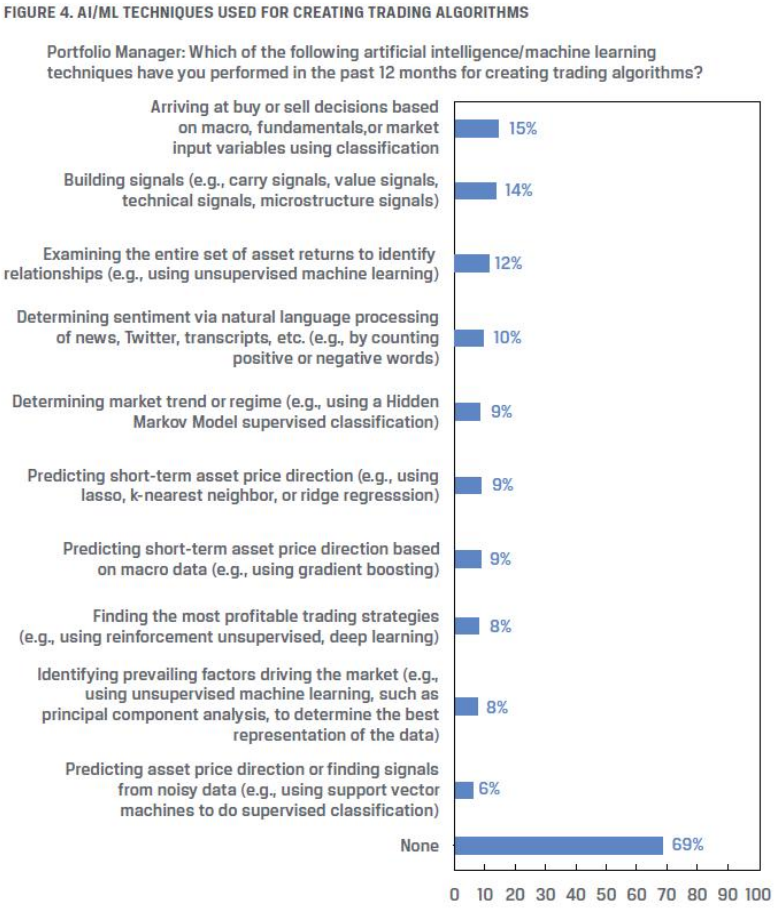
데이터명	세부내용
News Analytics	뉴스(영문, 일문)로부터 기업, 원자재 관련 감정점수 추출
Social Media Monitor	트위터, Stock Twits Meta Data 로부터 기업별 감정점수 추출
MarketPsych Analytics	뉴스와 소셜미디어로부터 추출한 감정점수를 사용하여 기업별 한달 후 주가 상승 가능성 제공
StarMine Text Mining Credit Risk Model	뉴스, 컨퍼런스콜, 사업보고서, 리서치보고서로부터 신용점수 추출
LinkUp	LinkUP 에서 제공하는 구인공고 데이터 제공
Satellites, Sensors & Models	농산물, 탄소, LNG, 금속, 기후, 선박 이동, 정유소 가동현황, 원유 저장고 현황, 원유 파이프라인 흐름, 농작물 현황 및 전망 등

자료: Refinitiv, NH 투자증권 리서치본부

3) 자산운용 분야

CFA Institute에 따르면 2019년 글로벌 자산운용역을 대상으로 한 설문에 응답한 230명 중 10%가 최근 12개월 동안 AI/머신러닝 기술 사용을 시도한 적이 있다고 답변하였다. **제작된 AI 모델은 방법론별로 상이하나, 크게 보면 ①시장 방향성 예측, ②매수 매도 타이밍 분석, ③가격을 주도하는 관계나 팩터 발견, ④소셜 미디어를 통한 감정 데이터 분석**으로 나눌 수 있다(그림14 참조).

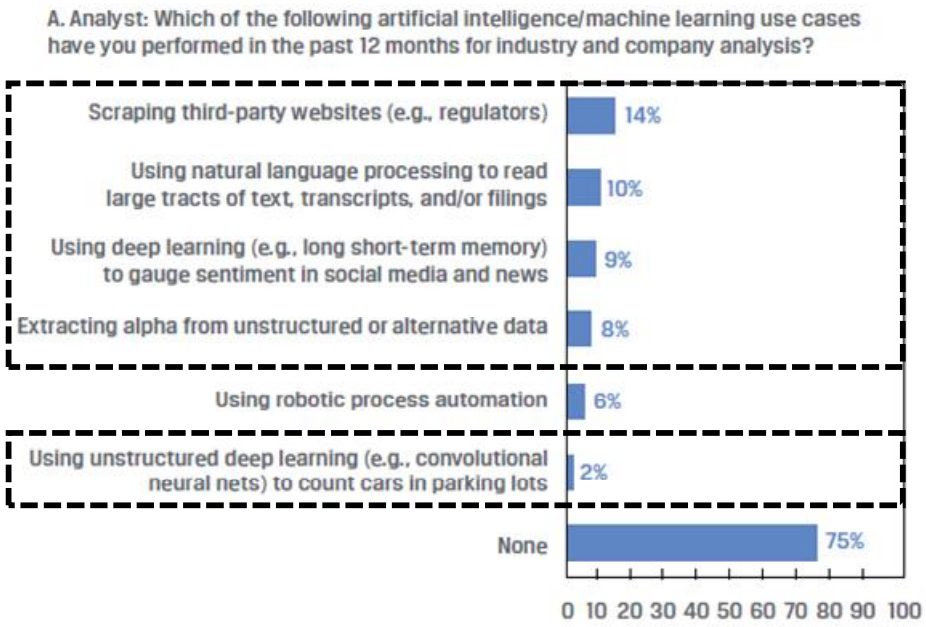
그림14. 글로벌 운용사에서 주로 제작한 트레이딩 AI 알고리즘



자료: CFA Institute(2019)., AI Pioneers in Investment Management

운용사들이 AI/머신러닝을 적용하여 산업이나 기업 분석한 경우에는 전통적인 데이터 대신 규제, 사업보고서, 소셜미디어 등 대체데이터를 사용하였다(그림 15 참조).

그림15. 운용사에서는 산업 및 기업 분석에 주로 텍스트 대체데이터 활용



자료: CFA Institute(2019)., AI Pionners in Investment Management, NH 투자증권 리서치본부

표6. 글로벌 운용사의 AI 도입 사례

기업명	학습목표	사용도구	입력 데이터
Man AHL	트레이딩과 주문 최적화	강화학습	Limit Order Book
New York Life Investments	투자 Signal 지표 생성 (Cycle, Value, Momentum, Sentiment)	-	Economic Indicators, Credit Quality, Leverage, Liquidity, Issuance Trend, Delinquencies 등등
State Street	거래량 예측	AutoEncoder + CNN	VIX, MOVE, S&P 500, Hang Seng, Historical Trading Volume, Regional Market Correlations, Specific Calendar Effect
Goldman Sachs	지역 시장 점유율 추정	NLP, Clustering, Sentiment Analysis	Analytics, Social Media, Satellite Imagery, Newspaper
American Century	어닝콜 기망 분석	NLP(Document Classification, Part of Speech Tagging, Entity Recognition, Topic Detection	Conference Call (Omission, Spin, Obfuscation, Blame 요소 추출)
China Life Asset	기업별 Credit Rate, Credit Report 자동화	NLP(BiLSTM+CNN)	Credit Risk Data + 10,000 Web sites

자료: CFA Institute(2019), AI Pionners in Investment Management, NH 투자증권 리서치본부

4) 머신러닝 기반 AI를 활용한 ETF

현재 미국에는 포트폴리오 구성 및 운용에 머신러닝 기반 AI를 사용하는 ETF가 약 16개 상장되어 있다. 이들은 크게 3가지 범주로 나눌 수 있다.

① **AI 모델형:** AI가 학습을 통하여 기업별 점수를 산출. 산출된 점수를 기준으로 포트폴리오 구성
② **대체데이터형:** 뉴스와 소셜미디어, 또는 리서치보고서의 텍스트를 분석하여 기업별 긍정/부정 점수 산출. 산출된 점수를 기준으로 포트폴리오 구성
③ **분류형:** 기업별 사업보고서, 실적 공시 등의 텍스트를 분석하여 사업내용이 유사한 기업끼리 군집화하여 섹터 구분

표7. 머신러닝 기반 AI를 활용한 ETF 목록

구분	브랜드	ETF 명	티커	설명
AI 모델형	ETFMG	AI Powered Equity ETF	AIEQ	AI가 경제, 재무, 뉴스 등 데이터를 분석하여 미국 기업별 점수 산출. 점수 기준으로 포트폴리오 분배
		AI Powered International Equity ETF	AIIQ	AI가 경제, 재무, 뉴스 등 데이터를 분석하여 미국 외 선진국 기업별 점수 산출. 점수 기준으로 포트폴리오 분배
	QRAFT	Qraft AI-Enhanced U.S. Large Cap Momentum ETF	AMOM	주식시장 데이터에서 레지듀얼 모멘텀 팩터 추출. 팩터 사용하여 미국 대형 기업별 향후 주가 상승 가능성 평가하여 포트폴리오 분배
		QRAFT AI-Enhanced U.S. High Dividend ETF	HDIV	주식시장 데이터에서 배당, 퀄리티 팩터 추출. 팩터에 따라 선별된 기업 중 AI를 통해 불확실성을 감안한 주가 상승 가능성을 기준으로 포트폴리오 분배
		QRAFT AI-Enhanced US Next Value ETF	NVQ	AI를 사용하여 무형자산 가치를 감안한 밸류 팩터 추출. 팩터 사용하여 미국 기업별 향후 주가 상승 가능성 평가하여 포트폴리오 분배
		QRAFT Ai-Enhanced U.S. Large Cap ETF	QRFT	매크로, 주식시장 데이터에서 팩터 추출. 팩터 사용하여 미국 대형 기업별 향후 주가 상승 가능성 평가하여 포트폴리오 분배
대체 데이터형	VanEckVectors	VanEck VectorsSocial Sentiment ETF	BUZZ	소셜미디어, 뉴스 등 온라인에서 가장 긍정적으로 평가된 75개 미국 대형주에 투자
	Advisor Shares	AdvisorShares Alpha DNA Equity Sentiment ETF	SENT	기업별 고객의 평가(소셜 미디어, 웹트래픽, 검색결과) 점수+재무점수 기준으로 기업 선정. 애널리스트 평가 기준 포트폴리오 분배
	SparkLine	Sparkline Intangible Value ETF	ITAN	무형자산 가치 반영 밸류 팩터 ETF. 브랜드, 인적자원, 지식자산, 네트워크효과 등을 자연어처리를 통하여 무형자산 측정
분류형	iShares	iShares Evolved US Consumer Staples ETF	IECS	BlackRock 분류 모델 기준 필수소비재 섹터로 군집된 기업에 투자
		iShares Evolved US Discretionary Spending ETF	IEDI	BlackRock 분류 모델 기준 경기소비재 섹터로 군집된 기업에 투자
		iShares Evolved US Financials ETF	IEFN	BlackRock 분류 모델 기준 금융 섹터로 군집된 기업에 투자
		iShares Evolved US Healthcare Staples ETF	IEHS	BlackRock 분류 모델 기준 필수 헬스케어 섹터로 군집된 기업에 투자
		iShares Evolved US Innovative Healthcare ETF	IEIH	BlackRock 분류 모델 기준 혁신 헬스케어로 군집된 기업에 투자
		iShares Evolved US Media and Entertainment ETF	IEME	BlackRock 분류 모델 기준 미디어/엔터 섹터로 군집된 기업에 투자
		iShares Evolved US Technology ETF	IETC	BlackRock 분류 모델 기준 IT 섹터로 군집된 기업에 투자

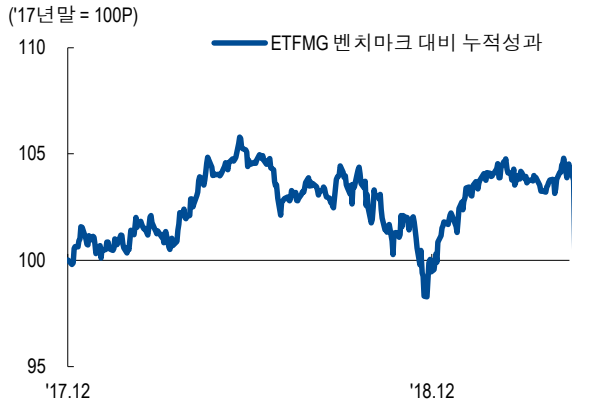
자료: NH투자증권 리서치본부, ETF별 투자설명서

머신러닝 기반 AI ETF는 2018년 ETFMG가 AIEQ, AIIQ 2종을 최초로 출시하였다. 출시 이후 6개월간 1억달러가 유입되며 큰 관심을 받았으나, 성과 부진으로 인하여 투자자들의 관심에서 멀어지며 자금이 유출되었다(그림16 참조). 지난해부터 성과가 개선되며 다시 자금이 유입되고 있으며 ETFMG의 AIEQ가 AI 모델형 ETF 중 가장 많은 자금을 운용하고 있다. QRAFT의 AI 모델형 ETF도 AMOM을 중심으로 빠르게 규모를 확대하고 있다(그림17 참조).

대체데이터형 ETF로는 SENT가 2021년 2월 처음 출시되었다. 이후 2021년 3월 상장한 BUZZ는 Reddit과 GameStop 사태을 이끌었던 Dave Portnoy의 마케팅에 힘입어 대규모 자금이 유입되었으나 부진한 성과를 기록하며 자금이 유출되는 추세이다(그림18 참조).

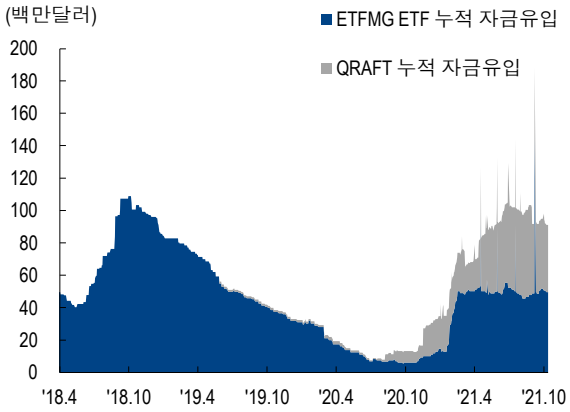
분류형 ETF는 2019년 iShares의 ETF가 최초로 상장하며 자금이 유입되기 시작하였다. 머신러닝을 통해 기업의 사업내용 변동을 빠르게 파악하여 보다 정확하게 섹터를 추종할 수 있어 상장 이후 자금이 꾸준히 유입되고 있다(그림19 참조).

