글로벌퀀트 R을 활용한 퀀트전략-(1) 판별분석

Global Quant
박우열
☎ (02) 3772-2324
✉ wooyeol.park@shinhan.com

- R을 활용한 데이터 수집과 시각화 및 통계처리의 유용성
- 판별분석에 사용할 변수 선정 및 백테스팅
- 증시 스타일 / 경기 국면 / 물가 국면 판별

투자자가 R 패키지 활용을 고려할 수 있는 이유

R패키지를 활용한 데이터 확보 기능의 유용성

투자자가 R 패키지를 활용해야 할 세 가지 이유는 1) 금융 데이터 수집, 2) 금융 데이터 수치적 가공, 3) 시각화에 갖는 압도적 장점 때문이다. 투자자가 직접 데이터 처리 로직을 구현하지 않고도 오픈소스를 통해 금융 데이터 분석의 효율성을 크게 높일 수 있다. 의미있는 데이터들이 범람하는 시대에 효율성과 생산성이 가진 무기는 더욱 높아질 전망이다. 우리는 여기에 주목해 R을 활용한 퀀트 전략을 소개해 나가려고 한다. 먼저, R이 가진 장점을 살펴보자.

1) 첫 번째는 오픈소스 금융 데이터 확보다. 블룸버그 등 유료 금융 데이터에 접근하기 어려운 개인투자자도 거래소나 야후 파이낸스에서 데이터를 확보할 수 있다. 전문투자자도 사무실 밖의 환경에서 데이터를 수집할 수 있다. 또한, R를 활용한 데이터 수집 자동화가 가능해 업무 효율성을 매우 높여준다.

R패키지를 활용하는 구체적인 예시

또한, 유료 계정에서도 접근이 쉽지 않은 대체 데이터를 크롤링할 수도 있다. 코스피의 편/출입을 고려한 장기 시계열 혹은 오픈소스로 제공되는 ESG 스코어, 특정 원자재와 관련된 날씨 데이터 등이다. 또한, 미국 기업들의 각종 재무상태표를 미국 증권거래위원회 홈페이지에서 더 짧은 주기로 직접 받아올 수도 있다.

구체적으로, 데이터 풀링에 가장 활용도가 높은 패키지 중 하나는 quantmod이다. 이를 활용해 데이터를 받고 지난번 자료에서 제시했던 미국 섹터ETF를 이용해 최적화 포트폴리오를 구축하기 위해 필요한 자산 비중을 시각화해봤다.

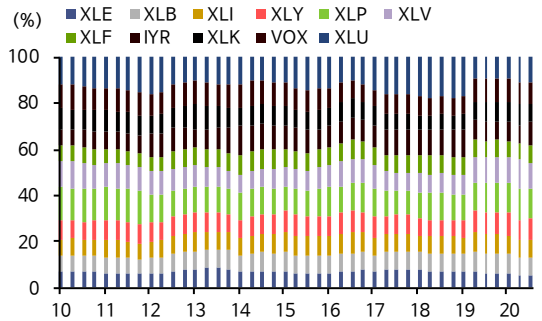
R을 활용한 데이터 수집과 백테스팅 코드 예시

```
# download price data
tick<-c("XLE", "XLB", "XLI", "XLY", "XLP",
"XLV", "XLF", "XLK", "XLC", "XLU")
US_ETF_Sector<-stockDataDownload(tick, from
="2010-01-01", to="2021-09-10")

# risk parity portfolio
risk_parity <- function(dataset) {
  prices <- dataset$adjusted
  log_returns <- diff(log(prices))[-1]
  return(riskParityPortfolio(cov(log_returns),
b=NULL, mu=NULL,w_lb=0)$w)
}
```

자료: R, 신한금융투자

포트폴리오 최적화에 필요한 자산별 비중 시각화



자료: R, Yahoo Finance, 신한금융투자

R패키지를 활용한 데이터 확보 기능의 유용성

2) R은 엑셀만으로 처리하기 힘든 빅데이터의 처리와 가공에 용이하다. 엑셀에도 상관분석이나 회귀분석을 해주는 고급기능은 존재한다. 하지만 제약조건을 추가한 정규화 회귀나 로지스틱 회귀분석, 주성분 분석을 활용해 차원축소를 하거나 예측 모형을 만들 수 있는 기계학습 기법을 활용하기에는 R이 더 적합하다.

