true friend 한국투자증권

퀀트 전략
이슈
2020.6.7

Korea Quant

인공지능은 어떻게 주식을 고를까?

팩터 투자의 new 알파: 인공지능

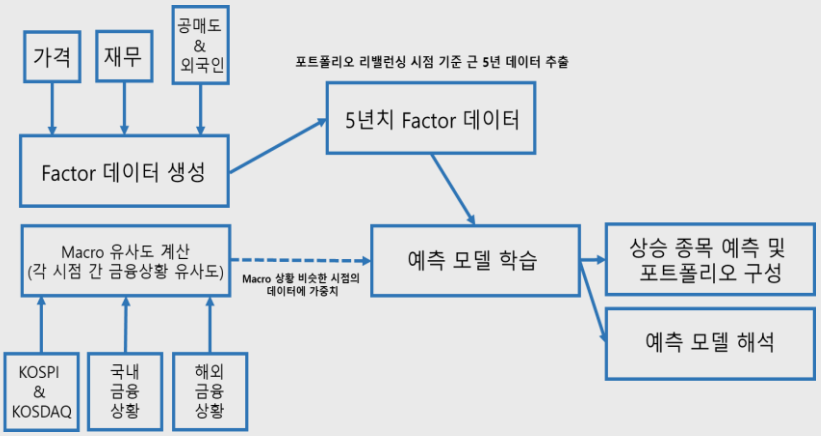
팩터 투자의 알파는 최근 들어 점점 소멸되고 있다. 금융 분야에서 정보 비대칭성이 해소됨에 따라 계량 투자 접근성이 높아지면서 투자자들에게 알려지기 시작했기 때문으로 추정된다. 이러한 현상은 기존 팩터들이 창출하던 알파를 더 이상 유효하지 않게 만든다. 다시 말해서 몇몇 우수한 개별 팩터에 투자하는 방식은 더 이상 초과수익을 내기 어렵다. 하지만 단일 개별 팩터에 투자하는 것이 아닌 **개별 팩터 여러 개를 조합했을 때**는 그렇지 않다. 본 팀은 인공지능(AI)을 이용해 팩터들 간 최적의 조합을 찾았다. AI는 매 기간 여러 가지 변수를 고려하여 최적의 팩터 조합을 찾아 포트폴리오를 구성했다. 그리고 결과적으로 기존 방식 대비 뛰어난 초과 수익을 창출하는 것으로 확인되었다.

해석 가능한 인공지능

AI기반 모형의 가장 큰 단점은 해석이 어렵다는 점이다. 인공지능이 아무리 좋은 주식을 잘 골라낸다고 하더라도, 인공지능이 '그 때 왜 A종목을 골랐는지' 설명할 수 없다면, 투자자들을 납득시키기 힘들 것이다. 이것이 오랫동안 AI가 퀀트 투자에 직접적으로 이용될 수 없었던 이유이며, 본 팀은 이러한 AI 모형의 문제점을 해결했다.

LIME이라는 통계 기법을 사용해서 AI의 사고 과정을 해석했으며, 해석한 내용을 바탕으로 모델을 개선시켰다. 이를 통해, 팩터 투자 알파가 사라졌던 최근 한국 주식 시장에서도 상당한 초과수익을 창출할 수 있었다.

전체 개요



자료: 한국투자증권

목차

권순규
purestar0509@gmail.com
황원영
jasonhw96@naver.com

I. 서론

1. 주제 선정 배경

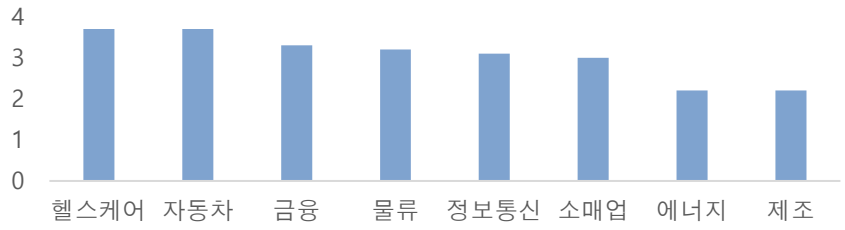
1) 커져가는 AI의 중요성

금융 시장에서도 커져가는 AI의 영향력

4차 산업 혁명을 맞아 주목을 받는 산업 중 단연코 화제는 인공지능(AI)이다. 이전과는 비교할 수 없는 데이터의 양과 이를 처리 가능하게 해준 컴퓨팅 능력의 비약적인 발전이 인공지능 산업을 발전시키고 있다.

금융시장에서도 인공지능을 활용하려는 움직임이 커지고 있다. 인공지능을 활용한 펀드 운용 규모의 증가, 로보 어드바이저 상품의 증가를 통해 이를 확인할 수 있다. 또한 PwC(Pricewaterhouse Coopers)에서 발표하는 AI 임팩트 지수에 따르면, 금융업은 헬스케어와 자동차산업에 이어 주요산업 중 세 번째로 **AI가 미치는 잠재적 영향력이 클 것**으로 예측하고 있다.

[그림 1] AI 임팩트 지수



자료: 한국투자증권

인간이 고려할 수 없는 수많은 경우의 수를 탐색할 수 있는 인공지능

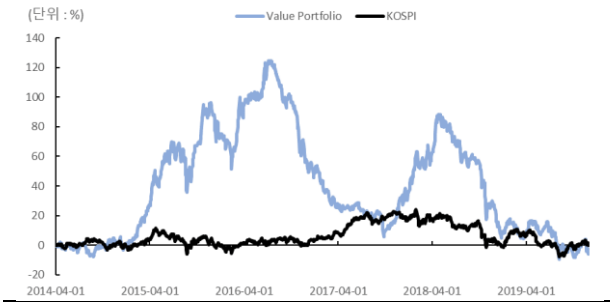
금융에 AI를 접목시켰을 때 가질 수 있는 가장 큰 장점은 투자를 함에 있어 **인간이 일일히 고려할 수 없는 경우의 수를 탐색할 수 있다는 점**과 **비선형적으로 얽혀 있는 금융시장의 수익 구조를 파악할 수 있다는 점**이다. 계량 투자를 예로 들자면, 투자에 활용할 수 있는 지표는 유한하지만 어떤 조합과 비중으로 투자하는지 결정해야 하는 측면에서는 무한한 조합이 존재할 수 있다. 인간이 고려할 수 있는 조건의 한계를 극복하는 데 도움을 준다고 이해하면 될 것이다.

본 보고서에서는 기존의 **전통 팩터투자에 AI를 접목**한 새로운 투자 모델을 소개한다. 여기서 말하는 팩터 투자란 자산의 수익률에 영향을 미치는 요인들을 기반으로 투자하는 투자 방식이다. 장기적으로 시장 대비 초과수익 창출을 목표로 한다. 수치에 기반하였기 때문에 일종의 퀀트 투자로 분류된다.