그림16. ETFMG의 AI 모델형 ETF들은 2019년까지는 크게 아웃퍼폼하지 못함



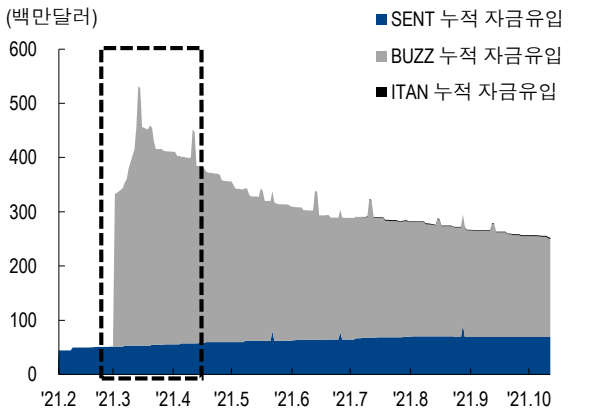
자료 FactSet, NH투자증권 리서치본부

그림17. AI 모델형 ETF 누적유입자산 추세



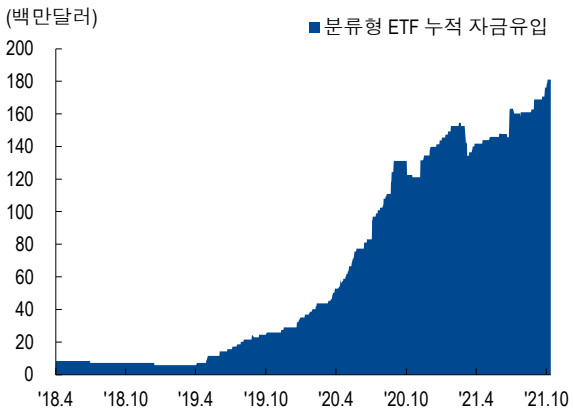
자료 FactSet, NH투자증권 리서치본부

그림18. 마케팅에 힘입어 대체데이터형 ETF에 자금 대량 유입



자료: FactSet, NH투자증권 리서치본부

그림19. 분류형 ETF에는 지속적인 자금 유입 추세



주: FactSet, NH투자증권 리서치본부

5) 기타 금융 관련 AI 서비스 기업

앞서 언급한 분야 외에도 리스크 관리, 트레이딩 툴, AI 모델 제작 서비스 등 다양한 핀테크 기업들이 금융 분야에서 AI를 활용하고 있다.

표8. 그 외 금융 관련 AI 서비스 기업

분류	기업명	세부내용
리스크 관리	TRADING TECHNOLOGIES	TT score 를 제공하여 트레이더의 개별 리스크를 관리 TT score: 트레이더의 거래 행위(알고리즘에 의하여 시리즈화)를 군집화하여 거래 행위의 리스크(시장 교란 가능성) 점수 측정
트레이딩	IMPERATIVE EXECUTION INC.	트레이딩툴 제공: AI 에 주식 호가 사이클을 학습시켜, 주가에 미치는 영향을 최소화하고, 평균 매매가가 설정된 호가에 최대한 근접하도록 도움
자산 수익률 예측	AUQUAN	자산 수익률 예측(변수 PCA 화, 장단기메모리 모델 사용, 손실함수 Mean-Variance Optimization) -> 2 년간 벤치마크대비 연율화 8% 초과 수익률 기록
	KAVOUT	펀더멘털, 가격, 거래량, 패턴 등의 정보를 머신러닝 모델을 통해 다음달 주식이 아웃퍼폼할 확률 계산
AI 모델 제작 서비스	EQUBOT	1. 사용자가 원하는 목적에 맞게 뉴스 데이터, 경제 데이터, 이벤트 데이터, 시장감정 데이터 등을 취합하여 머신러닝을 통한 최적화 모델 제작 2. AI Powered Equity Index 제공 중(AIEQ, AIIQ ETF 의 추종 지수) - EquBot 이 선택한 주식으로 구성한 지수. 1999 년부터의 데이터를 기준으로 딥러닝 모델 학습. 모델은 1 개월 후의 가격을 예측하는 것이 목적
AI 플랫폼	ACCERN	사용자가 분석하기 원하는 기업 관련 문서를 업로드하면, 문서에서 기업 정보, 이벤트 정보 등을 분석하여 데이터화 자료로 출력
	DATA ROBOT	사용자가 데이터를 플랫폼에 입력하고, 원하는 머신러닝 모델을 선택하여 학습 가능. 학습결과도 사용자 친화적으로 전환(그래프, 차트 등)하여 출력

자료: 각사 홈페이지, NH투자증권 리서치본부

AI의 한계는 데이터 의존성

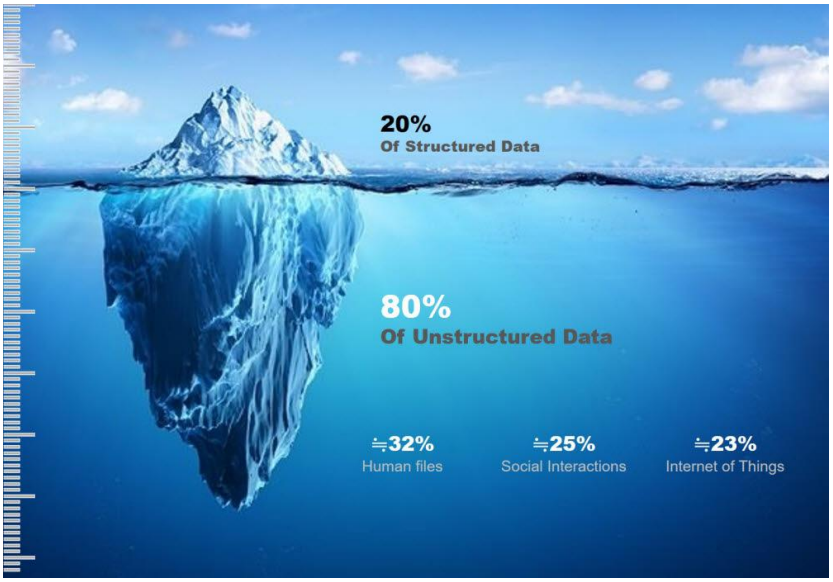
6) 금융산업 내 AI 활용의 한계점

금융 분야에서 머신러닝 관련 언급이 활발해지기 시작한 때는 2017년으로 아직은 초기 단계다. 로보 어드바이저도 머신러닝을 응용하고 있으나 계좌별 리밸런싱 자동화에 초점이 맞춰져있다. 일부 로보 어드바이저에서 포트폴리오 종목 선정 및 비중 계산에 머신러닝을 사용하고 있기는 하나, 대부분 전통적인 포트폴리오 분배 방법론을 따르고 있다.[주)]

자산운용 분야에서는 아직 AI의 성과가 부각되지 않았으며, AI의 운용역 대체에 대해서도 부정적인 입장을 견지하고 있다. 이는 AI의 데이터 의존성에 기인한다. AI의 학습은 철저히 데이터에 기반하므로 데이터 의존적이며, 데이터 외 환경의 학습은 어렵다. 시장에 영향을 주는 비정형 데이터가 많고, 데이터별 영향력을 명확히 구분하기 어려운 금융시장의 특성은 AI의 데이터 의존성 문제를 부각시킨다.

주) Beketov et al(2018)의 조사에 따르면 글로벌 로보 어드바이저사 중 대다수(약 219개)가 Markowitz의 포트폴리오 이론에 따른 자산배분을 수행하였다.

그림20. 전체시장에서 정형화된 데이터는 20%에 불과



자료: 전자신문

공개된 대체데이터에서는 알파가 소멸

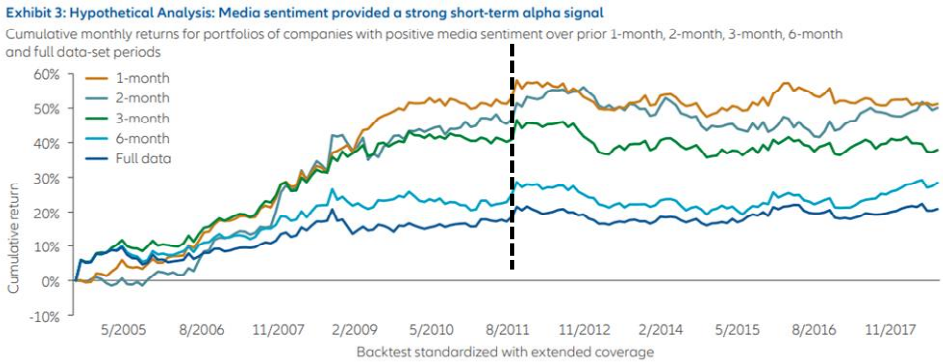
또 다른 문제는 대체데이터와 이를 이용한 AI 모델이 이미 알파를 상실했을 가능성이 있다는 점이다. Allianz Global Investors(2019)는 2004년부터 2018년까지 뉴스를 통해 분석한 S&P 500 기업별 감정 점수를 기반으로 투자한 포트폴리오의 성과를 분석하였다. 그 결과 2011년까지는 좋은 성과를 보였으나 이후로는 초과성과를 얻지 못하였다. Allianz는 2007년 이후 Reuters와 Dow Jones가 **뉴스 데이터를 대체데이터로 가공하여 대중에게 제공하면서 알파가 소멸**했을 것으로 추정하였다(다음 페이지 그림21 참조).

알파가 소멸된 데이터에서 AI가 추출할 수 있는 수익률은 제한적

AI를 활용한 미국 상장 ETF에서도 유사한 현상을 확인할 수 있다. 미국 대형주를 동일한 Universe로 삼는 ETFMG[주1)]의 AIEQ, Qraft의 AMOM, QRFT는 서로 다른 방법론으로 벤치마크 없이 액티브하게 운용된다. 그럼에도 성과는 0.96~0.99의 높은 상관계수를 보인다. 두 운용사 모두 매크로, 재무데이터, 뉴스 등 **대중에 공개된 데이터를 사용한다는 점에서 AI가 학습을 통해 데이터로부터 추출할 수 있는 최대 알파가 제한적**이며, 이로 인하여 유사한 성과를 기록하는 것으로 판단된다(다음 페이지 그림22 참조).

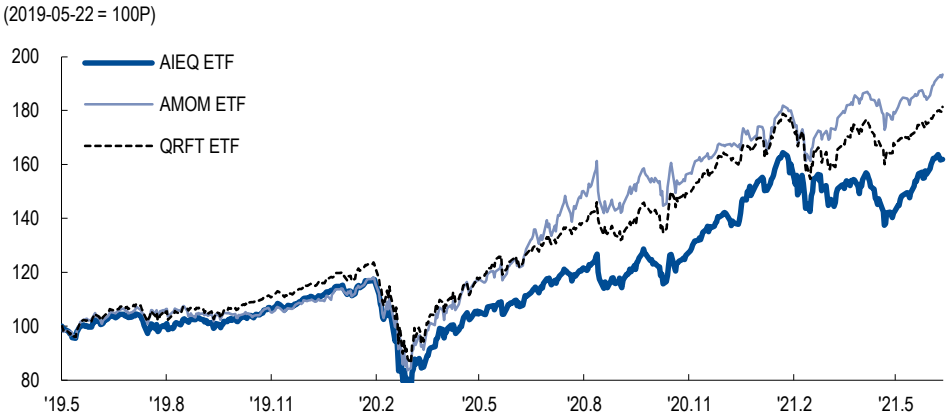
주1) ETFMG는 재무, 뉴스, 경영구조, 경제 점수를 통해 투자 종목을 선정하는 반면, QRAFT는 ETF별 전략팩터에 기반하여 투자 종목을 선정한다.

그림21. 2011년부터 뉴스 데이터의 알파가 소멸



자료: Michael Heldmann et al., "Artificial Intelligence in portfolio management: a tool, not a destination" Allianz Global Investors, 2019

그림22. AI Model형 ETF들의 성과가 유사한 추세를 보임



자료: FactSet, NH투자증권 리서치본부

IV. AI와 대체데이터를 활용한 KOSPI 예측 모델

앞서 언급한 한계에도 차별화된 데이터에 기반한 AI 활용 시 알파를 창출할 가능성이 있다. 본 보고서에서는 AI와 대체데이터를 활용한 KOSPI 예측 모델 구현을 통해 한국 주식시장에서 AI 전략의 유용성을 점검해보고자 한다.

1. 모델의 목표 및 KOSPI 전처리

모델의 목표는 주간 KOSPI 지수 방향 예측

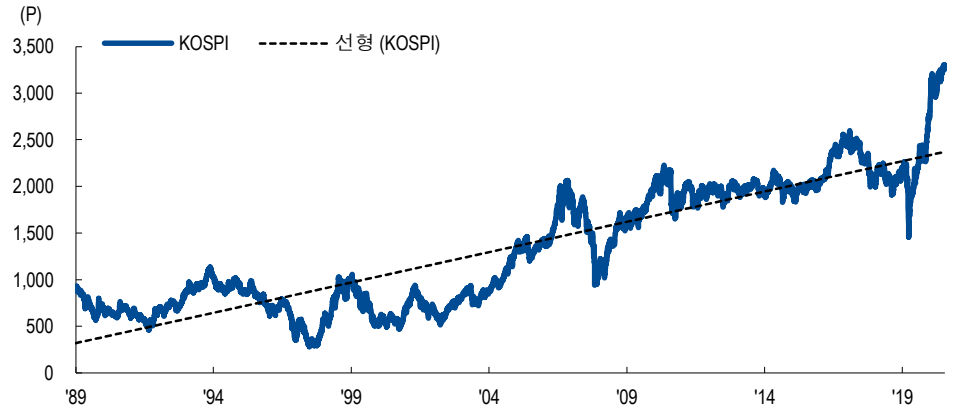
모델의 목표는 KOSPI의 매수/매도 시그널(Market Timing) 포착으로 설정하였다.

이 때, 예측해야할 지수로는 **원시계열 KOSPI가 아닌 장기 선형 추세 제거 진폭 조정 KOSPI (Adjusted Detrended KOSPI)를 설정**하였다.

원시계열 KOSPI는 GDP 상승, 인플레이션 등에 따라 장기적으로 우상향하는 특징(다음 페이지 그림23 참조)을 보이기 때문에 매수 시그널 편향이 발생할 수 있다. 또한 회귀분석은 대상 시계열의 정상성(Stationarity)[주1] 을 가정하므로 모델이 목적대로 작동하지 못할 가능성이 높다. 이에 0값을 기준으로 진동하는 조정 KOSPI(다음 페이지 그림24 참조)를 학습하는 것이 보다 적합하다고 판단하였다(자세한 방법론은 Appendix D 참조).