3) 엑셀로 구현하기 힘든 형태의 다양한 시각화 모듈을 제공하고 심지어 애니메이션으로 시각화할 수 있다. 예를 들어, 본 보고서에서는 두 개의 다른 집단에서 나온 산점도 데이터를 동일축에 겹쳐서 표현할 때 R의 시각화 기능을 사용했다.

R을 활용한 판별 분석 소개

실제 금융 시계열 데이터를 이용한 판별분석 원리 설명

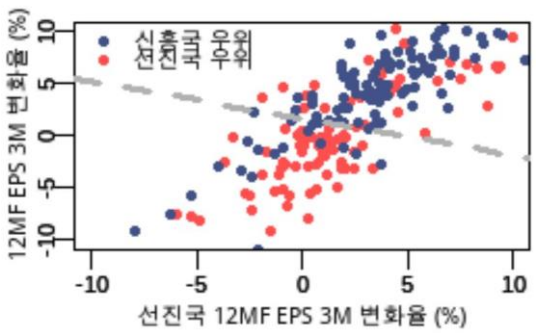
R의 통계 패키지를 활용해 금융 시계열 데이터로 투자 유망 지역과 경기를 판단하는 모델을 만들었다. 구체적으로, R에서 제공하는 판별분석 패키지를 사용한 뒤 앙상블 기법으로 결과값을 강화했다. 우선, 모델에 투입할 설명변수를 사용해 판별분석이 어떻게 작동하는 것인지 그 원리를 설명하겠다.

첫 번째로 왼쪽은 선진국과 신흥국의 12개월 선행 EPS 3M 변화율의 산점도 그래프인데 두 변수 간에 양의 상관관계가 있음을 알 수 있다. 산점도만을 보면 글로벌 이익 추정치 변화율에 동조화되고 있다고 해석할 수 있다.

그런데 이를 다음달 신흥국과 선진국 중 어느쪽이 수익률이 좋았는지를 기준으로 파랑/빨강색으로 두 집단으로 구분하면 새로운 정보를 얻을 수 있다. 즉, 글로벌 경기가 좋아서 모든 국가들의 이익추정치가 상향되는 시점에는 그 다음달 신흥국의 성과가 더 좋은 경우가 많았다는 것이다.

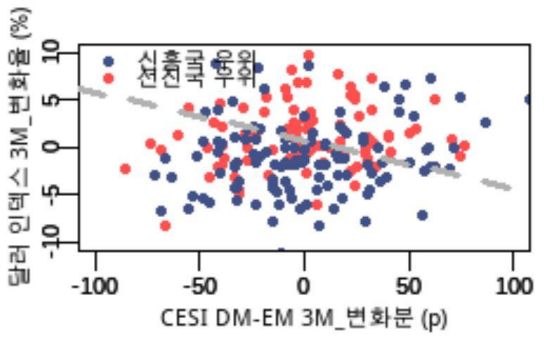
다음으로 오른쪽은 CESI의 선진국-신흥국 값의 3개월 변화율과 달러 지수의 3개월 변화율의 산점도이다. 산점도가 둥근 형태이므로 경기와 달러 방향성 간에 상관관계는 없다. 그런데 마찬가지로 데이터를 구분하면 새로운 사실을 알 수 있다. 이미 널리 알려진 사실이지만 달러가 약하고 신흥국 경기가 상대적으로 좋은 시점의 다음 달에는 나은 신흥국 성과를 기대할 수 있었던 것이다.

선진국과 신흥국 12MF EPS 3M 변화율 산점도



자료: Refinitiv, R, 신한금융투자

CESI_3M 변화율과 달러 지수 3M 변화율 산점도



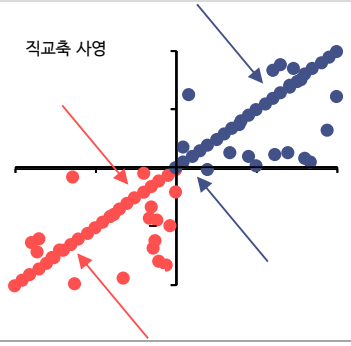
자료: Refinitiv, R, 신한금융투자

이 때, 위와 같이 회색 점선으로 두 집단을 구분하 것을 생각해볼 수 있다. 구체적으로, 이렇게 구분선을 그은 뒤 직교하는 축에 모든 점을 사영(projection)해 데이터를 두 집단으로 나눴을 때 동시에 겹쳐지는 데이터를 최소화해주는 기법이 바로 선형판별분석(LDA:Linear Discriminat Analysis)의 기본 아이디어다.