2) 전통 팩터 투자 전략의 한계

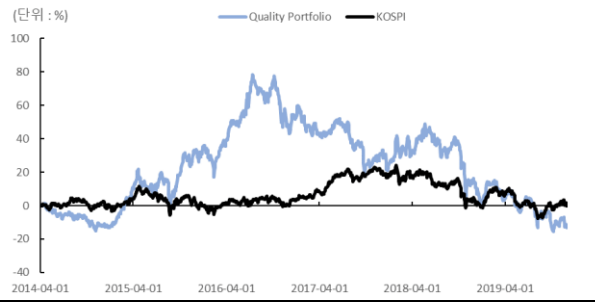
본 팀은 먼저, 개별 팩터 투자 성과를 확인해보았다. 한 개 혹은 여러 개의 팩터에 집중적으로 투자하는 퀀트 전략은 이미 알파를 상실하고 있음을 확인할 수 있었다.

[그림 2] Value Portfolio



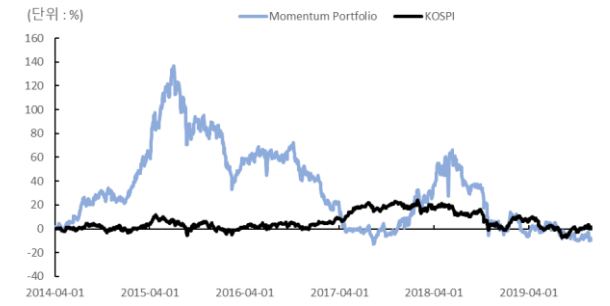
주: Value 팩터 (PER, PBR, PSR, CFY) 스코어 상위 50개 종목으로 포트폴리오 구성, 월간 리밸런싱, 동일 비중 포트폴리오
자료: 한국투자증권, FnGuide

[그림 3] Quality Portfolio



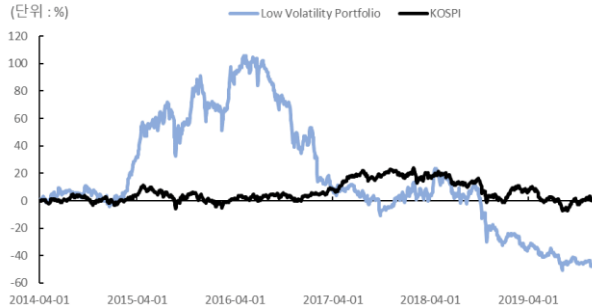
주: Quality 팩터 (ROA, ROE, CFO, Leverage, Liquidity 등) 스코어 상위 50개 종목으로 포트폴리오 구성, 월간 리밸런싱, 동일 비중 포트폴리오
자료: 한국투자증권, FnGuide

[그림 4] Momentum Portfolio



주: Momentum 팩터 (12_2개월, 6_2개월 모멘텀) 스코어 상위 50개 종목으로 포트폴리오 구성, 월간 리밸런싱, 동일 비중 포트폴리오
자료: 한국투자증권, FnGuide

[그림 5] Low Volatility Portfolio



주: Low volatility 팩터 (1년 주간&일간 volatility, 2년 주간&일간 volatility) 스코어 상위 50개 종목으로 포트폴리오 구성, 월간 리밸런싱, 동일 비중 포트폴리오
자료: 한국투자증권, FnGuide

기존 팩터 투자의 최근 저조한 성과

위의 그림은 전통적 팩터 투자 방식의 누적 수익률을 보여준다. 팩터는 크게 벨류, 퀄리티, 모멘텀, 변동성 4개에 대해 투자했고, 한 팩터에 전부 투자하는 방식과, 동시에 여러 개 팩터에 대한 비중을 다르게 하는 '팩터 통합'(Integrating)방식을 모두 사용하였지만, 네 개 팩터 모두 **2017년 이후로 수익률이 저조**한 것을 알 수 있었다.

이를 통해, 한국 주식시장에서 **단순히 몇몇 팩터를 선형적으로 조합하는 방식으로는 더 이상 초과수익을 창출하기 힘들다**는 결론을 내렸다. 본 팀의 AI모델은 이런 상황의 인식에서 출발하였다.

AI 해석을 통한 종목 선정 근거 탐색

'AI를 활용한 팩터투자' 자체는 새로운 아이디어가 아니다. 이를 활용한 연구가 활발히 진행중이고 관련 자료에 대한 접근도 쉽다. 하지만 AI모델에는 한 가지 치명적인 문제점이 있다. **바로 해석이 어렵다는 점이다.** AI가 어떤 논리로 최적의 팩터 조합 혹은 종목을 선정했는지 알 수가 없다는 뜻이다.

본 보고서에서는 이를 해결하고자 **AI를 해석할 수 있는 통계 기법을 적용**하였으며, 이를 통해 기존 AI를 개선해주었다.

다시 말해, 백테스트 기간 동안 우리가 만든 AI모형이 왜 이러한 기준으로 종목을 선정했는지 그 이유를 파악하고, 이 **피드백을 바탕으로 현재 시점에서 선택할 수 있는 최고의 투자 전략을 찾는 것**이 이 보고서의 궁극적인 목표라고 할 수 있다.

II. 본론

1. 데이터 수집 및 전처리

1) 데이터 수집

본 팀이 이번 분석을 위해 수집한 데이터는 크게 두 종류로 나뉜다. 첫 번째는 지난 20년간의 국내 상장 주식의 가격과 재무제표 데이터이며, 이는 한국거래소, DataGuide, Quantiwise, 그리고 KISVALUE 등의 프로그램을 통해 수집했다. 두 번째는 거시경제 및 매크로 데이터이며 DataGuide와 Bloomberg를 통해 수집했다.

2) 데이터 전처리

위의 과정을 거쳐 수집한 데이터의 길이는 20년이며, 수집한 변수는 수백개에 달하는 방대한 데이터이다. 따라서 **AWS 클라우드 서비스에 데이터베이스를 직접 구축하여 SQL로 데이터 저장/관리의 효율성을 높였다.** 안정된 DB를 바탕으로 보다 다양한 경우의 수의 백테스트를 진행할 수 있었다.

2. 팩터 생성

아무리 AI가 스스로 좋은 팩터 조합을 찾아준다고 해도, **팩터 자체의 퀄리티가 좋지 못하다면 포트폴리오 성과 역시 부진할 것**이다. 따라서, 본 팀은 가장 먼저 좋은 팩터를 생성하기 위해 다음과 같은 과정을 거쳐, 최종적으로 약 250개의 팩터를 완성했다.

중요한 것은 팩터들이 각 종목들의 특성을 얼마나 잘 반영하고 있는지 혹은 방해가 되는 노이즈가 포함되어 있는 지이다. 이를 고려하기 위해서는 각 종목이 속하는 산업군에 대한 이해, 시장 상황에 따라 어떤 영향을 받는지에 대한 이해 등이 필수적이다.