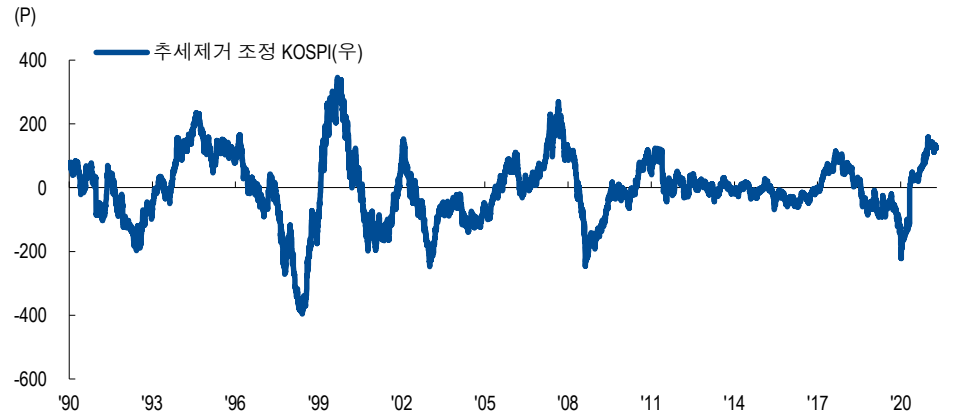
주1) 정상성(Stationarity): 시계열이 일정한 분산을 가지며, 특정 값으로 회귀하는 특성을 지닐 경우 정상성을 지닌다고 평가한다. 시간에 따라 시계열이 증가하거나, 분산이 변하는 등 특성이 변화하는 경우 비정상성을 지닌 시계열이라 부른다. 랜덤워크 이론에 따르면, 금융시계열의 미래값은 전기값을 기준으로 임의적으로 움직이므로 특정값으로 회귀하는 성질을 지니지 않은 비정상성 시계열이다.
주2) 랜덤워크(Random Walk): 금융시계열은 과거의 모든 정보가 반영되어 결정된다. 따라서 미래의 금융시계열은 오로지 미래의 정보에 의해서만 결정된다. 미래의 정보는 임의적으로 결정되므로 미래의 금융시계열은 임의값으로 볼 수 있다는 이론

그림23. KOSPI는 우상향하는 장기 선형 추세를 지님



자료: NH투자증권 리서치본부

그림24. 모델의 학습 목표인 조정 KOSPI



자료: NH투자증권 리서치본부

2. 모델의 입력 변수

1) 3가지 유형의 입력 변수

경제, 펀더멘털, 감성지표를 모델의 변수로 입력

모델을 학습시키기 위해 총 3가지 유형의 입력 변수를 사용했다. 우선 ①경제지표를 사용하였으며, 주가와 수익성, 배당과 같은 ②펀더멘털지표, 그리고 당사에서 발간된 주간 시장 전망 보고서에서 추출한 ③감성지표를 사용했다.
경제지표로는 통계청에서 공표하는 경기종합지수(선행, 동행, 후행) 17개를 사용하였다. 이 중 CP 유통금리는 한국실질금리(기준금리-CPI)로 변경하였다. 여기에 글로벌 경기 반영을 위한 미국 장단기 금리차 지표와, 통화정책과 주식시장 과열 정도를 판단하기 위한 시가총액/M2 통화량 지표를 더하여 총 19개 지표를 사용하였다.

펀더멘털지표로는 주식시장의 밸류에이션과 주가 움직임을 반영하는 변동성, 배당, 이익, 자산 등 총 8개 지표를 선정하였다.

감성지표는 당사에서 매주 발간하는 주간 시장 전망 보고서의 텍스트에서 추출한 긍정, 부정, 즐거움, 공포 등 10개 대체데이터로 구성하였다.

결과적으로 선정된 입력변수는 경제지표 19개, 펀더멘털지표 8개, 감성지표 10개와 조정 KOSPI로 총 38개이다(다음 페이지 표9~11 참조).

표9. 경제지표 19개

경제지표	세부설명
USTsprd	미국장단기금리차
KRTsprd	한국장단기금리차
RealRate	한국실질금리
InvCirc	재고순환지표
EconSent	경제심리지수
ProducersMachineryShipment	기계류내수출하지수(선박제외) %y-y
ConstructionOrders	건설수주액(실질) %y-y
TermsOfTrade	수출입물가비율
IndProd	광공업생산지수 %y-y
Service	서비스업생산지수(도소매업제외) %y-y
ConstructionCompleted	건설기성액(실질) %y-y
RetailSale	소매판매액지수 %y-y
Shipment	내수출하지수 %y-y
Imports	수입액(실질) %y-y
ServiceEmployed	비농림어업취업자수 %y-y
ProducersInven	생산자제품재고지수 %y-y
ImportConsumerGoods	소비재수입액(실질) %y-y
Employed	취업자수 %y-y
MVperM2	시가총액/M2 통화량

자료: NH투자증권 리서치본부

표10. 펀더멘털 지표 8개

펀더멘털지표	세부설명
K200C	KOSPI 200 콜옵션 프리미엄
R_Vol	KOSPI 실현변동성
ROE	자기자본이익률(12m fwd)
Payout	배당성향(12m fwd)
EP	당기순이익(12m fwd)/시가총액
5yEP	5 년평균 당기순이익(12m fwd)/시가총액
DP	배당수익률(12m fwd)
BP	순자산(12m fwd)/시가총액

주: Indicator 선정은 Miguel A. Ferreira, Pedro Santa-Clarara, NBER Working Paper(2008) 참고
자료: NH투자증권 리서치본부

표11. 감성지표 10개

부정적 감성	긍정적 감성
Negative	Positive
Fear	Joy
Sadness	Anticipation
Anger	Trust
Disgust	Surprise

자료: NH투자증권 리서치본부

2) 대체데이터: 감성점수

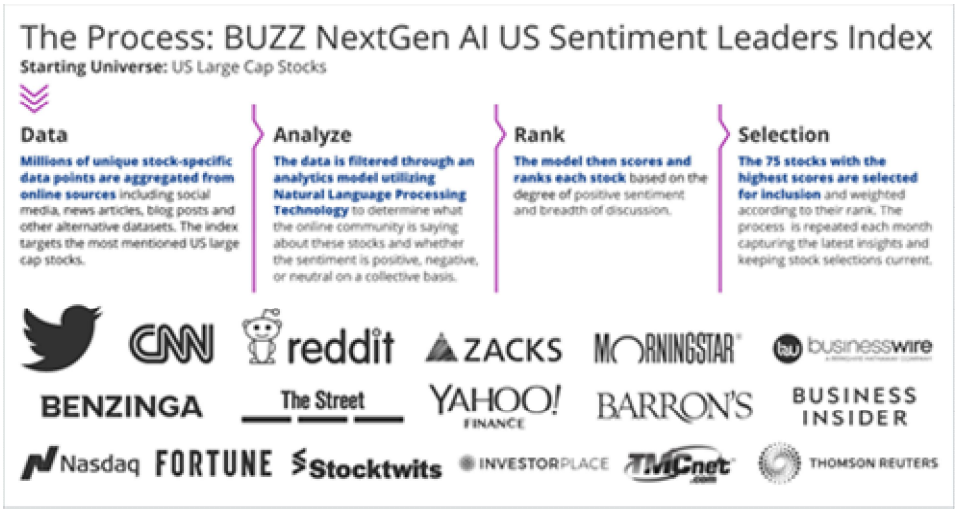
감성점수는 당사 주간 시장 전망 보고서에서 추출

앞서 언급한 바와 같이 대중에게 공개된 데이터는 알파가 고갈되었을 가능성이 높으므로 대체데이터를 활용해 알파를 찾을 필요가 있다.
Alianz Global Investors의 사례로 알 수 있듯이 대체데이터 역시 대중에게 공개되었다면 알파가 소멸되었을 가능성이 높다. 소셜 미디어를 활용해 종목을 선정하는 대체데이터형 ETF들(BUZZ, SENT)도 알파를 얻지 못하고 부진한 성과를 기록하였는데, 소셜 미디어가 갖는 비대표성[주]의 문제와 정보 후행성이 알파 소멸의 원인으로 판단된다.

당사 주간 시장 전망 보고서는 발간 목적이 국내 주식시장 전망으로 명확하고, 전문가의 의견이 반영되며, 발간 즉시 모델에 반영할 수 있다는 점에서 대체데이터로써 알파가 존재할 것으로 판단하였다.

주) 소셜 미디어의 비대표성: Garimella et al. (2018).에 따르면 소셜 미디어는 일반 대중의 의견을 반영한다 볼 수 없다. 또한 특정 여론에 치우쳐져있을 가능성이 높다. 트위터는 대립되는 견해를 가진 집단 간에 게시물이 고립되어 의견이 증폭, 고착화되는 Echo Chamber 현상이 존재한다.

그림25. 소셜 미디어 데이터를 활용해 종목을 선정하는 대체데이터형 ETF BUZZ



자료: VenEck Advisors, NH투자증권 리서치본부

텍스트 기반의 감성분석 방법론은 사전 기반과 머신러닝 기반으로 구분된다. 사전 기반은 단어별로 감성점수를 미리 분류해놓은 사전을 사용하는 방식으로 적용에 간편하다는 장점이 있는 반면, 동일 단어일지라도 문맥에 따라 감성점수가 변할 수 있다는 점을 반영하지 못한다.

머신러닝 기반 방법론은 시계열 처리 알고리즘을 사용하여 문장의 맥락을 고려한 감성점수 측정이 가능하다. 단점은 적합한 데이터셋을 사용한 학습이 요구되며 이로 인하여 사전 기반 대비 많은 자원을 요구한다.

본 보고서에서는 시중에 제공되는 **공용 감성어 사전(Lexicon)인 SentiWordNet과 감정분류 기능을 제공하는 NRC 감성사전 2종을 활용해 사전 기반 방법론을 적용하였다(표12 참조).**

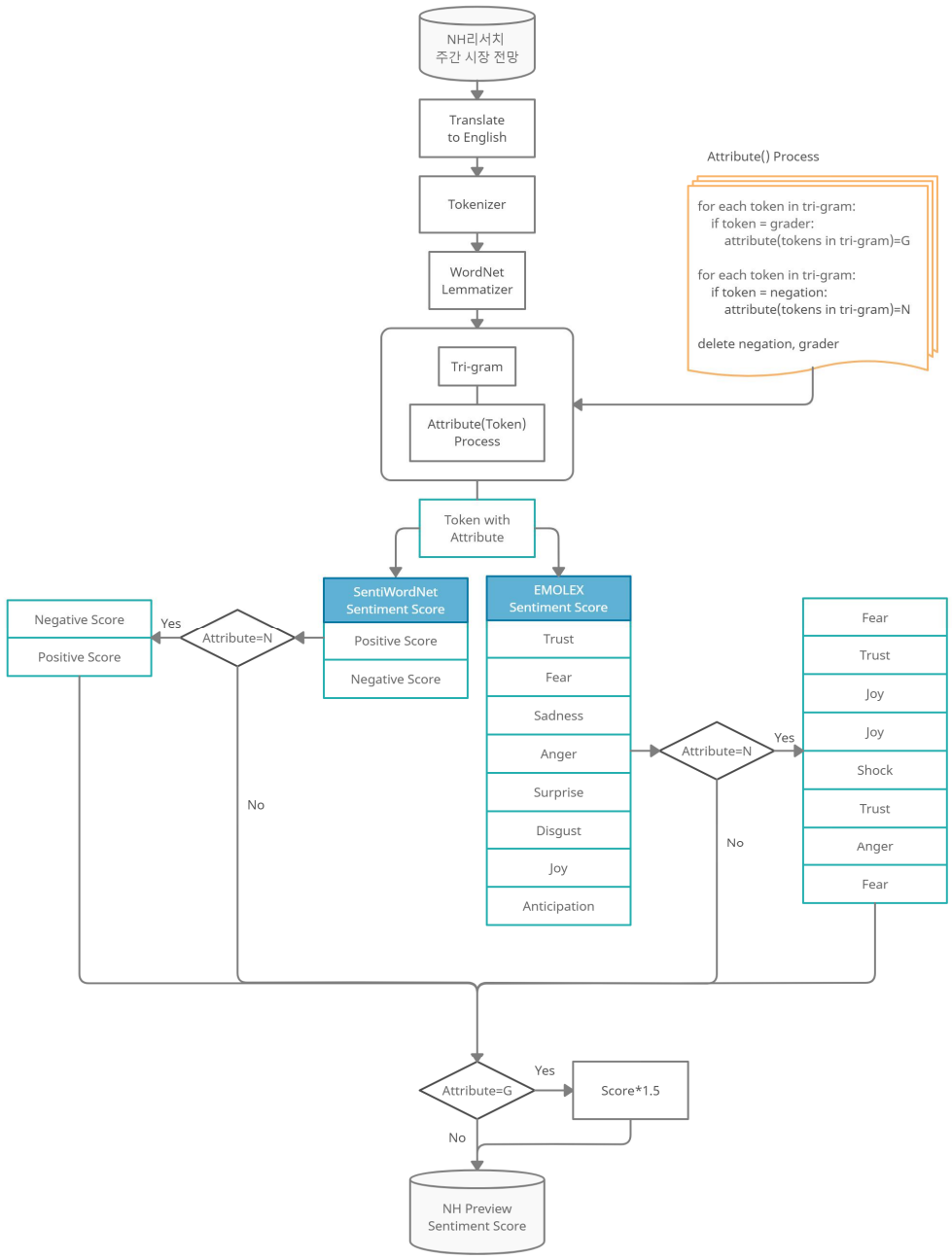
표12. 공용 감성어 사전 목록

목록	세부내용
Bing Liu's Lexicon	– 2014 년까지 업데이트 – Polarity(Positive, Negative) 분류 기능
AFINN Lexicon	– Finn Arup Nielson 이 2009~2011 년 수집한 Lexicon – Polarity+Intensity 분류 기능 제공
TextBlob Lexicon	– 2020 년 4 월까지 Change Log 존재 – Positive, Negative, Neutral(Polarity)+Intensity 와 주관성(Subjectivity) 분류 기능 – 품사 분류 기능 – Tokenization, Lemmatizaion, Stem 등 단어 처리 기능 제공 – WordNet 통합 기능제공
NRC Lexicon (EMOLEX)	– 제작자가 2021 년 3 월까지 논문 활동 지속 중 – 인터넷 투표 합산을 통한 감정 점수 측정 – Polarity 를 포함하여 감정(분노, 기대, 혐오, 공포, 즐거움, 슬픔, 놀람, 신뢰) 분류 제공
SentiWordNet Lexicon	– 2010 년에 목표치 11 만개 영어단어 정리 완료 – Polarity+Intensity 분류 기능 제공 – WordNet 에서 정의된 SysNet 을 통해 단어의 긍정 부정 점수를 전파 – 전파는 Random Surfer Model 사용 – *WordNet 은 1985 년 프린스턴 대학에서 시작된 영어 의미 어휘목록 구축 프로젝트. WordNet 은 현재 GWA(Global Wordnet Association)에 의하여 타 언어와의 연결화 진행. 단어와 단어간에 관계를 정의하여 Sysnet 을 만들었으며, 유의어와 반의어 등을 분류 가능
KNU Lexicon	– 2018 년 군산대학교 연구팀에서 제작 – 약 1 만 4 천개 단어의 감성어 보유 – Polarity 분류 기능 제공 – 평가자들의 투표를 통해 감정 점수 측정

자료: 각 감성어사전별 홈페이지, NH투자증권 리서치본부

감성어 사전은 영어를 기반으로 하므로 주간 시장 전망 보고서를 사전에 영어로 번역하는 과정을 추가하였다. 감성점수를 추출하기 위해 모든 단어는 품사를 감안하여 근원 단어로 변경시켰으며, 문장별로 very, really와 같은 부사나 not, wasn't와 같은 부정어가 있는지 파악하여 감성점수에 반영하였다. 전체 텍스트의 최종 감성점수는 (감지된 단어별 감성점수의 합 / 감지된 단어의 개수)로 표준화시켰다.