그런데 LDA는 결정 경계선을 선형으로 가정하기 때문에 공분산 구조가 상이한 데이터에 취약하다는 단점이 있다. 쉽게 말해 직선으로 딱 잘라서 두 집단을 구분하기 힘든 경우가 있을 수 있다는 것이다. 이때는 직선이 아니라 구불구불한 곡선을 사용해 구분하는 것을 생각해볼 수 있다.

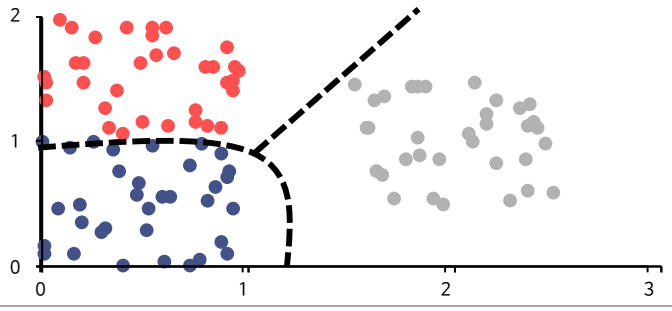
직선 대신 곡선을 사용해 구분하는 것은 이차판별분석(QDA)이다. 다만 유의미한 결과를 얻기 위해 요구되는 데이터 양이 더 많다. 따라서 장기 시계열에는 QDA를 사용하고 단기 시계열에는 LDA를 사용한 분석을 고려해볼 수 있다.

산점도 데이터의 직교축 사영



자료: Refinitiv, R, 신한금융투자

비선형 구분자를 사용한 이차판별분석 시각화



자료: Refinitiv, R, 신한금융투자

판별분석에 활용하기 위한 변수선정

경기, 이익, 달러, 금리, 수급 설명변수 선정 기준

구체적으로, 판별분석을 활용해 현재 시점에 대한 경기 판단 및 투자하기에 유망한 지역을 고르고자 한다. 경기, 이익, 달러, 수급, 금리 5개를 설명변수로 사용해 현재 시점의 국면을 판단하고 다음달 성과가 좋을 지역을 골랐다.

동일 분류 내에서는 독립 변수간의 상관관계가 높기 때문에 유의미한 지표들만을 선정했다. 상관관계가 높은 독립변수를 많이 사용하면 다중공선성 문제(Problem of multicollinearity) 발생해 모형의 설명력이 떨어지기 때문이다.

예를 들어, 경기지표 중에서는 OECD 경기선행지수의 설명력이 높지만 사후적으로 데이터가 수정되는 경향이 있고 지수 자체에 주가가 포함되어 있으며 발표주기도 2달이나 늦기 때문에 이를 설명변수에서 제외했다.

펀더멘털 지표 중에는 매출 추정치가 높고 수익성이 높으면 당연히 이익도 좋을 것이기 때문에 이익 추정치만을 활용해 가공했다. 동일한 카테고리 내에서는 적시성을 높이기 위해 발표 주기가 짧은 변수에 우선순위를 두었다.

밸류에이션 지표는 금리와 역의 상관관계에 있고 재무비율 자체에 밸류에이션이 포함되어 있기 때문에 모형에서 제외했다. (ROE= PBR/PER) 구체적으로, 변수 처리와 지표 입수 시점을 고려한 판별분석 모델은 다음과 같다.

판별분석 모델에 사용할 독립변수와 종속변수 설명

종속변수	관측	예측	독립변수	관측	차분	설명
R(DM-EM)	i기	i+1기	CESI(DM-EM)	i기	3	선진국과 신흥국 CESI 관측값 3개월 차
OECD CLI	i-2기	i기 ~	이익(DM-EM)	i기	3	선진국과 신흥국 12개월 선행 이익 추정치 관측값 3개월
			달러	i기	3	달러 관측값 3개월 차
			ETF_FF	i기	3	ETF 수급 관측값 3개월 차
			US10Y	i기	3	미국 10년물 금리 관측값 3개월 차

자료: Refinitiv, 신한금융투자

판별분석 모델1: 선진국 vs 신흥국 지역 판별

선진국vs신흥국 투자 유망한 지역 골라내기

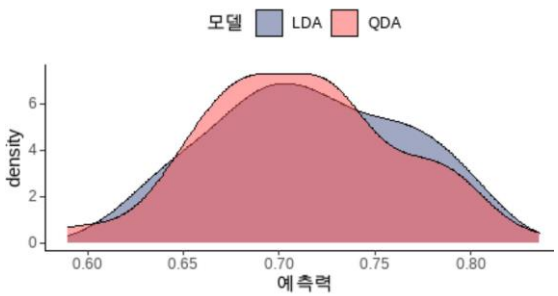
독립변수의 시계열 길이가 일정하지 않기 때문에 시계열이 짧은 변수를 일부 제외하여 2000년 1월부터 장기 백테스팅했다. 판별분석에 사용할 트레이닝 셋과 테스트 셋은 7:3 비율로 추출했고 시계열 내 시간의 흐름을 고려하기 위해 설명변수는 모두 차분값이나 변화율 값을 사용했다.