1) 기업 팩터

기업 팩터는 크게 밸류, 퀄리티, 모멘텀, 변동성

기업 팩터의 종류는 크게 4가지로 나눌 수 있다. 기업 주가의 상대적 가치를 나타내는 벨류팩터(Value), 기업 자체의 우량성을 나타내는 퀄리티 팩터(Quality), 전반적인 주가 상승 추세를 나타내는 모멘텀팩터(Momentum), 주가의 변동성과 위험성을 나타내는 변동성팩터(Volatility)이다. 본 보고서에서 설명될 팩터 모델은 보편적으로 활용되는 팩터를 최대한 활용하였다. 추가적으로, 위에서 언급한 점에 대한 고려를 반영하여 종목의 특성을 잘 나타낼 수 있는 정보를 선별하였다.

소형주 팩터는 상장폐지 데이터의 부재로 사용 X

대표적으로 자주 쓰이는 전통 팩터 중 하나인 소형주(Size)팩터는 사용할 수 없었다. 가장 큰 **이유는 상장 폐지 종목들에 대한 데이터를 구할 수 없었기 때문**이다. (하지만, 상장 폐지 데이터는 현재 기업정보 포탈서비스인 KISVALUE를 통해 수집 완료했으며, 빠른 시일 내에 AI모델에 추가할 예정이다.)

대표적으로 앞서 언급했던 4개 상위 팩터를 생성했던 본 팀의 방식을 소개한다.

<표 1> Value Factors (단위:)

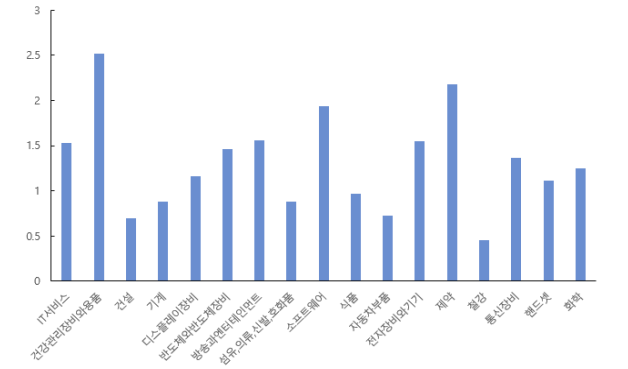
EY_R	(종목 EY – 업종 EY 중위값) / 업종 EY 중위값
PBR_R	(종목 PBR – 업종 PBR 중위값) / 업종 PBR 중위값
PSR_R	(종목 PSR – 업종 PSR 중위값) / 업종 PSR 중위값
CFY_R	(종목 CFY – 업종 CFY 중위값) / 업종 CFY 중위값

자료: 한국투자증권

Value: '동일 업종 대비' 해당 기업이 저평가 되있는가?

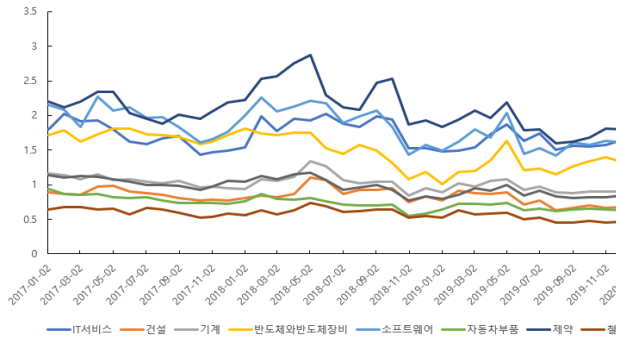
Value 팩터는 가치주를 선별하기 위해 활용되는 대표적인 팩터이다. 가장 많이 활용되는 팩터로는 PER, PBR 등이 있다. 핵심적인 의미는 주가에 비해 기업이 얼마나 고평가 혹은 저평가 되어 있느냐이며 주로 저평가된 종목을 발굴하는 데에 사용된다. 하지만 산업 군에 따라 고평가와 저평가의 기준이 상이하기 때문에 본 모델에서는 기존 Value 팩터를 **해당 종목이 속한 산업군과 비교한 값**으로 변형해서 생성했다.

[그림 7] 업종별 평균 PBR



주: 20년 1분기 실적 바탕
자료: FnGuide

[그림 8] 주요 업종 평균 PBR 흐름



자료: FnGuide

<표 2> Momentum & Volatility Factors (단위:)

Idiosyncratic_momentum	고유 수익 모멘텀 (6_2 모멘텀, 12_2 모멘텀, 24_2 모멘텀, 1달 모멘텀)
Idiosyncratic_volatility	고유 수익 변동성 (1년 일간 & 주간 변동성, 2년 일간 & 주간 변동성)
Momentum	6_2 모멘텀, 12_2 모멘텀, 24_2 모멘텀, 1달 모멘텀
Volatility	1년 일간 & 주간 변동성, 2년 일간 & 주간 변동성

주: _2 의 의미는 최근 1달의 수익률은 고려하지 않았다는 의미이다. 장기 모멘텀을 계산할 때 최근 1달을 계산하지 않는 이유는 근 1달의 모멘텀이 반전 효과를 가져오는 경우가 많기 때문에 따로 고려하였다.

Momentum & Volatility: 수익률과 변동성을 설명

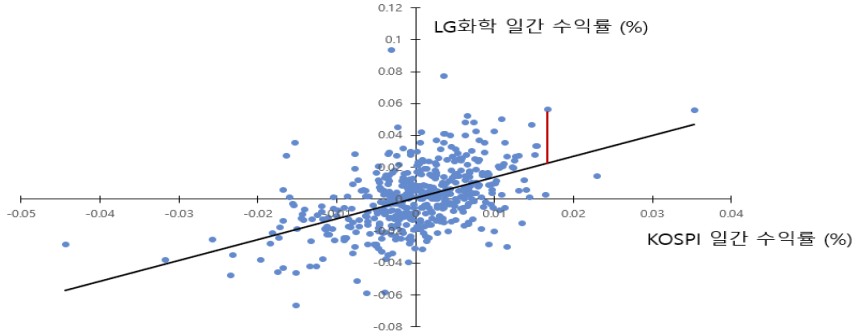
Momentum과 Volatility 팩터는 종목의 수익률을 바탕으로 계산되는 팩터이다. Momentum은 특정기간의 수익률이 얼마였는지, Volatility는 일간 수익률의 변동성이 얼마나 컸는지에 대한 정보이다. 종목의 수익률은 수 없이 많은 이유에 영향을 받겠지만 가장 큰 이유 중 하나는 시장 상황이다. 시장에 유입되는 자금이 많을수록 종목들의 수익률이 오를 확률이 높기 때문이다.

고유 수익률: 시장의 영향력을 제외한 수익률

해당 모델에서는 **각 개별 종목들의 수익률에서 시장의 영향력을 제외한 고유 수익을 활용하여 Momentum과 Volatility 팩터를 재생성**한다. 특정 종목의 고유 수익은 다음과 같이 계산된다. 종목의 일간 수익률을 시장(KOSPI, KOSDAQ)의 일간 수익률로 설명하는 회귀모형을 적합시킨다. 이 때 시장 수익률로 설명되지

않는 부분인 잔차항을 해당 기업의 고유수익으로 고려한다. 쉽게 말해, 고유수익은 **시장의 영향을 받지 않는 선에서 어느 정도의 수익률을 창출했는지**의 의미로 생각할 수 있다.