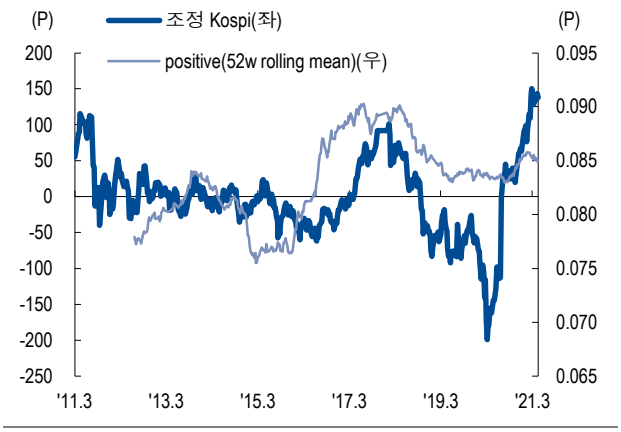
그림26. NH 주간 시장 전망 보고서 감성점수 추출 알고리즘 순서도



주: https://github.com/kyoujinkim/NH_KOSPI_Model/blob/main/SentimentExtractor.py 코드 참조
자료: NH투자증권 리서치본부

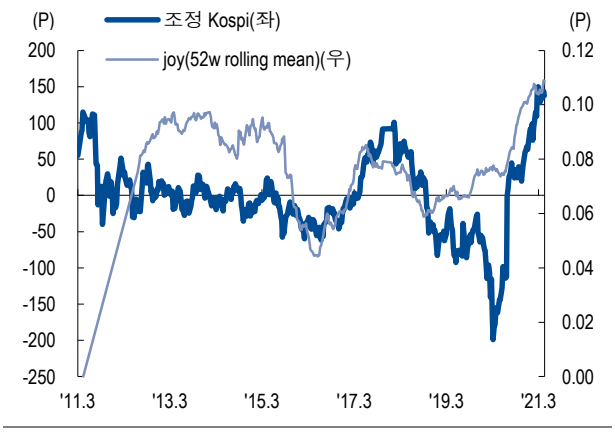
추출한 감성점수는 조정 KOSPI와 높은 상관관계를 보였다. 특히, Positive, Joy, Disgust, Anger 점수가 조정 KOSPI와 높은 상관계수를 기록하는 바, 해당 요인이 KOSPI 방향 예측에 유의미할 것으로 판단된다(그림27~30, 표13 참조).

그림27. 조정 KOSPI 지수와 Positive 점수



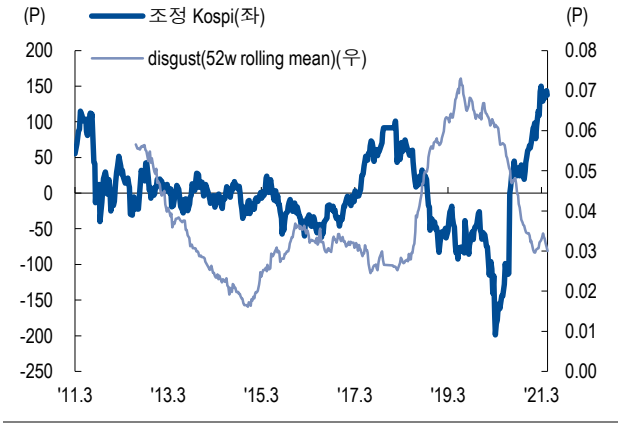
자료: NH투자증권 리서치본부

그림28. 조정 KOSPI 지수와 Joy 점수



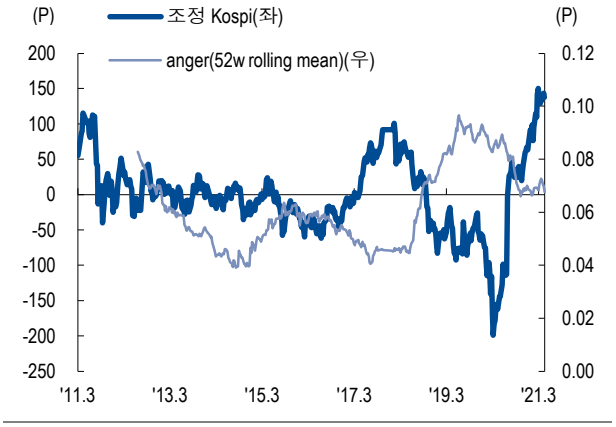
주: NH투자증권 리서치본부

그림29. 조정 KOSPI 지수와 Disgust 점수



자료: NH투자증권 리서치본부

그림30. 조정 KOSPI 지수와 Anger 점수



주: NH투자증권 리서치본부

표13. 감성점수(52w rolling mean)와 KOSPI 지표 간 상관계수표

구분	Negative (-)	Fear (-)	Sadness (-)	Anger (-)	Disgust (-)	Surprise (+)	Positive (+)	Joy (+)	Anticipation (+)	Trust (+)	Positive-Negative
Kospi	-0.09	**-0.28**	**-0.28**	0.02	-0.11	0.01	**0.51**	**0.28**	-0.03	0.02	**0.36**
조정 Kospi (추세제거)	0.02	-0.01	-0.1	**-0.42**	**-0.47**	-0.11	**0.29**	**0.44**	0.19	-0.03	0.14

자료: NH투자증권 리서치본부

3. 입력 변수 전처리

변수가 지나치게 많은 경우, 컴퓨터 역시도 학습에 어려움(머신러닝 시간 장기화, 과적합 문제, 차원의 저주 등)을 겪는다. 또한 경제지표의 경우, 전처리를 통해 변수간 상관관계를 충분히 감소시켰으나 경제지표 외 지표 간에도 의도치 않은 상관관계가 존재한다면 모델의 비중값이 왜곡되어 모델의 성능에 악영향을 미칠 수 있다.

이에 따라 **주성분 분석(PCA: Principal Component Analysis)[주] 을 통하여 변수의 개수를 최대한 감소시키고, 변수간의 상관관계를 제거**하였다. 기대 효과는 다음과 같다.

① 변수 차원의 축소(과적합 가능성 감소, 학습소요시간 감소)
② 노이즈 제거(전체 분산의 90%를 설명하는 주성분을 사용. 그외 주성분은 제거)
③ 입력변수 간 상관관계 제거(주성분의 축은 서로 직각을 이루며 교차, 즉 직교하므로 독립)

시간이 경과함에 따라 데이터의 성질과 상관관계가 변하므로 주성분 역시 변동될 수 있어 매달 마지막 금요일, 모델 학습기마다 주성분을 다시 측정하여 업데이트하였다.

주) PCA(Principal Component Analysis): 주성분 분석. 좌표 상에서 변수들을 가장 잘 설명할 수 있는 방향을 찾는다. 찾아낸 방향을 주성분 벡터라고 한다. 변수들을 주성분 벡터로 표현했을 때의 분산이 클수록 변수들을 가장 잘 설명하는 주성분이라 표현한다.

4. AI와 대체데이터를 활용한 KOSPI 예측 모델

1) 모델 설정

모델의 예측 정확도보다는 수익률 최대화에 초점

앞서 언급한 바와 같이 3가지 유형의 총 38개 입력 변수를 사용하여 KOSPI 상승/하락 시그널 도출을 위한 머신러닝 모델을 제작했다.

기간에 따라 입력해야하는 변수의 개수가 많아지기 때문에[주1] 다층의 딥러닝(Deep Learning) 모델을 사용해 학습을 진행했다. 출력 시그널의 종류는 [상승(Bull), 중립(Neutral), 하락(Bear)]으로 설정하였다.

학습방법으로는 강화학습을 사용하였다. 강화학습에 사용될 보상함수로는 포트폴리오의 변동성 대비 수익률을 최대화하는 함수를 사용하였다[주2]. **포트폴리오 변동성을 줄여 기간별 안정적인 수익률을 얻도록 보상함수는 샤프지수와 유사하게 설계**하였다.

$$p_t = \text{조정 } KOSPI_t$$
$$Reward_t = p_{t+5} - p_t$$
$$Gain_t = \frac{Reward_t}{STD\left(Series\ of\ Rewards(0\ to\ T)\right)}$$

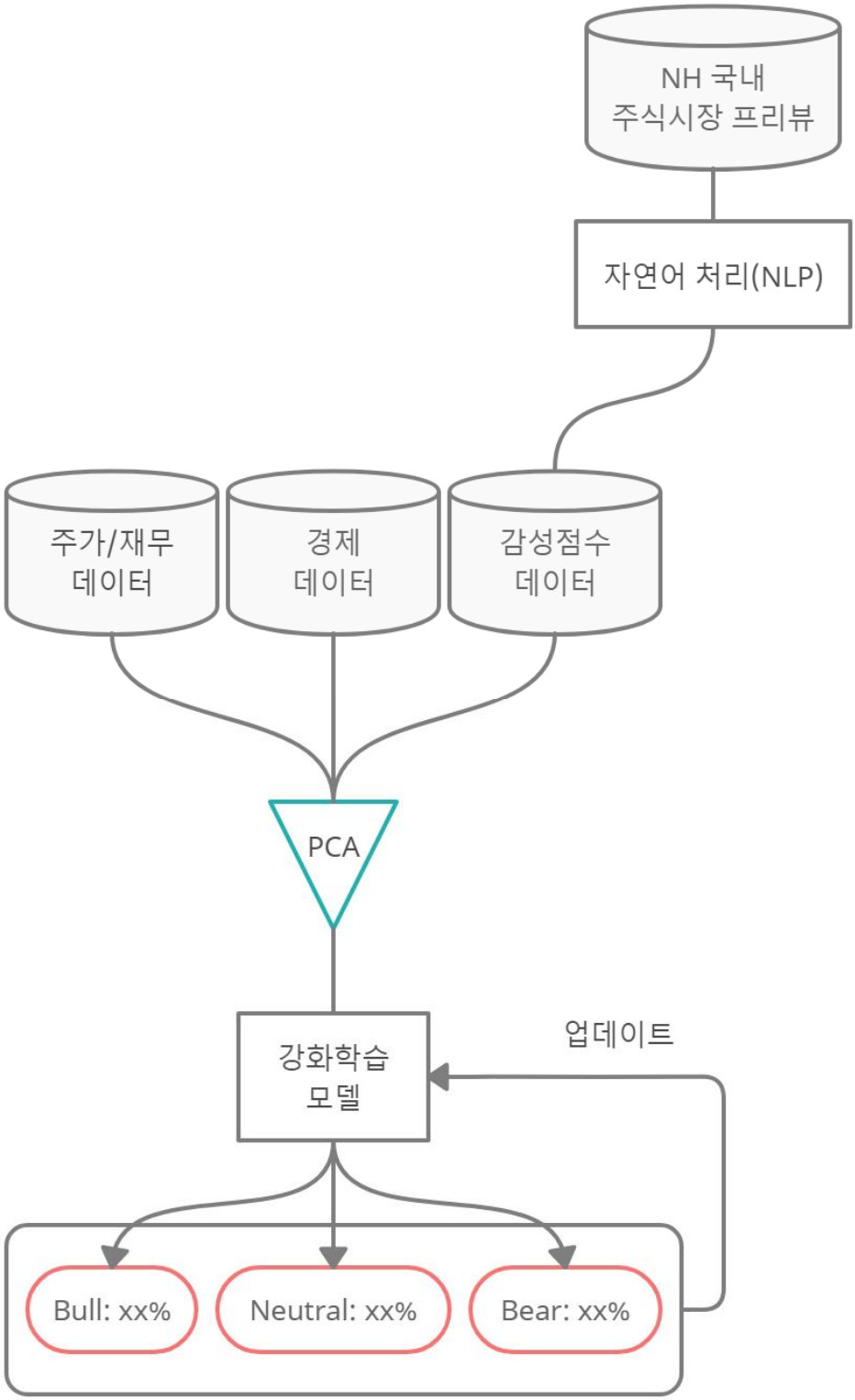
$* t$: 현시점, T: 총학습기간

주1) 입력변수가 길이가 1일인 경우에는 1(# of days) * 15(# of variables) = 15이나, 6개월인 경우에는 21*6(# of days) * 15(# of variables) = 1890개의 값을 입력해야한다.

주2) 변동성 대비 수익률 최대화를 통해 기대하는 효과는 다음과 같다.

1. 예측의 정확도보다는 포트폴리오의 수익률을 기준으로 행동을 선택하므로 변동성이 높은 시기가 전체 기간 대비 짧더라도 희석되지 않을 것(학습 Gradient 기준)
2. 단일 지수의 상승, 하락을 맞추는 것을 기준으로 변동성 대비 수익률을 최대화할 경우, 예측의 정확도를 상승시키는 것과 비슷한 효과를 지님. KOSPI 기준으로 예측률이 50%일 때, 예측률이 100%일 때보다 표준편차가 1.5배 높음
3. 변동성 대비 수익률을 최대화하는 함수를 사용하므로 KOSPI가 지속적으로 상승하는 시기에는 매매회전율이 낮아지고, KOSPI의 변동성이 증가하는 시기에는 매매회전율이 높아짐. 변동성이 높은 시기에 보다 적극적으로 대응

그림31. AI와 대체데이터를 활용한 KOSPI 예측 모델 알고리즘 순서도

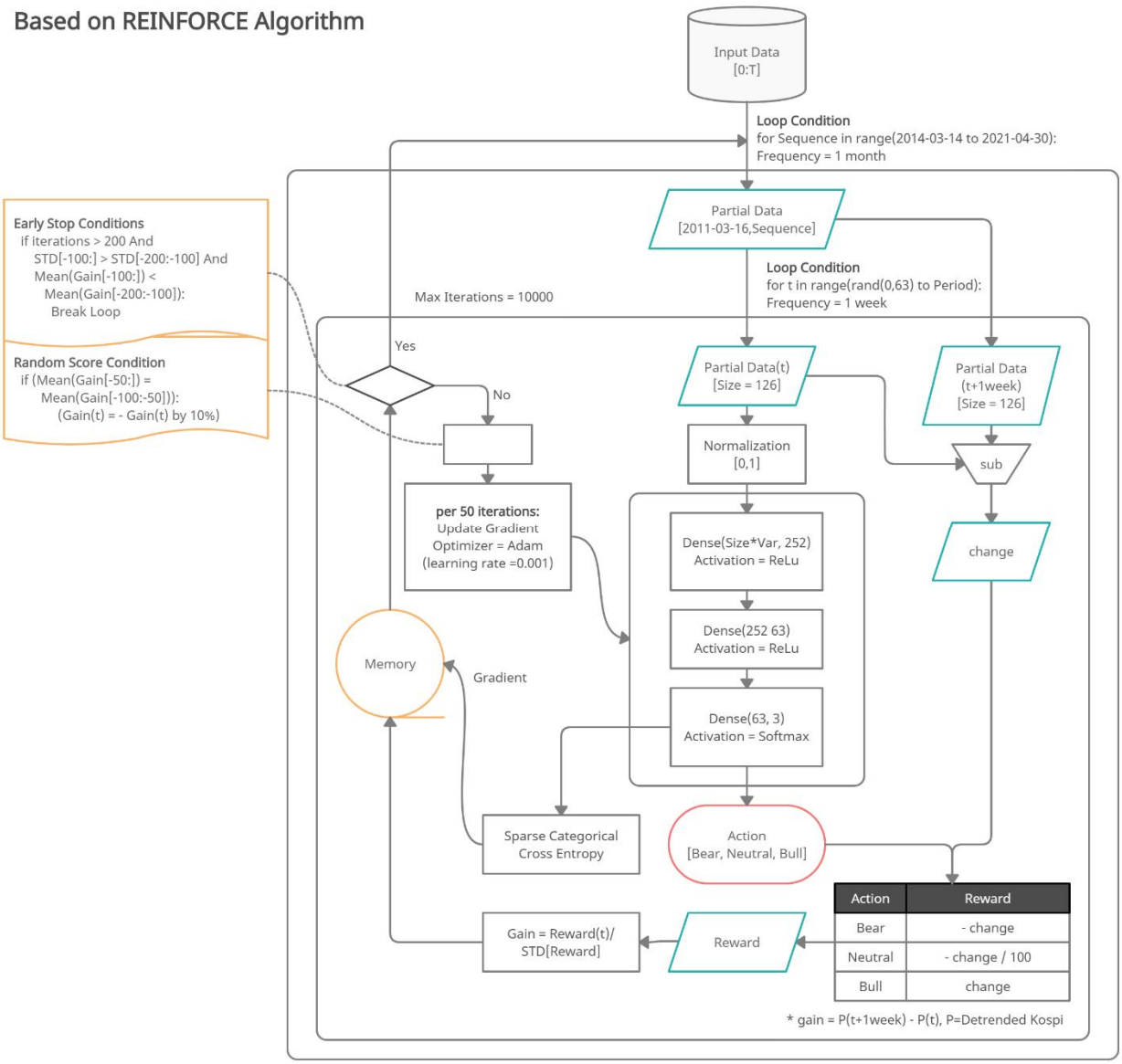


자료: NH투자증권 리서치본부

지도학습 알고리즘을 사용하여 모델의 예측 정확도를 극대화할 수 있게 설계할 수도 있으나, 이 경우 수익률 극대화에는 실패할 수 있다. 주식시장에는 낮은 변동성을 보이는 대부분의 기간과 짧지만 매우 높은 변동성으로 성과에 큰 영향을 미치는 시기가 존재한다. 정확도 극대화 모델은 변동성이 낮은 시기를 위주로 학습하기 때문에 변동이 높은 시기에 대응하는 능력이 떨어진다.