판별분석 모델별 반복 시뮬레이션 결과

트레이닝 셋이 무작위로 추출되기 때문에 판별분석을 수행할 때마다 예측력이 매번 다르다. 때문에 판별분석을 여러 번 수행했을때도 유의미한 예측력을 갖는지 100회, 1000회 반복해 시뮬레이션했다. 반복횟수가 높아질수록 LDA와 QDA 모델의 예측력이 서로 수렴했으며 평균치를 기준으로 종모양 분포가 되었다.

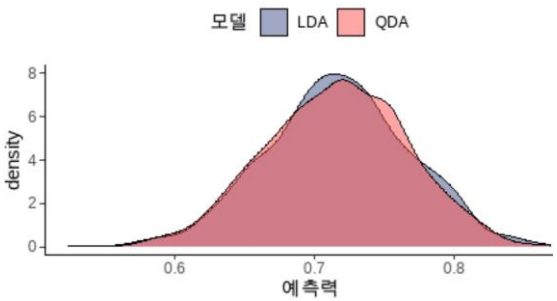
결과적으로, 판별분석 1,000회 반복값의 평균 판별력은 72%이었고 최소값은 61%, 최대값은 82%, 표준편차 5% 에서 다음달 수익률이 높을 지역을 판별해낼 수 있었다. 마지막으로, i-1기 까지의 판별변수를 사용해 다음달 선진국과 신흥국 중 수익률이 높을 지역을 골랐다. 이번달에는 선진국이 상승할 확률이 높다.

판별분석 100회 반복 예측력 분포



자료: Refinitiv, R, 신한금융투자

판별분석 1,000회 반복 예측력 분포



자료: Refinitiv, R, 신한금융투자

판별분석 모델2: 경기 국면 판단모델

경기, 이익, 달러, 금리, 수급 5가지 설명변수만을 사용한 이유

경기 국면에 대한 판단은 스타일과 지역별 자산배분의 핵심 변수이다. 다만 공식적인 지표 발표가 2달 늦기 때문에 실제로는 설명변수로 사용하기 어렵다. 따라서 판별분석을 활용해 지금 경기에 대한 판별을 해봤다. 지금 경기 국면을 미리 알 수 있다면 향후 투자할 지역과 스타일에 대해 유용한 정보를 제공할 것이다.

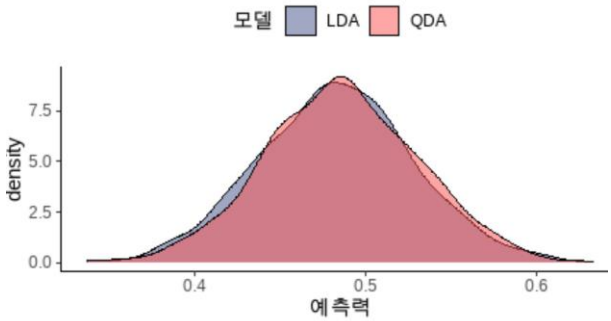
OECD경기선행지수는 통상 4개 국면으로 분류하는데 판별분석의 평균적인 적중률이 50% 수준이었다. 4가지 가능한 결과에 대한 무작위 예측값은 25%이므로 50% 까지 올리는 것도 의미는 있다. 하지만 이를 더 개선시킬 방법은 없을까?