[그림 9] 고유 수익 시각화



위의 그림은 고유수익의 계산 방식을 시각화 한 것이다. 검정색 선이 KOSPI의 일간 수익률로 LG화학의 일간 수익률을 설명하는 회귀 모형이다. 이 때 해당 회귀모형의 기울기가 흔히 말하는 beta이다. 2년치의 데이터를 가지고 위와 같이 회귀 식을 적합한 이후, 회귀 모형이 설명하지 못하는 부분(오차항, 그림에서 빨간 선)을 고유 수익으로 고려한다.

기존 팩터만 활용했을 때의 백테스트 성과 부진

이 4가지 상위 팩터들은 총 50개 정도의 하위 팩터들로 구성 된다. 본 팀은 이를 바탕으로 1차적인 백테스트를 진행해보았으나 2017년 이후 성과가 부진했다.

새로운 팩터 생성

이에 따라 공매도, 외국인수급, 신용융자, 대차잔고 등의 정보를 담고 있는 20개의 팩터들을 추가로 생성하였다.

<표 3> 공매도, 외국인, 신용융자, 대차잔고 팩터 (단위:)

활용 데이터	팩터
거래량	거래량 회전율 (누적 주식 거래량 / 총상장주식수)
외국인	외국인 지분율
공매도	공매도 잔고 비율, 공매도 체결 비율
대차거래	대차거래 체결비율, 대차거래 잔고금액 비율
신용융자	신용융자 잔고수량 비율, 신용융자 신규 비율

자료: 한국투자증권

2) 매크로 (거시경제)

국내외 금융상황을 고려한 매크로 변수

매크로 변수는 국내 금융 상황과 국외 금융 상황을 반영한 변수를 사용한다. 국내 금융 상황의 경우 크게 **금리, 금리 스프레드, 물가, 신용** 총량으로 나뉜다(신용 총량이란 유동성의 수요와 공급에 관한 정보를 제공하는 지표이다.) 그리고 해외 금융 상황의 경우에는 주요국의 금리와 환율을 사용하며 추가적으로 미국과 중국의 PMI 지수를 사용했다.

3. 모델링

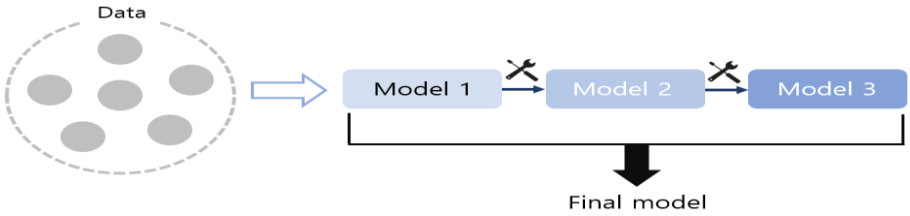
시장을 아웃 퍼폼할 수 있는 최적의 팩터 조합 학습

여러가지 머신러닝이 존재하지만, 우리는 그중 XG Boost라는 AI모델을 사용하였다. XG Boost는 분류(Classification)에 뛰어난 성능을 보이는 모델이다. 본 분석에서 XG Boost 모델은 **향후 1개월 시장대비수익률이 높을 것으로 예상되는 종목**들을 선정한다.

1) 모형 개요

XG Boost모델의 구체적인 구조는 본 보고서의 성격과 맞지 않으므로 간단하게 짚고 넘어가기로 한다.

[그림 10] XG Boost 설명



자료: 한국투자증권

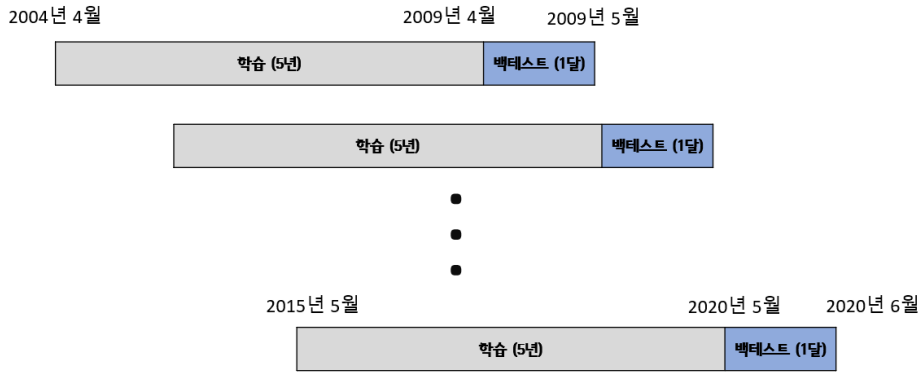
XGB모델의 여러가지 이점

XG Boost는 의사결정나무(Decision Tree)기반의 모델이다. 위의 [그림 10]과 같이 성능이 약한 모델을 단계적으로 발전시켜서 최종적으로 강력한 하나의 모델을 생성한다. 이 AI모델의 가장 큰 장점으로는 **우수한 예측 성능, 빠른 학습 속도, 유연성, 결측치의 자동 처리** 등을 꼽을 수 있다. 이러한 장점은 **결측치가 자주 발생하고 데이터 양이 방대한 금융 데이터에 이점을 가질 것으로 기대**한다.

2) 모델링 기간 설정 (백테스트 기간)

AI모형은 학습기간(Train)과 검증기간(Test)을 명확하게 설정해주는 것이 중요하다. 본 팀이 1차적으로 선정한 학습&검증 기간은 5년 학습, 1개월 검증이다.

[그림 11] 학습, 검증 기간 설정



자료: 한국투자증권

포트폴리오 리밸런싱 시점 기준 근 5년 정보 활용

학습기간의 구체적인 설정에 대해 많은 논의가 있었다. 일반적으로 **AI모형은 학습기간을 길게 할수록 다양한 상황을 학습**할 수 있기 때문에, 성능이 더 좋아지는 것으로 알려져 있다. 하지만, 금융시장의 특성상 **너무 먼 과거의 데이터는 현재 시점의 특성을 정확하게 반영할 수 없다**는 판단 하에 최종 학습기간은 5년으로

선정하였다. 오히려 불필요한 데이터를 추가했을 때 모델의 노이즈가 늘어날 우려가 있으며, 한국 금융 시장의 순환 주기 역시 4~5년 정도라는 사실을 참고했다.