Estrada(2008)에 따르면 15개의 국가별 대표 주가 지수에 평균 40년간 투자했을 때, 각 지수별로 가장 수익률이 안 좋았던 10일을 제외하면 평균 성과가 롱 포트폴리오 성과의 150.4%를 달성하였다. 반면, 수익률이 가장 좋았던 10일을 제외하면 롱 포트폴리오 성과의 절반 수준을 기록하는 것에 그쳤다. 이는 변동성이 높은 시기 대응력의 중요성을 시사한다. **강화학습은 수익률 값이 모델 학습에 영향을 주므로 지도학습 대비 변동성 상승기의 학습 비중이 높아져 변동성이 높은 시기에 대한 대응력을 갖출 수 있다.**

그림32. 강화학습 알고리즘의 순서도



주1: 강화학습 모델 최적화 과정은 Appendix E 참조
주2: https://github.com/kyoujinkim/NH_KOSPI_Model/blob/main/REINFORCE.py 코드 참조
자료: NH투자증권 리서치본부

2) 최종 모델 출력: 앙상블

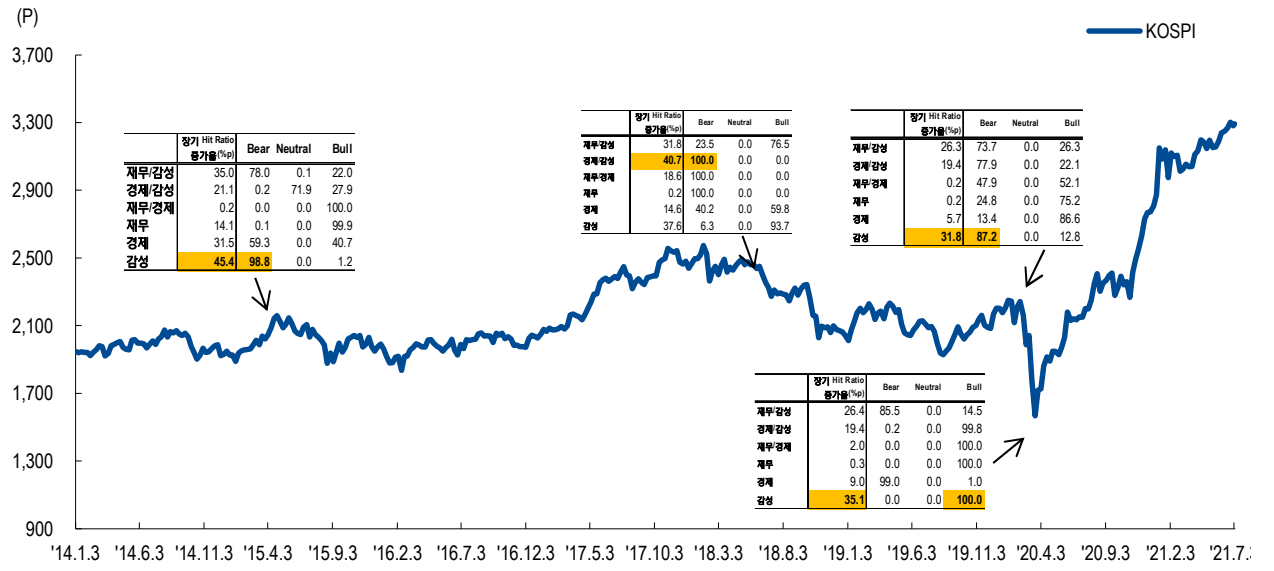
모델의 최종적인 출력은 감성+펀더멘털, 감성+경제, 펀더멘털+경제, 감성, 펀더멘털, 경제 총 6가지 데이터 집합으로 학습시킨 각 모델의 결과값을 종합하여 결정한다. 이를 앙상블 기법이라 한다. 이를 통해 모든 데이터를 입력했을 때 발생할 수 있는 과적합, 또는 차원의 저주 문제를 일부분 해소할 수 있다.

앙상블 기법은 모든 모델 중에서 각 시기별로 최적인 모델을 선택하는 방식을 택한다. 최적 모델의 기준은 '최근 3년간 모델의 예측률이 그 이전과 비교하여 얼마나 상승했는지'이다.

2015년 6월 15일 중국 주식시장이 급락한 시기에 최근 3년간 예측률이 가장 크게 상승한 모델은 감성 데이터 모델이었다. 이 때 감성 데이터 모델이 하락에 가장 높은 확률을 부여했으므로 매도 포지션이 선택되었다. 한편, 2018년 6월 15일 미국이 500억 달러 규모의 중국 수입품에 대하여 관세를 25%를 부과하였을 때 장기 예측률 증가율이 가장 높았던 모델은 경제/감성 데이터 모델이었다. 이 때 모델은 하락에 가장 높은 확률을 부여했으므로 매도 포지션이 선택되었다.

모델은 코로나 시기에도 좋은 대응력을 보여주었다. 코로나19의 초입기인 2020년 2월 28일에 선택된 감성 모델은 매도 포지션을 선택하였고, 반등이 시작된 3월 20일에는 감성 모델에 의하여 매수 포지션이 선택되는 것을 관측할 수 있었다(그림 33 참조).

그림33. 앙상블 기법을 적용한 모델의 세부 동작 양상



자료: NH투자증권 리서치본부

5. AI와 대체데이터를 활용한 KOSPI 예측 모델 성과

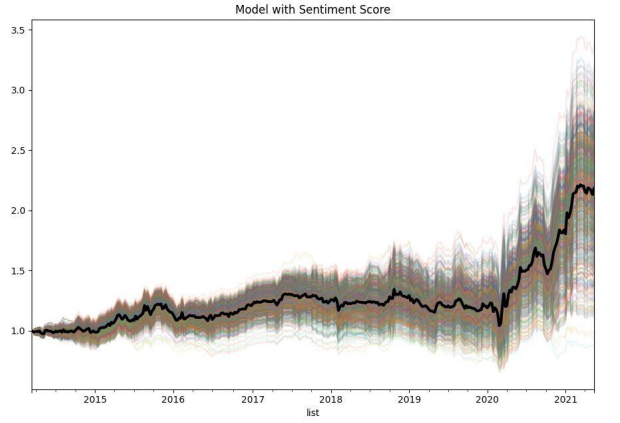
1) 모델 예측력

모델의 각 시점별 매수/매도 포지션 선택은 확률에 의해 선택되므로 결과값은 각 시행 차수마다 다르다. 이에 가장 확률이 높은 포지션에 투자하는 방식으로 백테스팅을 진행하였다.

평균 55%의 횡단면 예측률(시점별로 KOSPI의 방향을 맞춘 시행 차수의 비율(52주 평균))을 보여주었으며 52주 평균 최대 예측률은 67.4%(2021년 2월 19일)를 기록하였다.

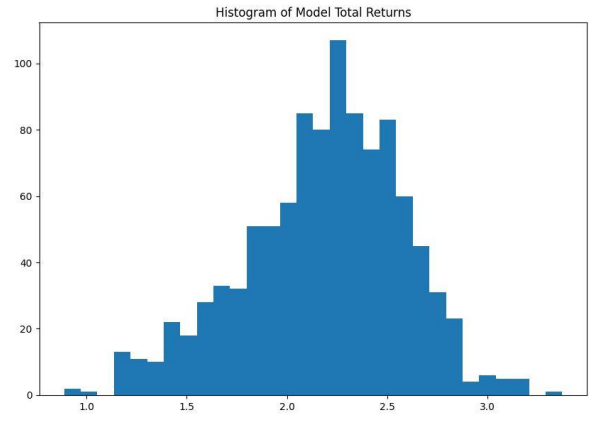
누적수익률은 2014년 3월 14일부터 2021년 10월 15일까지 139.2%(코스피 대비 82.2% 아웃퍼폼)를 기록하였다. 한번 결정된 포지션은 평균적으로 약 7~8주 유지되었다.

그림34. 시행별 모델 결과값과 평균(검은선)



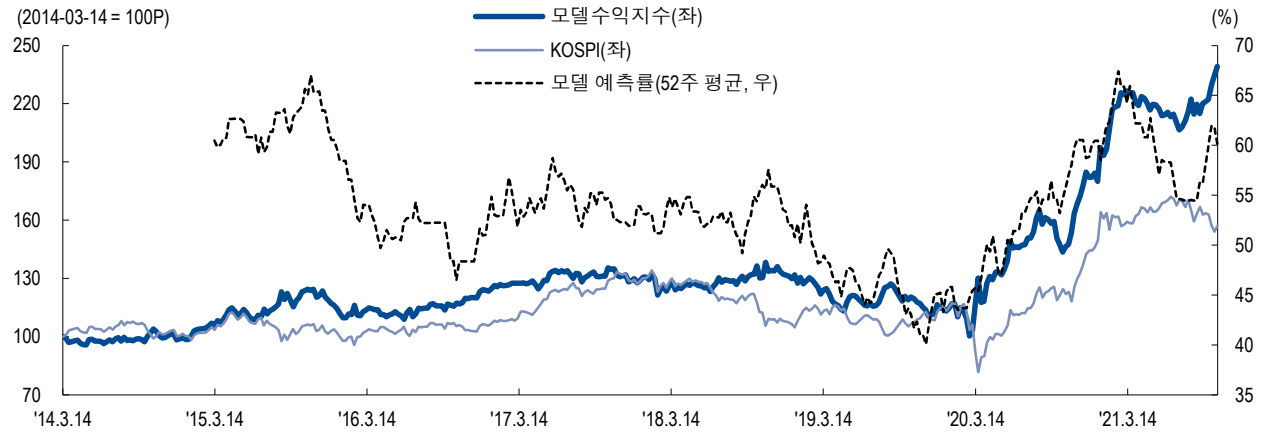
자료: NH투자증권 리서치본부

그림35. 모델 결과값 분포



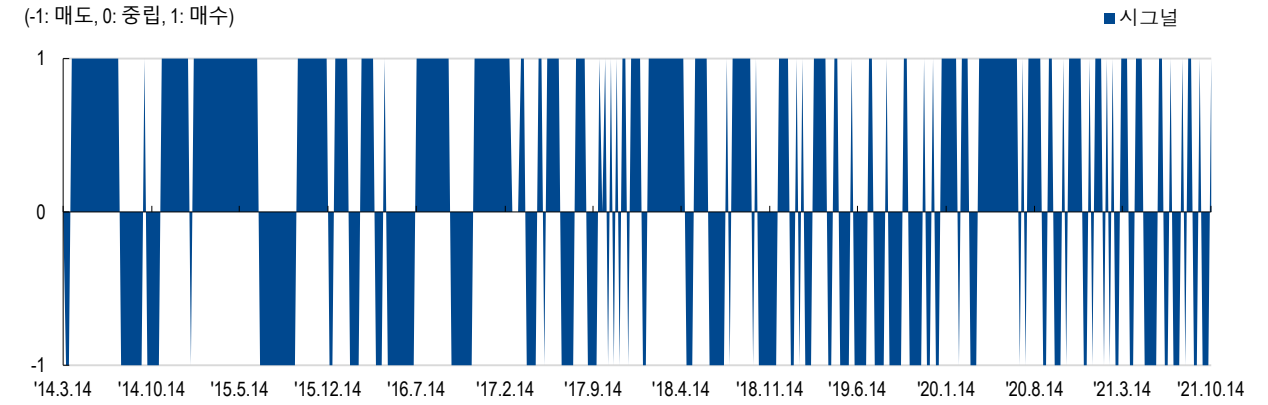
주: NH투자증권 리서치본부

그림36. 모델 수익 지수 평균 추이



자료: NH투자증권 리서치본부

그림37. 모델의 시그널 추이



자료: NH투자증권 리서치본부

수익률 측면에서 **모델은 KOSPI 대비 연평균 6%p의 초과수익률을 기록하였다. 샤프지수(Sharpe Ratio)는 KOSPI 대비 0.37p 개선**되었으며, 최대손실폭(MDD, Max Drawdown)은 11.8%p 개선됐다. 상방 변동성은 KOSPI 대비 2.5%p 높은 반면, 하방 변동성은 3.1%p 낮았다(표14 참조).

연도별로는 2015년과 코로나19 사태 이후인 2020년, 2021년에 특히 수익률이 좋았다. 다만, 2019년에는 예측률이 40%로 하락하며 KOSPI 대비 성과가 저조하였다(그림38, 41 참조).

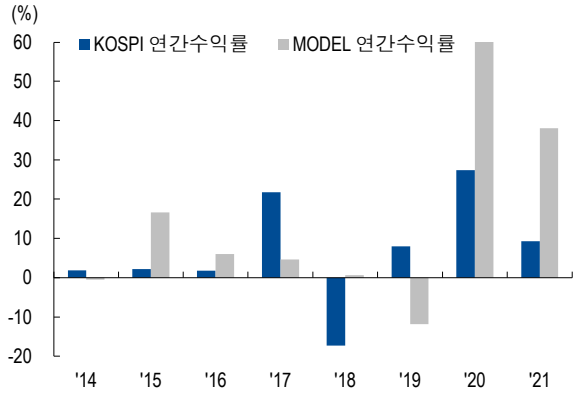
KOSPI 주간 수익률 변동폭이 큰 구간에서는 변동폭이 낮은 구간 대비 예측률이 높았으며, 변동폭이 중위인 구간에서의 예측률도 양호하였다. 또한 상승장에서 하락장 대비 높은 예측률을 기록하였다(표14, 15 참조).