판별분석 결과: 분류 매트릭스

		예측			
		1국면	2국면	3국면	4국면
실제	1국면	14	9	0	1
	2국면	8	15	2	5
	3국면	1	10	12	2
	4국면	4	7	3	5

자료: Refinitiv, R, 신한금융투자

판별분석 1,000회 반복 실행 결과 예측값의 분포



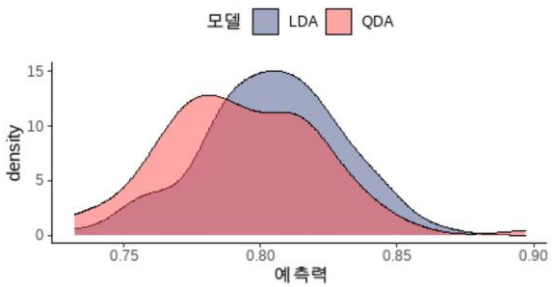
자료: Refinitiv, R, 신한금융투자

경기판단은 가장 최근 관찰된 경기선행지수 값을 기준으로 이번 국면이 유지될지 다음 국면으로 넘어갈지로 분류하는 문제이다. 즉, 국면전환 시 다른 국면으로 급격하게 바뀌는 것이 아니라 국면별로 순서가 존재한다.

국면을 이진 분류하여 판별하면 예측력이 크게 향상

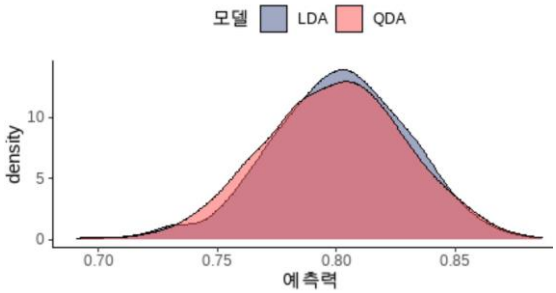
따라서 가능한 결과값이 4개가 아니라 (국면유지,국면전환)이라는 두 가지 결과 중에 하나를 판별하는 이진 분류로 축소가 가능하다. 종속변수를 2개로 축소한 뒤 판별분석을 수행했더니 예측력이 크게 올라갔다. LDA와 QDA 모델 시뮬레이션 예측력은 반복할수록 유사해졌으며 평균적인 예측력은 81% 수준이었다.

판별분석 100회 반복 예측력 분포



자료: Refinitiv, R, 신한금융투자

판별분석 1,000회 반복 예측력 분포



자료: Refinitiv, R, 신한금융투자

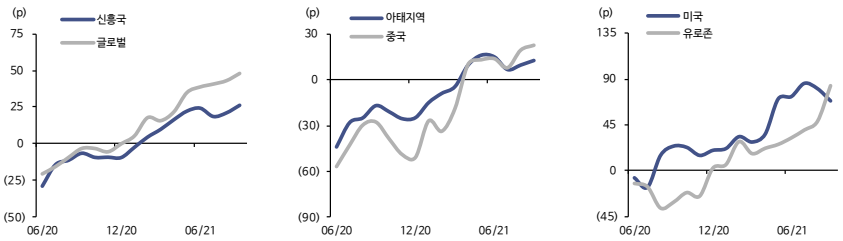
판별분석 모델3: 물가 판단모델

물가 국면 판단 모델

물가에 대한 관심이 급증하고 있다. 공급망 차질로 각종 원자재 가격 상승이 상승하고 있으며 유가는 연중 최고치를 돌파했다. 천연가스 가격 또한 유럽 친환경 규제, 러시아의 수출 감소 등의 이유로 급등하고 있다. 유럽의 봄철이 평년보다 추웠고 아시아의 여름은 평년보다 더웠기에 에너지 수요가 늘었던 것 또한 이유였다. 이제는 곧 겨울철이 다가오고 있어 난방 수요가 더 늘어날 전망이다.

이에 따라 각국 인플레 서프라이즈 지수도 꾸준히 우상향 중이다. 물가가 급등하면 원가상승을 판가에 전가시킬 수 있는지에 따라 산업별 수혜가 엇갈릴 수 있어 물가에 대한 판단은 증시에 접근하는 데 있어 매우 중요한 변수이다.

Citi 인플레이션 서프라이즈 지수 추이



자료: Refinitiv, 신한금융투자

이에 따라 주요 원자재 가격을 독립변수로 사용해 뉴욕 FED에서 월간으로 발표하는 UIG(UIG:Underlying Inflation Gauge)의 방향성을 판별할 수 있을지 모델링해봤다. 구체적으로, 모델링에 사용한 독립변수와 종속변수는 다음과 같다. 시계열은 S&P 케이스실러 주택가격지수의 시계열이 시작하는 2001년 4월부터 월간 데이터를 사용했다.