모델 업데이트 주기와 포트폴리오 리밸런싱 주기는 '1개월'

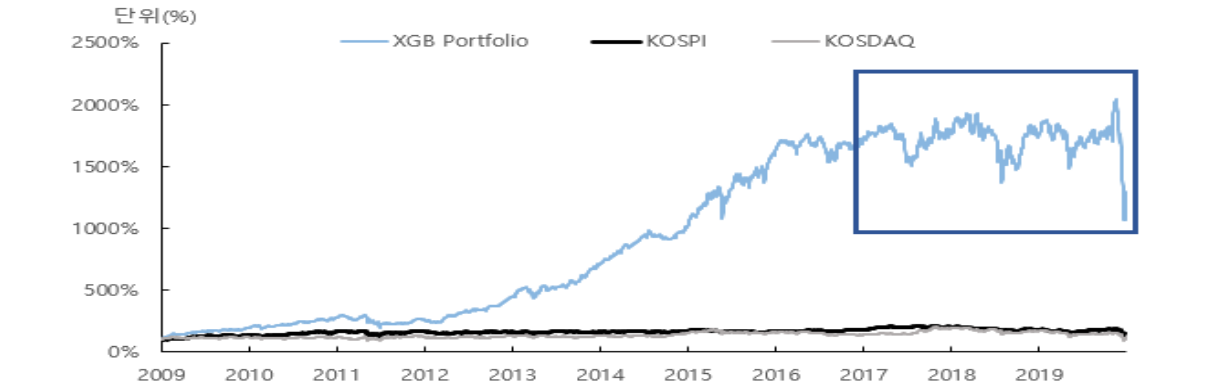
포트폴리오 선정 종목 수와 리벨런싱 방식은 다음과 같다. 5년의 학습을 마친 모형은 **이후 한 달에 한 번씩 시장대비 아웃퍼폼할 종목 상위 30개를 선정**한다. 첫 번째 생성한 모델로 1개월 동안 투자를 진행한 후, 1개월이 끝날 때 즈음에 업데이트된 데이터로 두 번째 모델을 생성해주는 방식이다. 즉, **모델 업데이트 주기와 포트폴리오 리밸런싱 주기는 모두 '1개월'**이다.

4. 백테스트 결과

1) 포트폴리오 성과 비교

백테스트 전체 기간에 대한 누적 수익률은 다음과 같다.

[그림 12] 1차 백테스트 결과 (누적수익률)



자료: 한국투자증권

2009년 이후 AI모델이 선정한 포트폴리오가 시장대비 상당한 초과수익률을 창출해낸 것을 확인할 수 있다.

2) 결과 피드백

AI 모델이 최근 저조한 성능을 보인 이유는?

하지만, [그림 12]의 노랑색 박스 안에 있는 최근 3년정도의 수익률은 그렇지 못하다. 이에 대한 원인은 크게 다음 2 가지로 파악된다. **1)팩터들이 투자자들에게 알려지기(well-known) 시작하면서 알파가 사라지는 현상**과 **2)거시경제 상황에 대한 이해의 부족** 때문이다.

따라서, 위의 2가지 문제를 해결하고자 본 팀은 다음과 같은 해결책을 제시했다.

해결책1. 매크로 거리를 고려하여 모델 개선
해결책2. AI 머릿속으로 들어가 보기

첫 번째 해결책은, 현재와 유사한 매크로 상황을 가진 과거 데이터에 가중치를 주는 것이다. 다시 말해, 모델이 현재 상황과 비슷한 과거 정보를 더 고려하는 방향으로 학습한다는 것이다. 아무리 AI가 똑똑해도, 현재 상황이 과거에 학습했던 데이터와 너무 다른 상황이라면 좋은 성능을 내기 어렵다. 이 해결책을 실행하기 위해 앞서 간단히 언급했던 매크로(거시 경제) 변수를 활용한다.

해석가능한 AI모델: LIME

두 번째 해결책은, 직접 AI의 머릿속으로 들어가보는 것이다. 서론에서도 언급했지만, **AI 모델의 가장 큰 단점은 해석하기가 어렵다는 점**이다. 이는 XG Boost를 비롯한 대부분의 모형이 블랙박스(Black-Box)모형이라고 불리우는 이유이기도 하다. 하지만, 본 팀은 **AI 모델을 해석하기 위해 LIME이라는 새로운 알고리즘을 적용하여 모델을 개선**하였다.

5. 모델 개선

1) 매크로 디스턴스 모델

매크로 상황의 유사성을 고려한 AI 모델 학습

본 팀의 AI모델을 바탕으로 총 12년간 백테스트를 진행했다. 구체적인 기간은 기간은 2009년부터 현재까지였지만, 위에서 살펴본대로 2017년 이후부터는 AI가 찾는 팩터 조합으로도 더 이상 유의미한 알파를 창출하지 못하고 있는 것을 확인할 수 있었다.

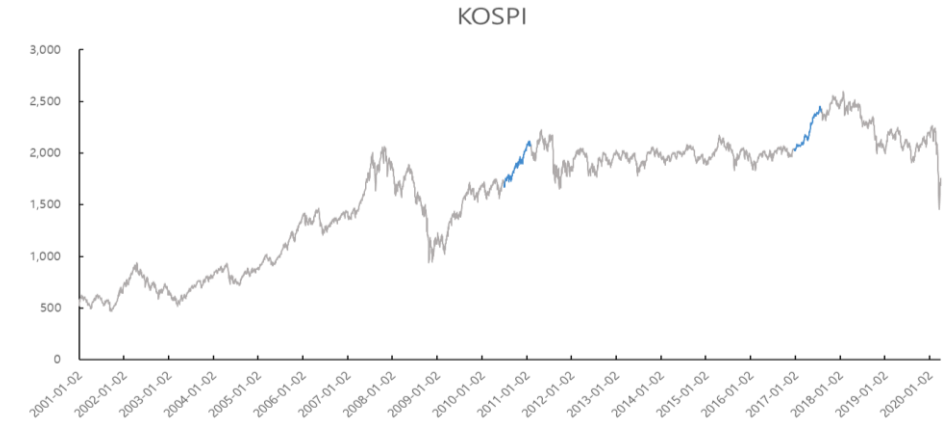
좋은 팩터라면 꾸준히 시장에서 알파를 창출할 수 있어야한다. 하지만 현실적으로 **상황에 관계없이 언제나 수익을 냈던 팩터는 존재하기 어렵다.** 매크로 상황을 고려한다는 것은 이와 같은 논리에서 출발한다. 꾸준히 수익을 냈던 팩터는 없을지라도, 특정 경기 국면에서 상대적으로 우위에 있는 팩터는 존재할 수 있다. **현재 매크로 상황과 비슷했던 과거 시점의 정보를 더욱 고려하여 유효한 팩터를 탐색**한다면 매크로 정보를 활용하지 않았을 때 보다 더 나은 수익률을 기대할 수 있다.

매크로 상황은 크게 시장 상황, 국내 금융 상황, 해외 금융 상황으로 구성

매크로 상황 반영은 다음과 같이 이루어진다. 크게 시장 상황(KOSPI, KOSDAQ 흐름), 국내 금융 상황(금리, 금리 스프레드, 신용 총량, 물가), 해외 금융 상황(주요국 금리, 환율 및 PMI)을 고려한다. 포트폴리오 리밸런싱 시점 기준으로 매크로 상황이 비슷한 과거의 데이터에 가중치를 주어 팩터 모델을 학습하며 이를 종목 선정에 활용한다.