표14. KOSPI 대비 모델 성과 개요

성과요약표	KOSPI 예측 모델	KOSPI
연율화 수익률(%)	12.10	6.09
샤프 지수(x)	0.74	0.37
소티노 지수(x)	1.21	0.46
최대낙폭(%)	-27.42	-39.17
표준편차(%, 52w)	16.32	16.44
표준편차(상방, %, 52w)	12.89	10.42
표준편차(하방, %, 52w)	9.99	13.13
예측(상승)률(%)	55.67	56.93

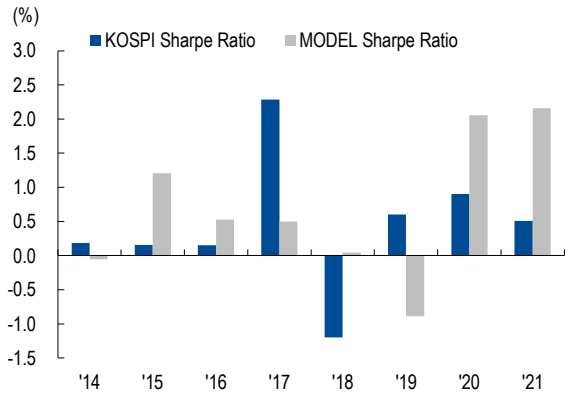
주1: 2014년 03월 14일부터 2021년 10월 15일까지 백테스트 결과
주2: 샤프 지수 = 연간수익률/연간변동성
주2: 소티노 지수 = 연간수익률/연간하방변동성
자료: NH투자증권 리서치본부

그림38. KOSPI 대비 Model 연간수익률



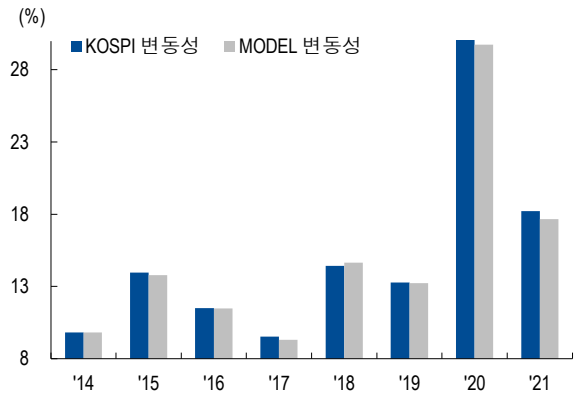
자료: NH투자증권 리서치본부

그림39. KOSPI 대비 Model Sharpe Ratio



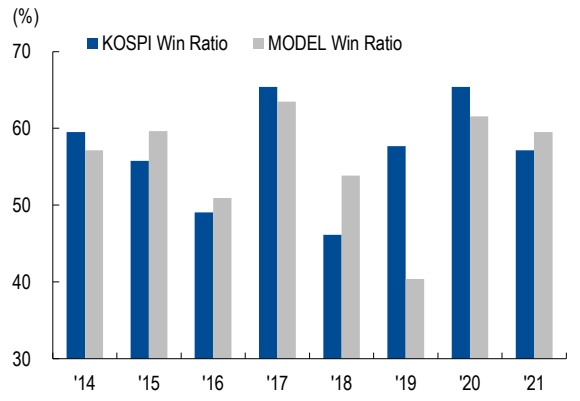
주: NH투자증권 리서치본부

그림40. KOSPI 대비 Model 연간변동성



주: NH투자증권 리서치본부

그림41. KOSPI 대비 Model Win Ratio



주: NH투자증권 리서치본부

표15. 모델 수익률 구간별 예측률

수익률구간	정답횟수(회)	총횟수(회)	예측률(%)
상위 10%	27	39	69.2
상위 10~20%	21	36	58.3
상위 20~30%	20	38	52.6
중위 30~70%	87	151	57.6
하위 30~20%	17	38	44.7
하위 20~10%	15	36	41.7
하위 10%	19	36	52.8

자료: NH투자증권 리서치본부

표16. 모델 장세별 예측률

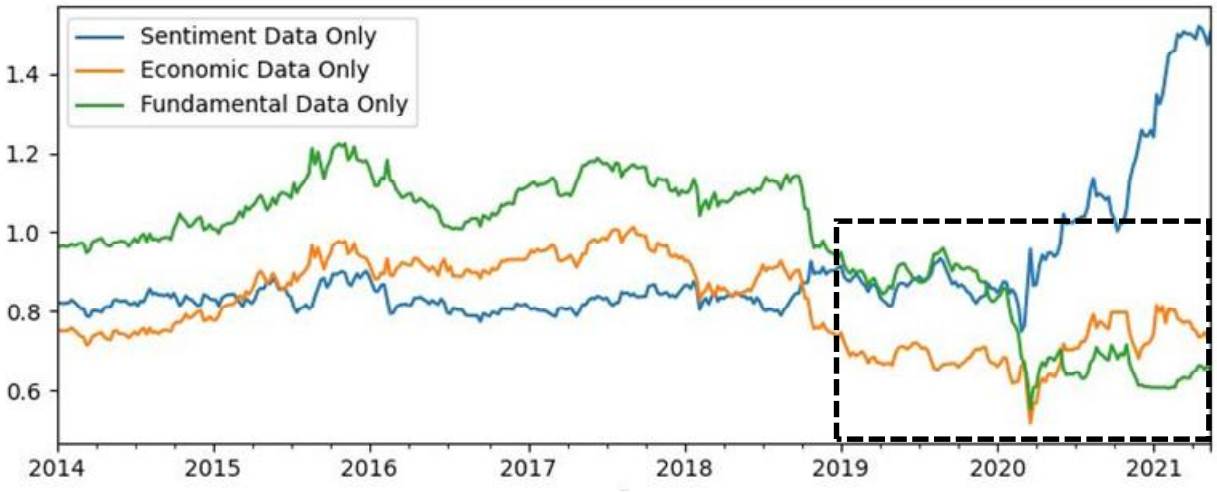
매트릭스		예측		
		상승	하락	
실제	상승	140.0	86.0	재현도 61.9
	하락	90.0	81.0	
		정밀도 60.9		정확도 55.7
F1-Score : 61.4				

주: 중립포지션은 상승에 포함하여 계산
자료: NH투자증권 리서치본부

2019년부터 전통 데이터의 알파 소멸

2019년 성과 부진은 경제지표와 펀더멘털지표와 같은 전통 데이터 변수에서 알파가 급격히 감소하며 발생한 현상으로 추정한다. 3가지 변수를 각각 단독으로 학습시킨 모델들의 성과는 2019년 모두 부진했다. 2019년 이후에는 전통 데이터인 경제지표와 펀더멘털지표의 성과가 지속적으로 악화되며 알파가 소멸하는 모습을 보였고, 대체데이터인 감성지표에 의해 모델의 알파가 발생하였다.

그림42. 각 변수를 단독으로 학습시킨 모델의 성과, 2019년부터 전통 데이터의 알파 소멸



자료: NH투자증권 리서치본부

2) 감성지표의 모델 성과 기여도

감성지표는 특히 2018년 이후로 모델 성능에 기여

다음으로 감성지표가 모델의 성과에 미치는 영향을 파악해 보기 위하여 총 6가지 모델 중에서 감성 데이터를 학습한 모델 3가지를 제외한 펀더멘털+경제, 펀더멘털, 경제로만 앙상블 기법을 적용해보았다.

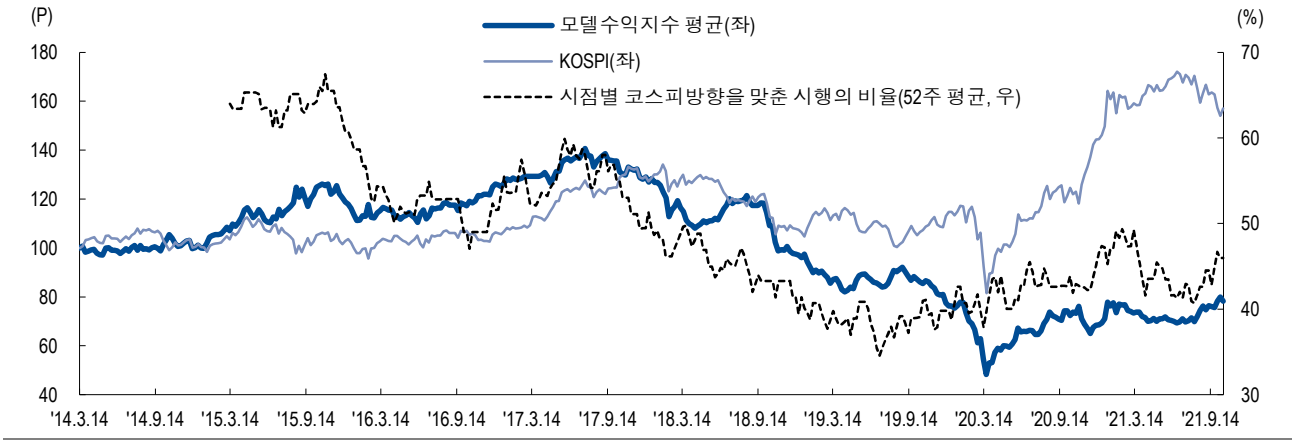
결과적으로는 2017년까지는 양호한 성과를 보였으나 2018년 이후 예측력이 50% 이하로 하락하였다. KOSPI 대비 하방변동성도 높아져 **전반적인 모델의 성능이 저하**한 것을 확인할 수 있었다(그림43 참조). 결론적으로 감성지표가 모델의 초과 수익에 기여한 것으로 파악된다.

표 17. 기존 모델 대비 감성지표 제외 모델 개요

성과요약표	감성지표 제외 모델	KOSPI 예측 모델
연율화 수익률(%)	-3.16	6.09
샤프 지수(x)	-0.19	0.37
소티노 지수(x)	-0.25	0.46
최대낙폭(%)	-65.72	-39.17
표준편차(%, 52w)	16.38	16.44
표준편차(상방, %, 52w)	10.99	10.42
표준편차(하방, %, 52w)	12.48	13.13
예측(상승)률(%)	51.13	56.93

주1: 2014년 03월 14일부터 2021년 10월 15일까지 백테스트 결과
주2: 샤프 지수 = 연간수익률/연간변동성
주3: 소티노 지수 = 연간수익률/연간하방변동성
자료: NH투자증권 리서치본부

그림43. 감성지표 제외 모델의 수익 지수 평균 추이



자료: NH투자증권 리서치본부

6. 결론

AI를 활용한 금융시장 예측은 아직 검증되지 않은 분야다. 다만, 이미 금융산업 전반에서 AI 활용을 시도하고 있으며, AI를 활용한 ETF들 역시 벤치마크 대비 아웃퍼폼하고 있는 상황이다. 이에 머신러닝 AI를 활용한 KOSPI 예측 모델 제작을 시도하였다.

우선 알파가 훼손되지 않은 입력변수를 얻기 위해 당사 주간 시장 전망 보고서의 감성지표를 추출했다. **추출한 감성지표는 KOSPI와 0.36~0.51 수준의 유의미한 상관관계를 보였다.**

이를 대입한 KOSPI 예측 모델은 입력한 모든 변수에서 알파가 존재하지 않았던 2019년을 제외하면 **수익률과 변동성 측면에서 양호한 결과**를 얻을 수 있었다. 추후 보다 다양한 대체데이터 사용과 입력변수 품질 개선, 개선된 머신러닝 알고리즘 대입을 통해 모델의 예측력을 더욱 개선할 수 있을 것으로 기대한다.

V. Appendix

Appendix A. 머신러닝의 학습 원리

알고리즘에 의하여 학습할 기준이 정해지면, 기계는 기준에 맞추어 학습(Learning)하며 최적화(Optimization) 과정을 진행한다. 최적화란 Agent가 설정한 기준과 최대한 일치하거나 기준치가 최대가 되도록 행동 규칙을 조정하는 과정이다. 예를 들어 자율주행을 목표로 하는 AI라면 최대한 빠르게 목적지에 도착하는 것이 목표가 될 것이고, 도착시간이 지연되거나 사고가 발생하여 도착에 실패할 경우에 원인이 되는 행동을 멈추도록 최적화될 것이다.

지도학습에서 회귀 알고리즘의 최적화 과정을 통해 머신러닝의 학습 원리를 파악해 보자. 회귀 알고리즘에서 대표적으로 사용되는 기준인 MSE(Mean Squared Error: 예측치 대비 관측치 오차의 제곱합 평균)의 측정식은 다음과 같다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

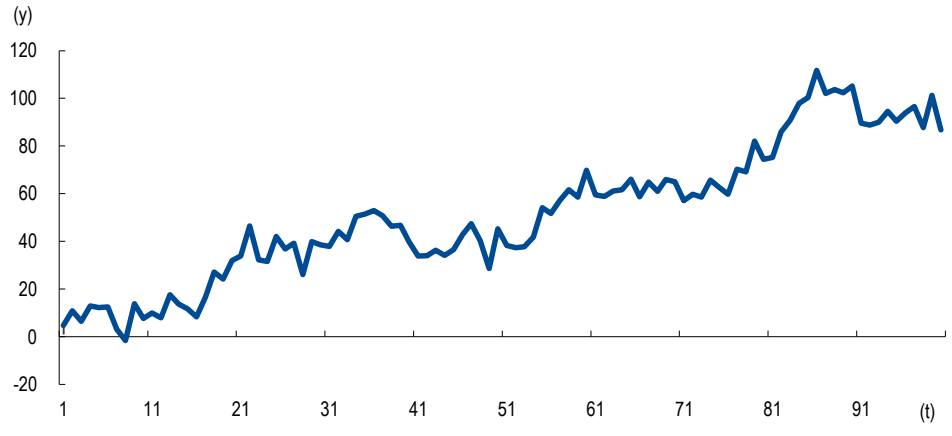
$* \hat{y}_i =$ 예측값, $y_i =$ 정답

수식에 따르면 예측값과 정답이 일치할수록 MSE는 작아진다. 따라서 머신러닝은 MSE가 최소값을 가지도록 최적화를 진행할 것이다.

MSE를 통한 최적화 과정의 예시를 들기 위해 기울기가 1인 직선에 무작위 숫자를 더하거나 뺀 시계열을 작도하였다. 작도된 시계열의 수식은 다음과 같다.

