이벤트 기반 주가예측 서비스 시스템

[아기호랑이]팀 민찬홍, 백현우, 이수빈

목차

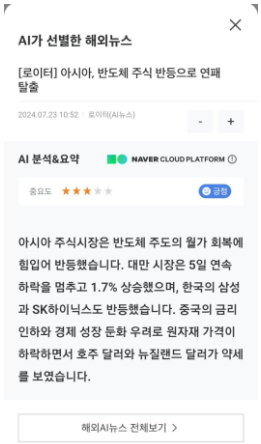
I. 서론

1. 주제선정 배경 및 필요성

최근 금융 시장은 기술의 급속한 발전과 함께 빠르게 변화하고 있다. 특히, 생성형 인공지능(Generative AI)의 도입은 금융 서비스 분야에 혁신적인 변화를 가져오고 있으며, 사용자들에게 맞춤형 정보 제공과 개인화된 투자 전략 제안을 통해 큰 가치를 제공할 수 있는 잠재력을 가지고 있다. 생성형 AI 는 방대한 금융 데이터를 분석하고 패턴을 인식하여 사용자가 놓치기 쉬운 중요한 정보를 실시간으로 제공할 수 있다.

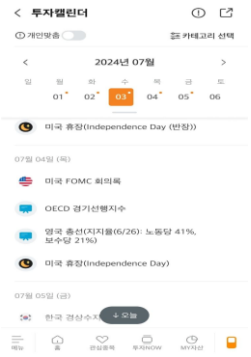
현대의 투자자들에게 신속하게 중요 데이터를 선별 및 파악하는 능력은 필수적이다. 그러나 수 많은 정보들 중 어떤 정보가 중요한지, 어떻게 해석해야 하는지에 대한 어려움은 해결되지 않은 것이 현실이다. 이에 AI 를 이용하여 실적 발표나 경제 지표 발표와 같은 중요한 이벤트를 감지하고, 이를 기반으로 주가 변동 예측 및 투자 전략을 제안하여 투자자들이 신속하고 정확한 의사결정을 할 수 있도록 돕는 서비스를 제안하고자 한다. 여기에 미래에셋증권 사용자들에게 맞춤형 알림 기능을 결합한다면 신속성 또한 확보할 수 있을 것이라 자신한다.

2. 기존 서비스의 한계



1) 이벤트와 주가 간 관계 정보 부족

미래에셋증권 앱에서 제공하는 ai 가 선별한 뉴스의 경우 단순히 긍/부정 제시 형태로 해당 이벤트가 어떤 근거로 어떤 주식에 영향을 미치는지에 대한 정보가 부족하여 설명력과 신뢰성이 떨어진다. 특별히 투자 시 가장 중요하다고 할 수 있는 주가에 대한 영향에 대한 설명은 없다.



2) 캘린더 내용 부족

캘린더 내 정리된 이벤트의 경우 실적 발표, 공모주, 신규 상장 등과 같은 이벤트는 있으나 신제품 출시와 같은 주가에 영향을 미치는 일부 중요 이벤트가 생략된 경우가 있었다.

3) 개별 이벤트의 영향 분석 부족

위의 사진에서 해당 날짜에 진행되는 다양한 주식 이벤트가 있긴 하나 미국 FOMC 와 같은 주식 이벤트가 무엇이고, 발표 결과는 어떠며 해당 이벤트가 주식 시장에 어떠한 영향을 미치는지에 대한 정보는 언급되고 있지 않다.

3. 데이터 수집 및 전처리 방안

1) 데이터 수집

본 프로젝트에서는 과거 주가 데이터, 전문가 예측 의견, 실적 발표 결과, 금융 용어, 신제품 리뷰 데이터 등을 수집하였다.

1)-1. 과거 주가 데이터

다양한 주식의 과거 주가 데이터를 수집하기 위해 웹 크롤링 기법을 사용하였다. 이때, 데이터는 PyKrx 모듈 API 를 사용하여 추출하였다. (https://github.com/sharebook-kr/pykrx?tab=readme-ov-file)

수집된 데이터는 CSV 파일로 변환하여 저장하였다.