시장 상황은 KOSPI와 KOSDAQ의 최근 흐름을 바탕으로 평가한다. 이 때 시장을 단순히 상승장, 하락장, 보합장으로 나누지 않고 시계열 흐름의 유사도를 활용한다. 자세히 설명하자면 특정 시점 기준 이전 150일 간의 KOSPI, KOSDAQ 흐름을 해당 시점의 시장 상황 지표로 판단한다는 뜻이다. 이렇게 표현된 시장 상황은 각 시점 간의 시장 상황 유사도를 계산하는 데에 사용된다.

[그림 13] 코스피 국면 유사도 예시



자료: 한국투자증권

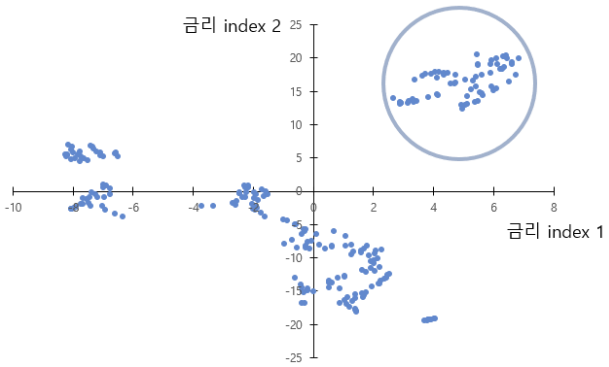
위의 그림은 2017년 8월 1일을 기준으로 이전 150일의 코스피 흐름이 가장 비슷한 국면을 탐색한 예시이다. (파랑색으로 색칠된 두 구간이 서로 유사하다고 탐색되었다.) 가장 코스피 흐름이 가까웠던 시점은 2011년 2월 1일이다.

같은 카테고리에 속하는 매크로 지표들의 경우 공통된 정보를 추출

국내 금융 상황과 해외 금융 상황의 경우 각 카테고리에 속하는 매크로 지표들의 수가 적지 않다. 때문에 모든 지표를 고려해서 경기 국면을 판단하기에는 무리가 있다. 예를 들어, 금리 카테고리에는 콜금리(1일물 중개거래), CD 유통수익률, 국고채 3년 금리 등이 포함 되어있는데 이를 모두 고려하여 매크로 유사도를 고려하기에는 직관적이지 못하다.

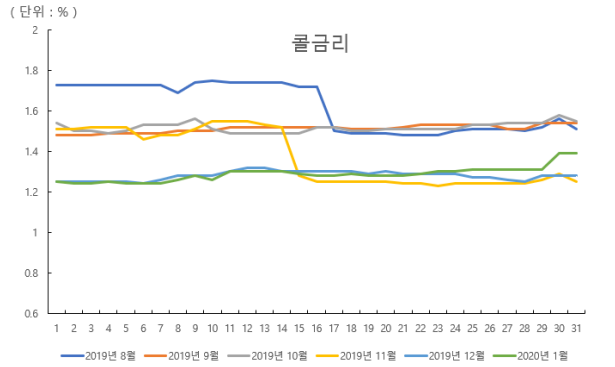
따라서 같은 카테고리에 속하는 매크로 지표들의 경우 PCA 라는 기법을 활용하여 공통으로 지니고 있는 주요 정보를 추출하여 사용한다. **PCA는 차원을 압축**시킬 때 사용될 수 있는 통계 분석 기법의 하나이다.

[그림 14] 금리 상황 인덱스



자료: 한국투자증권

[그림 15] 금리 유사 국면의 콜금리 흐름



자료: 한국투자증권

위의 두 그림 중 왼쪽은 지난 20년간의 금리 흐름을 2차원으로 압축해서 표현해 놓은 산점도이다. 한 달간 금리의 흐름을 하나의 점으로 표현하였다고 이해할 수 있다. 이렇게 여러 개의 금리 지표를 활용해 금리 상황 인덱스를 만들어 2차원으로 압축하게 된 결과, 각 시점의 금리 상황 유사도를 구할 수 있게 되었다. 간단히 말해 **가까운 점들이 한 달간 비슷한 금리 상황을 가지고 있는 시점들**이라고

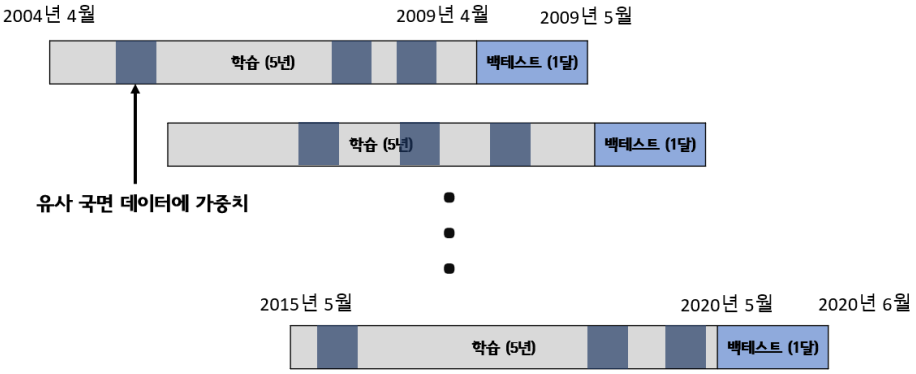
보면 된다.

흥미로운 점은 왼쪽 그림 우측 상단 원 안에 들어 가 있는 시점들이다. 최근 시점의 금리 상황은 거의 모두 원 안에 포함된 점들이다. 현 상황과 같은 저금리 기조가 이전에는 찾아볼 수 없는 상황이라고 이해할 수 있다. 오른쪽 그림은 원 안에 포함된 시점들의 콜금리 상황(한 달간 콜금리)이다.

매크로 상황이 비슷한 정보에 가중치를 두어 모델 학습

이렇게 모든 시점 간 시장 상황(KOSPI & KOSDAQ) 유사도, 국내 금융 상황 유사도, 해외 금융 상황 유사도를 계산한 뒤 종합적인 매크로 국면 유사도를 계산한다. 유사도라는 것은 일종의 거리 개념이라고 볼 수 있다. 이를 활용하여 [그림 16]과 같이 **모델을 학습시킬 때 현재 시점(포트폴리오 리밸런싱 시점)과 매크로 거리가 가까운 데이터에 가중치**를 준다.

[그림 16] 학습, 검증 기간 재설정



자료: 한국투자증권

2) 해석가능한 AI모형: LIME

AI 해석을 통해 모델 개선

AI 모델이 전 기간에서 좋은 주식을 골라 투자한다는 사실은 어느정도 확인할 수 있었다. 하지만 더 중요한 것은 '이 AI를 어떻게 믿을 수 있는가?'이다. 매 기간마다 모델이 왜 해당 종목들을 골랐는지 그 이유를 파악할 수 있다면, 이를 통해 최근 들어 부진한 팩터 성과를 개선할 수 있을 것이다.