$y_t = 1 * t + \varepsilon$, ($\varepsilon$: 무작위수)

그림A1. 기울기가 1인 곡선에 무작위 숫자를 더하거나 뺀 시계열



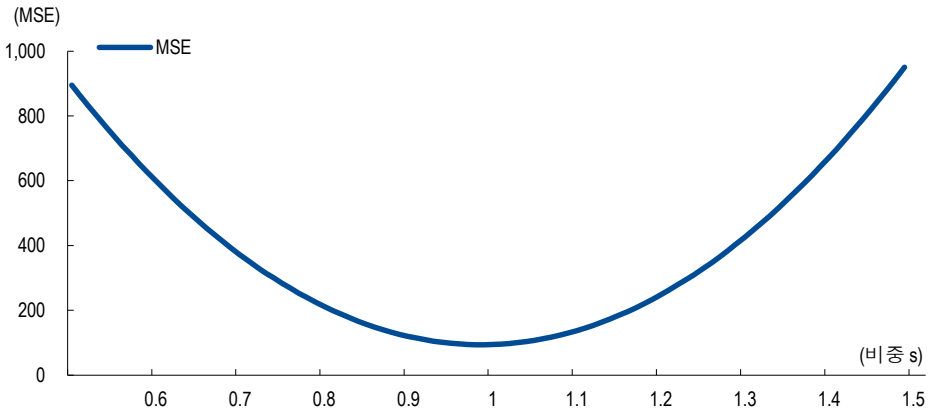
자료: NH투자증권 리서치본부

목표는 시계열을 가장 잘 표현하는 선형방정식을 찾는 것이며, 이를 위하여 MSE를 최소화하는 과정을 거칠 것이다.

정답식을 $y_t = s * t$로 추정하고, MSE를 통해서 기울기 s를 추정한다. 앞서 말한 바와 같이 MSE를 최소화할수록 예측치와 정답이 일치하므로, MSE를 최소화하는 s가 위 시계열을 가장 잘 설명하는 기울기일 것이다.

MSE를 최소화하는 s를 찾기 위하여 s와 MSE간의 관계를 그래프로 나타내면 다음과 같다(그림A2 참조). 곡선을 따라 내려가며 s가 1에 가까워질수록 MSE도 최저치에 가까워지는 것을 확인할 수 있다.

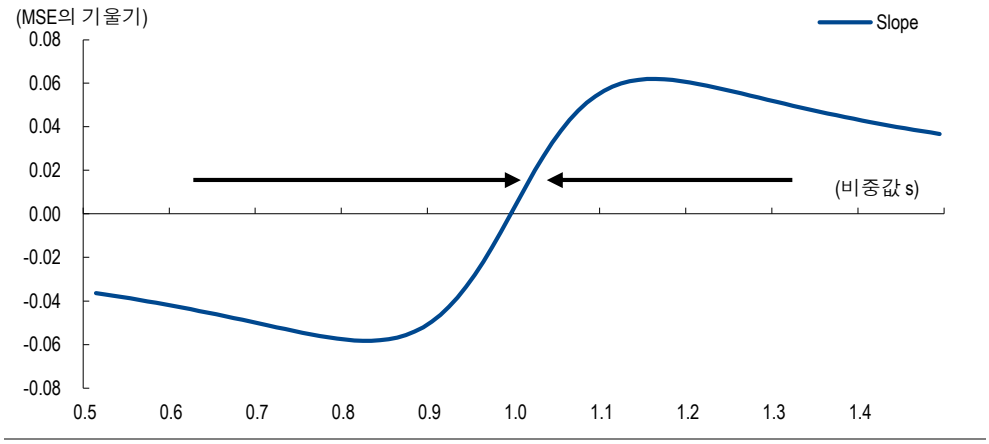
그림A2. 비중값 s가 x축이고, MSE가 y축인 곡선. s가 1일 때 MSE가 가장 작다



자료: NH투자증권 리서치본부

그렇다면 기계는 어떻게 가장 낮은 지점으로 비중값 s를 조정할 수 있을까. MSE 대신에 MSE 곡선의 기울기를 사용하면 된다(그림A3 참조).

그림A3. MSE의 기울기에 따라 방향을 조정하여 최적화 진행



자료: NH투자증권 리서치본부

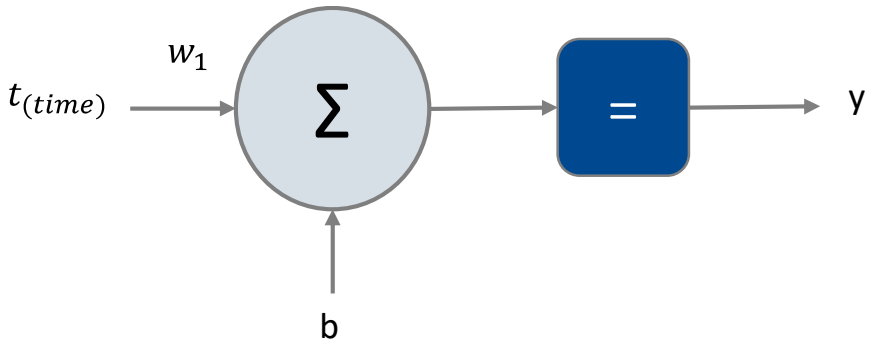
기울기가 (-)일 때에는 s를 증가시키고, 기울기가 (+)일 때에는 s를 감소시키면 MSE값을 최소화하는 s값으로 수렴하는 것을 알 수 있다. 이와 같이 곡선을 따라 가장 최적화된 비중값을 찾는 과정이 최적화이며, 머신러닝에서 AI가 스스로 정답을 찾아가는 작동원리이다. 앞서 소개한 곡선을 타고 내려가는 방법론은 경사하강법(Gradient Descent)이라고 표현하며, 머신러닝에서 가장 많이 사용되는 최적화방법 중 하나이다.

Appendix B. 인공신경망(Artificial Neural Network) 세부구조

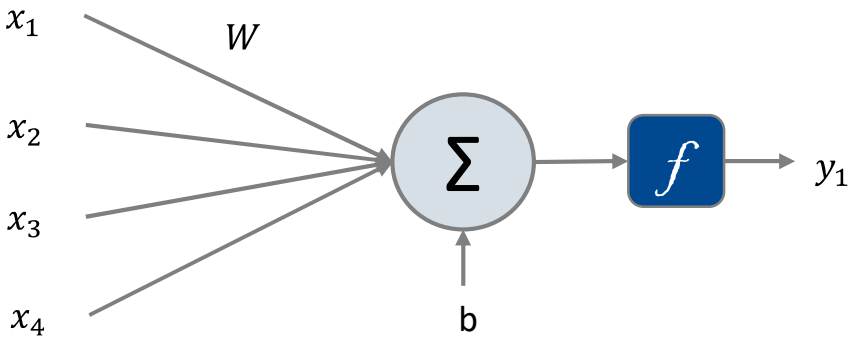
APPENDIX A에서 설정한 모형식 $y_t = s * t$의 변수를 범용적으로 표현하면 다음과 같다.

$y = w_1 * x + b$

이를 도형으로 표현하면 다음과 같다.

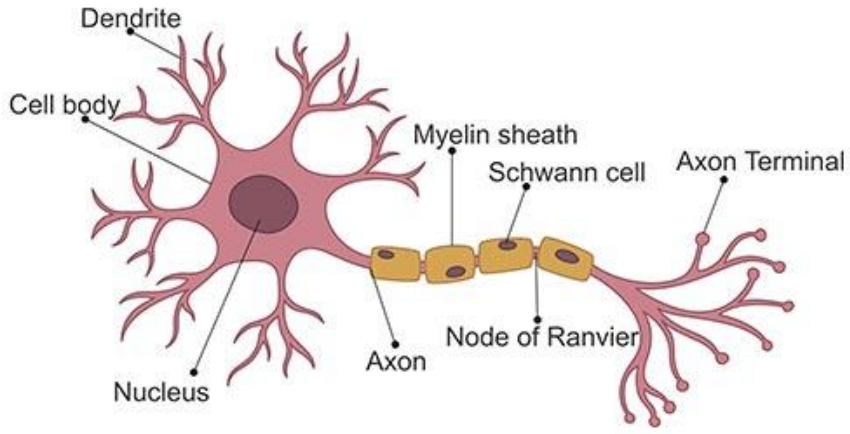


조정해야할 비중(w)과 입력값(x)이 더 많아지면 다음과 같은 모습으로 표현된다.



이와 같은 도형식의 모습이 인간 신경계의 모습과 비슷하여 인공신경망이라 표현하게 되었다. 인공신경망은 현재 머신러닝에 사용되는 가장 대표적인 방법론인 딥러닝의 근간을 이룬다.

그림B1. 인간의 신경계(Neural Network)를 구성하는 신경세포(Neuron Cell)



자료: CUSABIO, NH투자증권 리서치본부

Appendix C. 심층신경망(Deep Neural Network)

Appendix B에서 소개한 인공신경망을 응용한 AI가 현재에는 뛰어난 성능을 보여주고 있으나, 초기에는 컴퓨터 연산력 부족과 1969년 Marvin Minsky와 Seymour Papert가 발견한 문제점이 야기하는 한계점으로 인하여 정체를 겪었다.

한계점은 인공신경망이 XOR 구조와 같이 단순해 보이는 문제를 처리하지 못한다는 점이다. XOR 구조란 컴퓨터 회로에서 사용되는 논리식으로, A와 B에 서로 다른 값이 들어갔을 때에만 참값이 나오는 구조이다(그림C1 참조).

그림C1. XOR게이트의 표현식과 입/출력 관계표

표현식: $y = x_1 \oplus x_2$

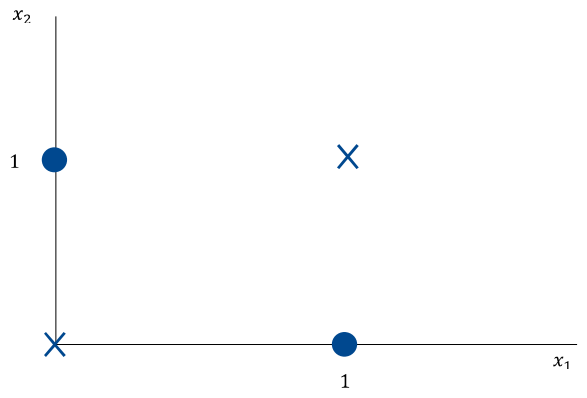
입력 (x_1, x_2)에 따른 출력값(y):

x_1	x_2		y
0	0		0
0	1	→	1
1	0		1
1	1		0

자료: NH투자증권 리서치본부

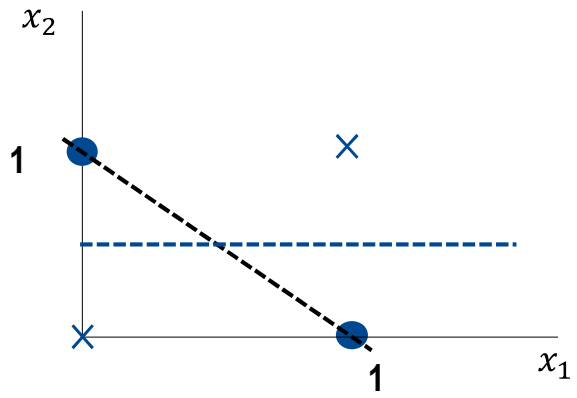
좌표상에 표현된 XOR 논리 게이트는 다음과 같다. 참은 O, 거짓은 X로 표현하였다(그림C2 참조). 직선을 사용해서는 어떤 방식으로도 참과 거짓을 완벽하게 분류할 수 없다.(그림C3 참조)

그림C2. XOR 논리 게이트의 좌표상 표현



자료: NH투자증권 리서치본부

그림C3. 그 어떤 직선도 O와 X를 완전히 구분해내지 못한다

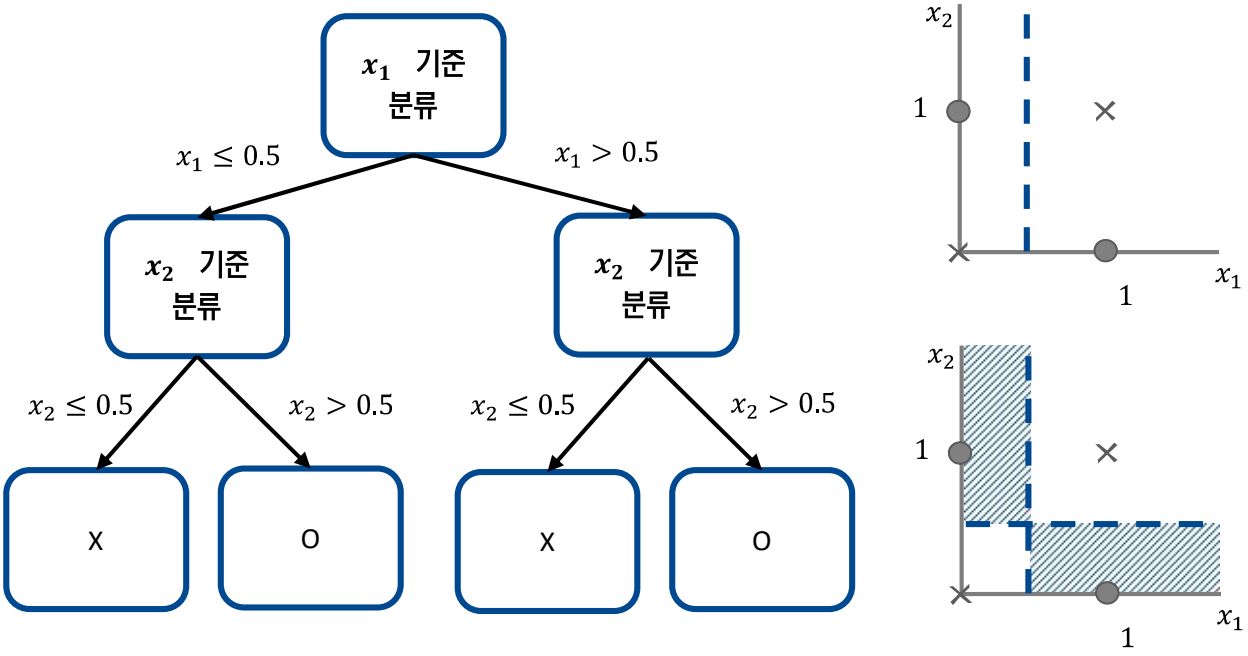


자료: NH투자증권 리서치본부

실제로 XOR문제에 대하여 단층의 인공신경망 모델을 사용하여 머신러닝을 수행하면, 아무리 많은 횟수를 학습하여도 분류 최대 정확도는 50%를 넘지 못한다. 따라서 머신러닝에 용이하도록 단순한 모델을 사용하면서 동시에 XOR 문제를 해결할 수 있는 방법을 새롭게 고안해야했다.

첫 번째 방법은 값이 잘 분류될 때까지 여러 단계에 걸쳐 영역을 나누는 것이다. 첫번째 단계에서는 x_1이 0.5보다 큰지 작은지를 기준으로 분류한다. 두번째는 x_2가 0.5보다 큰지 작은지를 기준으로 분류하면 원하는대로 분류가 가능하다. 이와 같은 방식을 의사결정나무 학습법이라 한다(그림C4 참조). 의사결정나무는 분류 과정을 직관적으로 확인할 수 있는 설명 가능한 머신러닝이라는 것이 장점이다. 그러나 단계별 분류마다 한 개의 변수만을 사용하기 때문에 변수가 많은 경우 모델이 매우 복잡해지며 과적합 문제가 발생하기 쉽다.

그림C4. 의사결정나무 학습법(Decision Tree)



자료: NH투자증권 리서치본부

두 번째 방법은 인공신경망 모델을 2단으로 구성하는 것이다. 이 경우 첫 번째 인공신경망 단에서 데이터를 구분짓는 특징값을 추출하고, 두 번째 인공신경망 단에서는 추출된 특징값을 기반으로 문제를 처리하게 된다.