판별분석 모델에 사용할 독립변수와 종속변수 설명

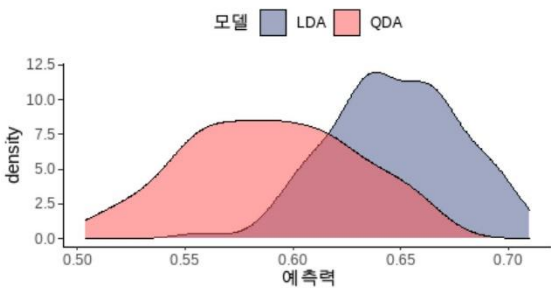
종속변수	관측	예측		독립변수	관측	차분	설명
UIG	i-1기	i기		GSCI 원자재지수	i기	3	3개월 변화율
			~	S&P 케이스실러	i기	12	12개월 변화율
				WTI가격	i기	3	3개월 변화율
				미국 실업률	i기	0	레벨값

자료: Refinitiv, 신한금융투자
주) UIG(Underlying Inflation Gauge): 뉴욕 FED에서 월간으로 집계하는 기저 인플레이션 지수

그 결과는 다음과 같았다. 시뮬레이션 숫자가 적은 경우 LDA와 QDA 분포에 차이가 있는 것은 독립변수 간에 공분산 구조가 상이한 것으로 해석할 수 있다. 하지만 마찬가지로 시행 횟수를 늘릴수록 LDA와 QDA의 예측력이 수렴했는데

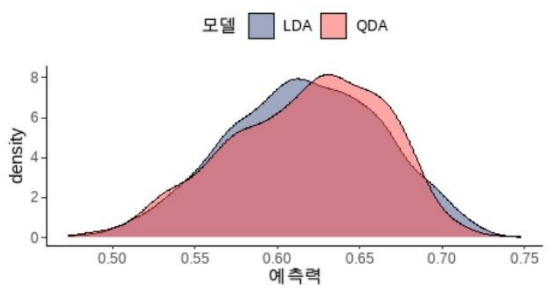
최종적으로 LDA의 평균적인 예측력은 65% 수준이었다.

판별분석 100회 반복 예측력 분포



자료: Refinitiv, R, 신한금융투자

판별분석 1,000회 반복 예측력 분포



자료: Refinitiv, R, 신한금융투자

결론

선진)신흥

경기국면: 유지

물가: 상승

R을 활용할 경우 얻을 수 있는 이점이 많다. 다양한 금융 시계열 데이터 확보에 있어 편의성을 제공하고 그 외에도 비정형 데이터, 대체 데이터 수집에 활용할 수 있다. 엑셀에서는 구현하기 힘든 각종 통계 패키지와 시각화 기능을 고려하면 R을 사용하지 않을 이유가 없다.

이번 보고서에서는 R에서 제공하는 여러 기능 중 판별분석을 활용해 증시 스타일, 경기 국면, 물가 방향성에 대한 모델을 만들었다. 결론적으로 이번달 증시는 선진국이 신흥국보다 성과가 좋을 확률이 높다. 경기는 현재 국면이 계속되는 가운데 물가는 올라갈 전망이다. 따라서 현재 국면에서 물가가 상승할 때 수혜를 볼 수 있는 업종과 스타일을 선별하는 접근이 유효하다.

Compliance Notice

- 이 자료에 게재된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다(작성자: 박우열).
- 본 자료는 과거의 자료를 기초로 한 투자참고 자료로서, 향후 주가 움직임은 과거의 패턴과 다를 수 있습니다.
- 자료 제공일 현재 당사는 상기 회사가 발행한 주식을 1% 이상 보유하고 있지 않습니다.
- 자료 제공일 현재 당사는 지난 1년간 상기 회사의 최초 증권시장 상장시 대표 주관사로 참여한 적이 없습니다.
- 자료제공일 현재 조사분석 담당자는 상기회사가 발행한 주식 및 주식관련사채에 대하여 규정상 고지하여야 할 재산적 이해관계가 없으며, 추천의견을 제시함에 있어 어떠한 금전적 보상과도 연계되어 있지 않습니다.
- 당자료는 상기 회사 및 상기회사의 유가증권에 대한 조사분석담당자의 의견을 정확히 반영하고 있으나 이는 자료제공일 현재 시점에서의 의견 및 추정치로서 실적치와 오차가 발생할 수 있으며, 투자를 유도할 목적이 아니라 투자자의 투자판단에 참고가 되는 정보제공을 목적으로 하고 있습니다. 따라서 종목의 선택이나 투자의 최종결정은 투자자 자신의 판단으로 하시기 바랍니다.
- 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 어떠한 경우에도 당사의 허락없이 복사, 대여, 재배포될 수 없습니다.