이 부분은 **본 보고서의 가장 핵심적인 부분**이다. 이를 위해 'LIME'라는 통계 기법을 사용하였으며, 이 피드백을 바탕으로 기존 모델의 한계점을 개선해보기로 하였다. LIME 알고리즘은 모두 블랙박스 모형을 해석하기 위해 고안되었으며, 이미 여러 공신력 있는 학술지에 의해 그 타당성이 입증되었다.

입력 값에 따른 출력 값의 변화로 판단하는 모델의 논리

LIME은 우리가 설명하고자 하는 어떤 데이터와 유사한 여러 개 인공 데이터를 생성하고, 여기에 해석 가능한 선형 모형을 적합함으로써 해당 데이터를 해석하는 방식이다. 좀 더 쉽게 말하자면, **모델 자체는 블랙박스지만 모델의 입력 값을 바꾸어 가면서 출력 값이 어떻게 바뀌는지를 확인함을 통해 모델의 논리를 확인**할 수 있는 것이다. LIME 방식을 통해, 우리는 모델이 어떤 기준으로 탑30 종목들을 선정할 수 있었는지를 **개별 기업별로 파악**할 수 있다.

LIME의 동작 원리

<설명하려고 하는 데이터: A기업>

1. 관심있는 데이터인 A기업을 설명하기 유사한 인공 데이터를 n 개 생성한다.
2. 생성된 인공 데이터와 기존 데이터 간의 유사도를 구한다
3. 위에서 생성된 총 n+1개의 데이터를 기존 블랙박스 모형에 대입시켜 예측 확률을 뽑아낸다.
4. 이후 해당 데이터를 가장 잘 설명하는 m개 변수 조합을 추출
5. 4번에서 구한 m개 변수와 2번에서 구했던 유사도 행렬을 이용하여 최종 회귀모형 적합
6. 이 회귀식을 바탕으로 A기업에 대한 해석을 진행

자료: 한국투자증권

전체를 부분으로 분할해서 해석하는 LIME

LIME의 기본 동작 원리는 위와 같으며, **핵심은 비선형적으로 이루어져있는 전체 모델을 부분 부분으로 분할해서 선형적인 해석을 하는 것**이다. 이는 원의 둘레 길이를 구하기 위해 원을 작게 쪼개서 직선의 거리 구하는 원리다.

따라서 위의 위 알고리즘을 사용하여 **전체 백테스트 기간동안 AI가 어떤 기준으로 종목들을 골랐는지**를 알아보았다. 이를 통해 우리는 크게 2가지 사실을 알아냈다.

Fact 1. 거시 국면에 따라 모델이 중요하게 여기는 팩터들이 다르다.
Fact 2. 국면에 따라 중요하게 나왔던 변수들을 바탕으로 **종목 간의 비중을 조절**해 주었더니, 최근 포트폴리오 성과가 유의하게 개선되었다.

위의 두 가지 사실을 바탕으로 팩터 AI모델을 업그레이드했다.

Fact1. 거시 국면에 따라 모델이 중요하게 여기는 팩터들이 다르다.

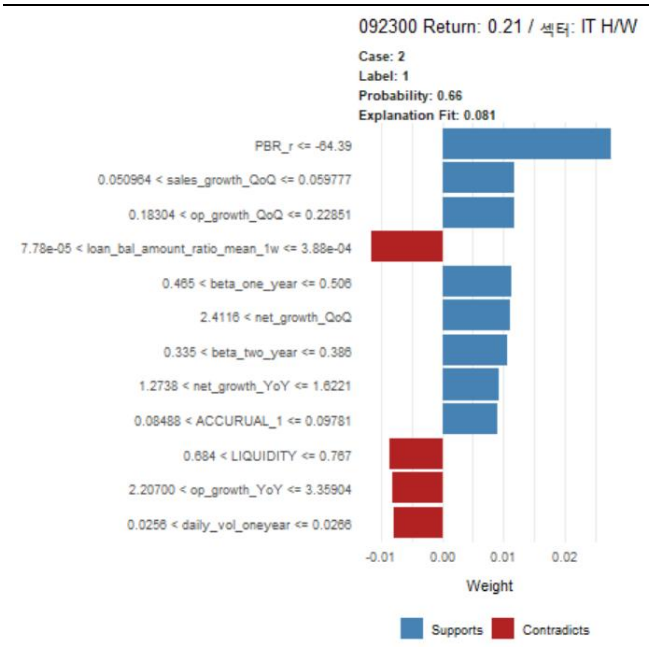
우선, 전체 코스피 시장을 아래 그림처럼 상승, 보합, 하락 크게 3개 국면으로 나누어서, **각 국면마다 모델이 중요하게 여기는 팩터들이 무엇인지 살펴보았다.** (여기서의 국면 분할은 앞선 매크로 거리 국면 파트와 관련이 없다. 코스피 장세에 따라 AI가 어떻게 생각하는지를 직관적으로 보기 위해 국면을 새로 나눈 것이다.)

[그림 17] 코스피 시장 국면 분리

자료: 한국투자증권

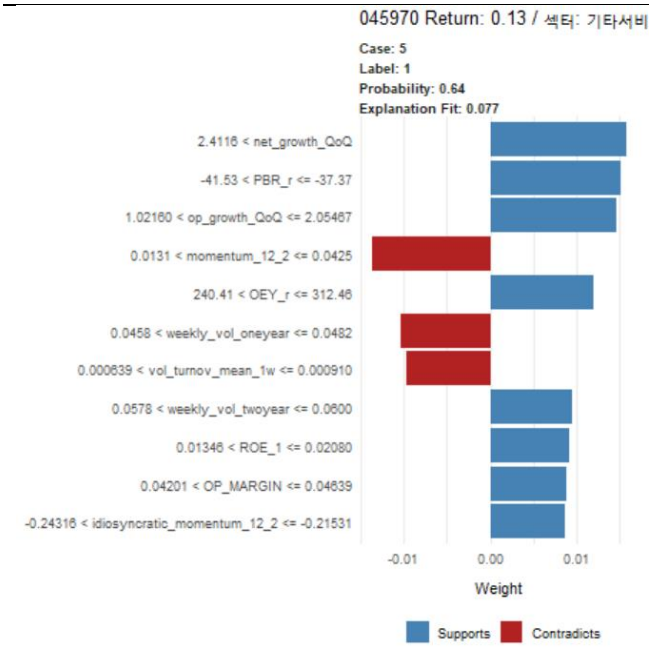
아래 [그림 18, 19]는 KOSPI가 본격적으로 박스권 구간 진입했던 2016년 7월 한 달 동안 AI모델이 아웃퍼폼할 것으로 예상했던 '현우산업(092300)'과 '코아시아(045970)'의 LIME 플랏이다.