이와 같은 특성으로 인하여 인공신경망 모델의 단을 더 깊게 구성할수록 데이터의 특징을 세부적으로 추출하는 것이 가능해지며, 더욱 복잡한 문제를 해결할 수 있다. 이와 같은 해결 방식이 현재 머신러닝에서 주로 사용되는 방법론인 딥러닝(Deep Neural Network, 또는 Multi Layer Perceptron)이다.

그림C5. 두 개의 신경망을 연결한 딥러닝 모델식

첫 번째단 ANN:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{111} & w_{112} \\ w_{121} & w_{122} \\ w_{131} & w_{132} \\ w_{141} & w_{142} \\ w_{151} & w_{152} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \\ b_{15} \end{bmatrix} \right),$$

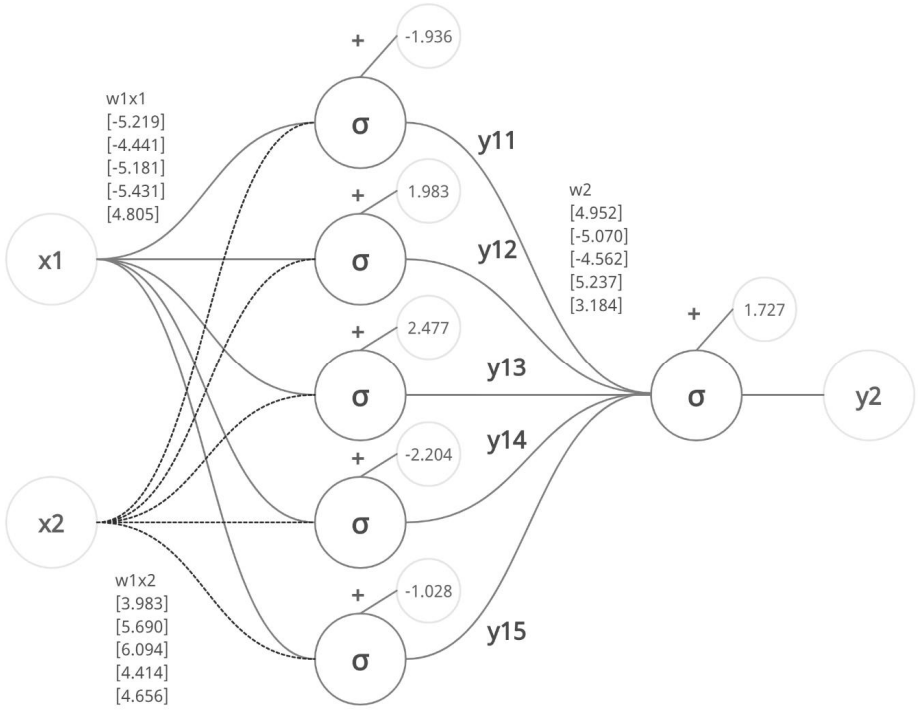
$$Sigmoid\ \ Function\colon\ \sigma(x) = \frac{1}{1+e^{-x}}$$

두 번째단 ANN:

$$y_2 = \sigma\left(\begin{bmatrix} w_{21} \\ w_{22} \\ w_{23} \\ w_{24} \\ w_{25} \end{bmatrix} [y_{11}\ \ y_{12}\ \ y_{13}\ \ y_{14}\ \ y_{15}] + b_2 \right)$$

자료: NH투자증권 리서치본부

그림C6. 두 개의 신경망을 연결한 딥러닝 모델 Diagram



자료: NH투자증권 리서치본부

표C1. 첫 번째단의 ANN을 통해 추출된 특징값

(x1,x2)		(0,0)	(0,1)	(1,0)	(1,1)
변환식 1	→	0.12614	0.885734	0.00078	0.040251
변환식 2	→	0.878973	0.999535	0.078865	0.962031
변환식 3	→	0.922493	0.999811	0.062724	0.967394
변환식 4	→	0.099405	0.901193	0.000483	0.038404
변환식 5	→	0.263442	0.974111	0.977625	0.999782

자료: NH투자증권 리서치본부

표C2. 두 번째단의 ANN을 거치며 높은 정확도의 분류값이 출력된다

최초입력값	최종출력값
(0,0)	0.007 ≈ 0
(0,1)	0.987 ≈ 1
(1,0)	0.985 ≈ 1
(1,1)	0.018 ≈ 0

자료: NH투자증권 리서치본부

Appendix D. TLH 전략 예시[주)]

TLH 전략 예시1

1년 이상 보유한 펀드 A: 미실현손실 130,000달러
1년 미만 보유한 펀드 B: 미실현손실 100,000달러

1년 이상 보유 실현이익: 200,000달러
1년 미만 보유 실현이익: 150,000달러

TLH 전략 미적용 시:
과세액 = 200,000*20% + 150,000*37% = 95,500달러

TLH 전략 적용 시:
미실현손실을 실현(혹은 수확(Harvesting)이라 표현)
과세액 = (200,000-130,000)* 20% + (150,000-100,000)*37% = 32,500달러
총절세액: 63,000달러.

TLH 전략 예시2

연초에 산업 A에 투자하는 ETF에 100,000달러 투자. 산업 A는 중도에 급락하여 10% 손실이 발생(-10,000달러). 그러나 이후 반등하여 연말에 8% 이익으로 마감(+8,000달러). 해당 이익은 미실현. 그외 포트폴리오에서 발생한 단기실현이익은 15,000달러

TLH 전략 미적용 시:
과세액: 15,000*37% = 5,550달러

TLH 전략 적용 시:
10% 손실 발생 시 즉시 손실을 수확. 산업 A를 추종하는 비슷한 펀드나 ETF를 사용하여 재매수. 이 경우 수확한 손실(-10,000달러)을 사용하여 기말에 절세 가능
과세액: (15,000-10,000)*37% = 1,850달러
총절세액: 3,700달러

주) 출처: Investopia, NH투자증권 리서치본부

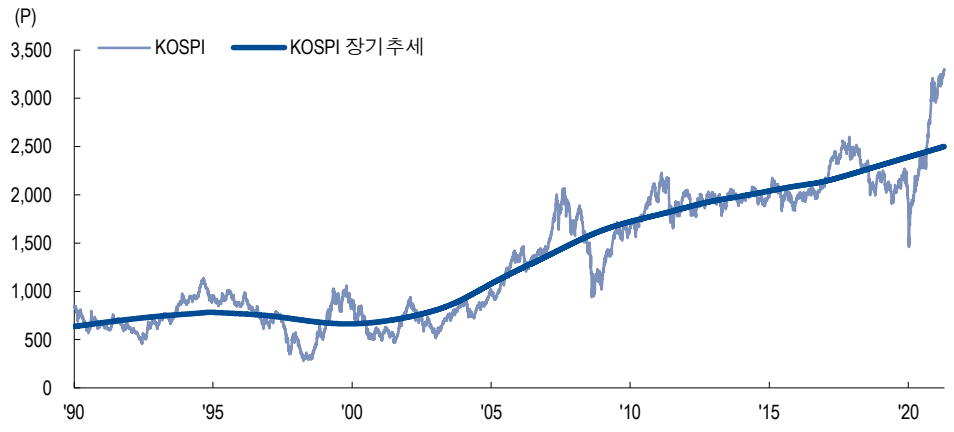
Appendix E. KOSPI 시계열 분해 방법론

KOSPI를 장기추세 + 추세제거 KOSPI로 시계열 분해하였다. 이 때 계절성은 고려하지 않았다. 시계열 분해에는 STL(Seasonal and Trend decomposition using Loess) 기법을 사용하였다.

STL은 장기 선형 추세를 국소 선형회귀분석(Locally Estimated Scatterplot Smoothing, Loess)을 사용하여 감지하는 기법으로 이동평균기법과 대비하여 입력 기간이 짧더라도 평탄화된 장기 선형 트렌드를 추출하는 데에 유리하다.

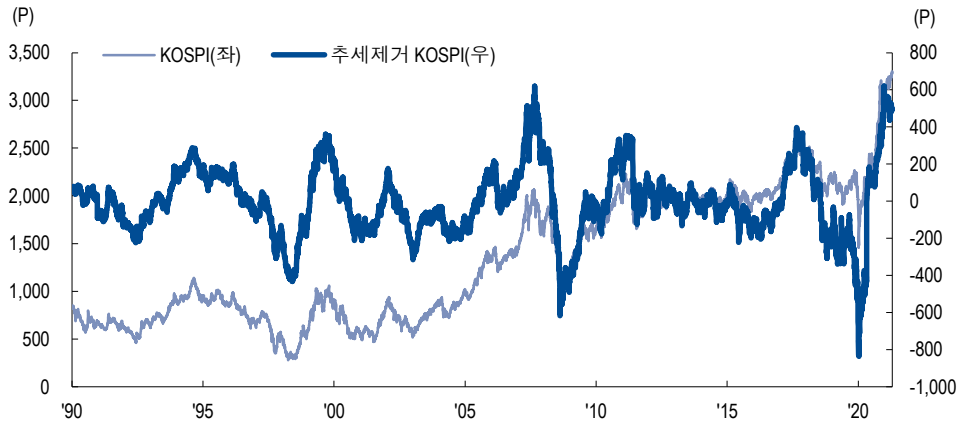
이와 같은 방법으로 분해한 KOSPI 시계열 데이터는 각각 그림E1, E2와 같다.

그림E1. KOSPI와 장기추세



자료: NH투자증권 리서치본부

그림E2. KOSPI와 Detrended KOSPI



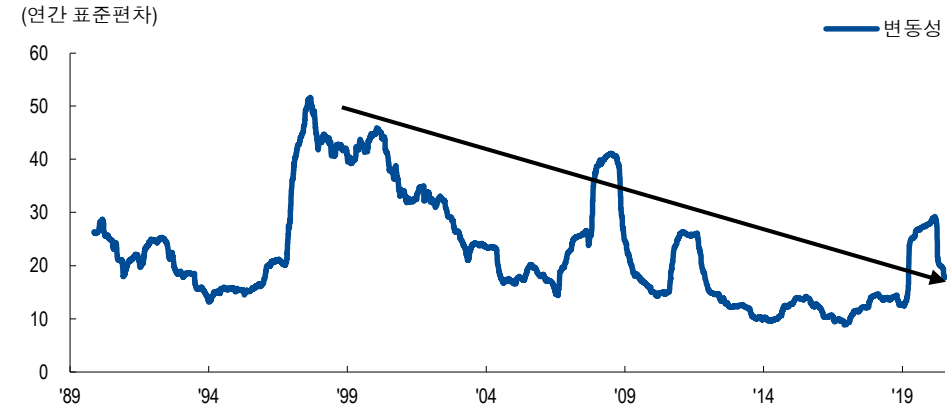
자료: NH투자증권 리서치본부

이 때 KOSPI의 레벨에 대한 고려 없이 시계열을 분해했기 때문에 추세제거 KOSPI의 진폭은 시간이 경과할수록 점차 확대되는 양상을 보인다. 주가 수익률의 변동성이 점차 증가하고 있다면 진폭의 확대가 타당하지만, 실제로는 KOSPI 주가 수익률의 변동성은 점차 축소되는 추세를 보인다(그림 E3 참조). 따라서 진폭의 증가분을 조정해주어 시계열이 정상성을 갖도록 하였다.

$$\text{조정 } KOSPI_t = \text{추세제거 } KOSPI_t * \frac{\text{장기 추세}_t}{\text{장기 추세}_0}$$

장기 추세_0: 기준시점의 장기 추세

그림E3. KOSPI 일간 수익률의 연간변동성은 감소 추세



자료: NH투자증권 리서치본부

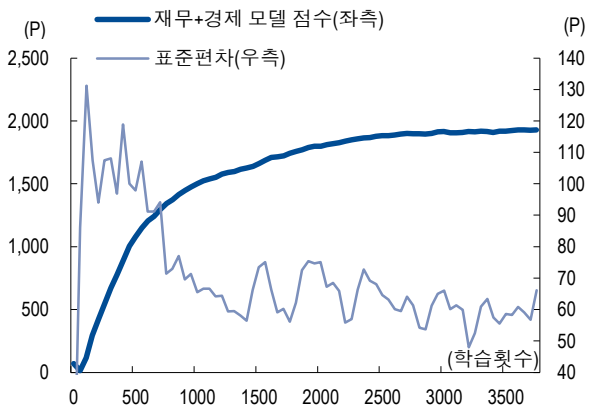
Appendix F. 강화학습 모델 최적화 과정

모델은 매주 마지막 영업일에 다음주 KOSPI 포지션을 결정한다. Out of Sample 오류를 피하기 위하여 조정 KOSPI 데이터와 주성분 계수 등은 모델 학습 시점, KOSPI 포지션 선택 시점마다 계산했다.

각 입력값에 대하여 최대 만 번의 시뮬레이션을 진행하며, 시뮬레이션 50번마다 메모리에 기억된 시뮬레이션 결과를 모델에 학습시켰다. 과적합을 피하기 위하여 과거 100번의 시행이 과거 101번부터 200번까지의 시행보다 최대 수익률이 하락하는 동시에 분산값이 증가한다면 학습을 종료하였다. 학습 중에 국지최적값에 빠져 동일한 결과값이 반복되는 경우가 발생하는 상황을 방지하기 위하여 과거 50번 시행에서의 결과값 평균이 과거 50번에서 100번까지와 동일할 경우, 일시적으로 10%의 확률로 랜덤 선택을 하도록 설정하였다. 그럼에도 모델의 성능이 개선되지 않는다면 학습을 종료하였다.

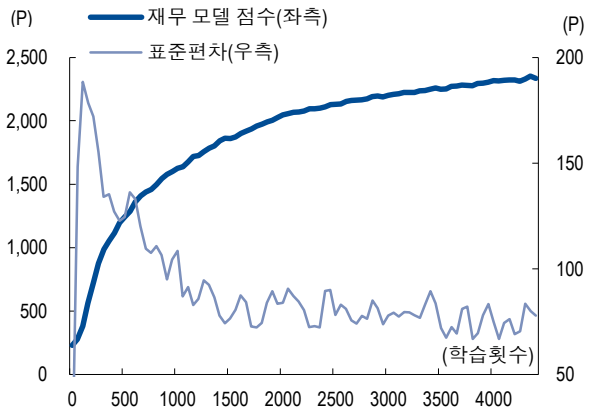
평균적으로 시뮬레이션이 1,600회 가량 시행되었을 때 학습이 종료되었으며, 많을 경우 4,400회까지 시행되었다. 시뮬레이션은 감성, 재무, 경제 데이터의 조합(감성+재무, 감성+경제, 재무+경제, 감성, 재무, 경제 총 6가지)별로 각각 시행되었는데 전반적으로 테스트 기간이 길수록, 입력값의 차원이 높을수록 최적화되는데에 더 오랜 시간과 회차가 소요되었다. 데이터별로는 재무 → 재무+경제 → 경제 → 감성+경제 → 감성+재무 → 감성 데이터 모델 순으로 학습에 더 많은 시간이 소요되었다.

그림F1. 재무+경제 데이터 학습 모델의 학습 과정 사례



자료: NH투자증권 리서치본부

그림F2. 재무 데이터 학습 모델의 학습 과정 사례



자료: NH투자증권 리서치본부