[그림 18] 현우산업(092300) LIME플랏



자료: 한국투자증권, R

[그림 19] 코아시아(045970) LIME플랏



자료: 한국투자증권, R

[그림 19]과 [그림 20]을 통해 AI가 왜 두 종목을 선정했는지 이유를 파악할 수 있다.
위의 자료를 해석하는 방법은 다음과 같다. 일단 두 종목 모두 label 이 1 이기 때문에 시장 대비 아웃퍼폼할 것으로 예측된 종목이다. 그리고 **Supports(파란색)로 표시된 부분은 예측 결과에 긍정적인 방향으로 작용한 요인, Contracts(빨간색)은 예측 결과에 부정적인 방향으로 작용한 요인**이다.

두 종목 모두 1) PBR이 업종 평균 대비 30%이상 낮고, 2) 실적 성장률이 높은 점 등의 요소가 해당 종목들이 AI에 의해 선택되는데 크게 기여했음을 알 수 있다. 반면, 현우산업(092300)의 경우 대차거래 잔고비율이 낮았던 점과, 코아시아(045970)의 최근 12개월 주가 모멘텀이 부진했던 점은 AI가 두 종목을 선택하는데에 부정적으로 작용했음을 알 수 있다.

이처럼, LIME을 통해, AI모델이 '해당 종목을 선정하게 된 이유'를 직관적이고 쉽게 파악할 수 있다. 다시 말해, **LIME을 사용해서 AI모형이 어떻게 사고하고 판단하는지 파악할 수 있는 것**이다.

이와 같은 방식으로, 나머지 국면에 대해서도 AI가 어떤 팩터를 중요하게 생각하는지를 파악해 보았고, 이를 다음과 같이 표로 정리할 수 있었다.

<표 4> 국면별 AI가 중요하게 여겼던 팩터

국면 1(상승)	국면 2(하락)	국면 3(보합)
거래량회전율	일별 주가변동성	동일업종평균대비 PBR
1 개월주가모멘텀	주간 주가변동성	영업이익성장률
대차거래체결비율	시장베타	매출액성장률
신용융자신규가입비율	공매도잔고비율	자기자본대비이익
매출액대비영업이익	공매도체결비율	매출액대비영업이익

국면에 따라 주요하게 활용되는 팩터가 변함

보합장에서는 주로 벨류와 퀄리티 팩터들이 중요하게 여겨졌고, 하락장에서는 변동성(Volatility)관련 팩터들이 주요하게 다뤄졌다.
이를 통해, 모형을 설계할 때 **국면을 고려해주는 것**의 필요성을 확실히 인지할 수 있었다.

Fact2. 국면에 따라 중요하게 나왔던 변수들을 바탕으로 종목 간의 비중을 조절해 주었더니, 최근 포트폴리오 성과가 유의하게 개선되었다

비슷한 국면에서 주요하게 활용되는 팩터를 활용하여 포트폴리오 내 종목 비중 조절

현재까지 백테스트를 진행할 때는 종목간의 비중은 전부 동일하게 가져갔다. 예를 들어, AI가 이번 달에 투자할 30개의 종목을 골랐다면, 모든 종목 간의 투자 비중은 동일하게 유지했던 것이다. 하지만, 본 팀은 **AI의 사고 과정을 해석할 수 있는 장점을 보다 적극적으로 활용하기 위해, AI가 중요하게 여겼던 팩터들을 바탕으로 종목간의 비중을 다르게 해서** 2차 백테스트를 진행했다.

종목간의 비중을 다르게 하는 과정은, 우선 현재 투자하는 시점과 가장 유사하다고 판단되는 과거 시점을 찾는다. 그리고, 해당 시점에서 AI가 종목을 고를 때 중요하게 판단했던 팩터를 기준으로 종목간의 비중을 조절해준다.

예를 들어, 투자하는 기간이 2020년 6월 한 달 동안이라고 한다면, 먼저 가장 유사한 과거 국면을 탐색한다. (가장 유사했던 국면은 2014년 7월이라고 하자) 이 때 AI가 종목을 고를 때 중요하게 여겼던 팩터는 'idiosyncratic Momentum_6_2'이다. 이를 기준으로, AI가 골랐던 30개 종목에 대한 비중을 조절해준다.

위와 같은 방식으로 2017년부터 2차 백테스트를 진행하였고, 그 결과는 [그림 20]에 나타나있다.

[그림 20] 개선된 모형 백테스트 결과(누적 수익률)

자료: 한국투자증권

위와 같이, 개선된 AI모델이 1차 모델에 비해 훨씬 더 안정적이고 높은 수익률을 제공할 수 있게 되었다.

3) 2020년 6월 추천 종목

앞으로의 전략

지금까지 해석가능한 AI모델을 통한 본팀의 퀀트 투자 전략에 대해 소개했다. 이를 바탕으로 6월 한 달 동안 아웃퍼폼할 것으로 기대되는 상위20개 종목이다.

종목코드	종목명	섹터	마켓
42700	한미반도체	기계	코스피
214680	디알텍	의료 · 정밀기기	코스닥
990	DB 하이텍	전기전자	코스피
11070	LG 이노텍	전기전자	코스피
47810	한국항공우주	운수장비	코스피
100120	뷰웍스	의료 · 정밀기기	코스닥
139670	키네마스터	IT S/W & SVC	코스닥
251370	와이엠티	화학	코스닥
9420	한올바이오파마	의약품	코스피
64290	인텍플러스	IT H/W	코스닥
131390	피앤이솔루션	일반전기전자	코스닥
242040	나무기술	출판 · 매체복제	코스닥
178920	SKC 코오롱 PI	화학	코스닥
5950	이수화학	화학	코스피
40910	아이씨디	IT H/W	코스닥
11170	롯데케미칼	화학	코스피
207940	삼성바이오로직스	의약품	코스피
140860	파크시스템스	의료 · 정밀기기	코스닥
12610	경인양행	화학	코스피
83450	GST	IT H/W	코스닥

III. 결론

해석가능한 AI 모형을 통한 퀀트 투자 전략

활용 방안 및 기대효과

지금까지 해석가능한 AI모형을 통해 시장 지수 대비 상당한 초과 수익률을 창출할 수 있는 퀀트 모형에 대해 설명하였다.

AI 팩터 모형의 핵심적인 특징은 **1)해석 가능한 인공지능과 2)매크로 정보 활용**이다. 이 두 가지 특징은 다음과 같은 기대효과를 갖는다.

첫 번째, **액티브 펀드 매니저들의 운용에 도움을 줄 수 있을 것**으로 기대한다. 그저 오를 종목을 추천하기보다 모델이 어떤 논리로 해당 종목이 오를 것이라고 예측했는 지의 정보를 제공하기 때문이다.

두 번째, **매크로 정보 활용의 다양성**이다. 본 모델은 크게 시장 상황, 국내 금융상황, 해외 금융상황을 종합적으로 고려했지만 각 상황을 얼마만큼의 비중으로 고려할지는 다양한 방면으로 선택할 수 있다.