1)-2. 기업 공시자료 데이터

본 프로젝트에서는 DART(전자공시시스템) (https://dart.fss.or.kr/main.do) 의 공시자료를 크롤링하여 데이터를 수집하였고, 이를 기업의 실적 변화와 이에 따른 주가 변동을 분석하는 데 사용하였다.

1)-3. 과거 주요 경제지표 데이터

과거 주요 경제지표 데이터통계청, 한국은행 등의 공공 데이터 포털과 국제금융센터(KCIF) (https://www.kcif.or.kr/)에서 제공하는 경제지표 데이터를 활용하였다. 주요 경제지표에는 실업률, 소비자물가지수(CPI), 금리, 경상수지, 소비자신뢰지수 등을 선정하였으며, 여러 지수 중 주가에 미치는 영향의 크기를 주로 고려하여 총 6 개의 변수를 선정하였다. 2021 년 6 월 1 일부터 2024 년 7 월 30 일까지의 3 년 2 개월 간의 데이터를 수집하였고, 이렇게 수집된 데이터는 주가 변동과의 연관성을 분석하고 예측 모델에 통합하여 사용하였다.

1)-4. 신제품 리뷰 데이터

신제품에 관한 리뷰 데이터를 가상의 기업을 가정하여 ChatGPT 를 이용하여 생성하였다.

2) 데이터 전처리

수집된 데이터는 다양한 형식과 구조를 가지고 있기 때문에, 분석에 적합한 형태로 전처리하는 과정이 필요하다. 데이터 전처리는 데이터의 품질을 높이고, 분석 결과의 신뢰성을 확보하기 위해 중요한 단계이다.

2)-1. 주가 데이터 정제 및 필터링

수집된 데이터 중에서 날짜, 종가, 거래량 열만을 남기고 불필요한 열을 제거하였다. 데이터 분석의 정확성을 높이기 위해 결측치와 이상치를 처리하는 과정도 포함되었다.

2)-2. 공시자료 데이터 전처리

각 기업의 재무제표 데이터를 주가 예측에 사용하기 위해 전처리를 수행하였다. 먼저, 결측치와 이상치를 모두 제거하였고, 그 후, 추후 클러스터링을 수행하기 위해 필요한 PER, PBR, 영업이익증가율, 자기자본이익률, 부채비율 등 5 개의 주요 Feature 를 추출하는 과정도 수행하였다.

2)-3. 주요 경제지표 데이터 전처리

위의 '데이터 수집' 단계에서 설명한 6 개의 주요 경제지표를 데이터를 주가 예측에 사용하기 위해 전처리를 수행하였다. 각 6 개의 지표 데이터를 기반으로, 각 날짜의 지표 변동치를 산출하여 절대 수치와 함께 입력변수로 사용하였다. 이는 각 경제지표가 주가에 미치는 절대적인 영향과, 경제지표가 발표되었을 때의 주가 변동을 모두 고려하기 위함이다.

II. 본론

4. 서비스 사용 방안

1) 재무제표 실적 발표에 따른 예상 등락률

1. 현재 미래에셋증권 '투자캘린더' 서비스의 문제점

미래에셋증권의 '투자캘린더' 페이지에는 다양한 기업의 실적 발표 일정이 정리되어 있다. 그러나 실적 발표 일정 클릭 시 해당 종목의 정보 페이지로 넘어가며, 단순히 실적 발표 결과만 나열될 뿐, 실적이 주가에 미치는 영향에 대한 설명은 부족하다. 이로 인해 주식 투자자들은 실적 발표 결과가 주가에 어떤 영향을 미치는지 이해하기 어려워 투자에 어려움을 겪고 있다.

따라서 '투자캘린더'에서 특정 경제 지표를 클릭할 경우, 다음과 같은 정보를 제공하는 서비스를 제안한다:

- 과거 매출액, 영업이익 등의 기본 실적 발표 결과 및 예상치를 그래프로 제공

- 과거 실적을 기반으로 한 미래 실적 예상치에 따른 주가 변동률 제공

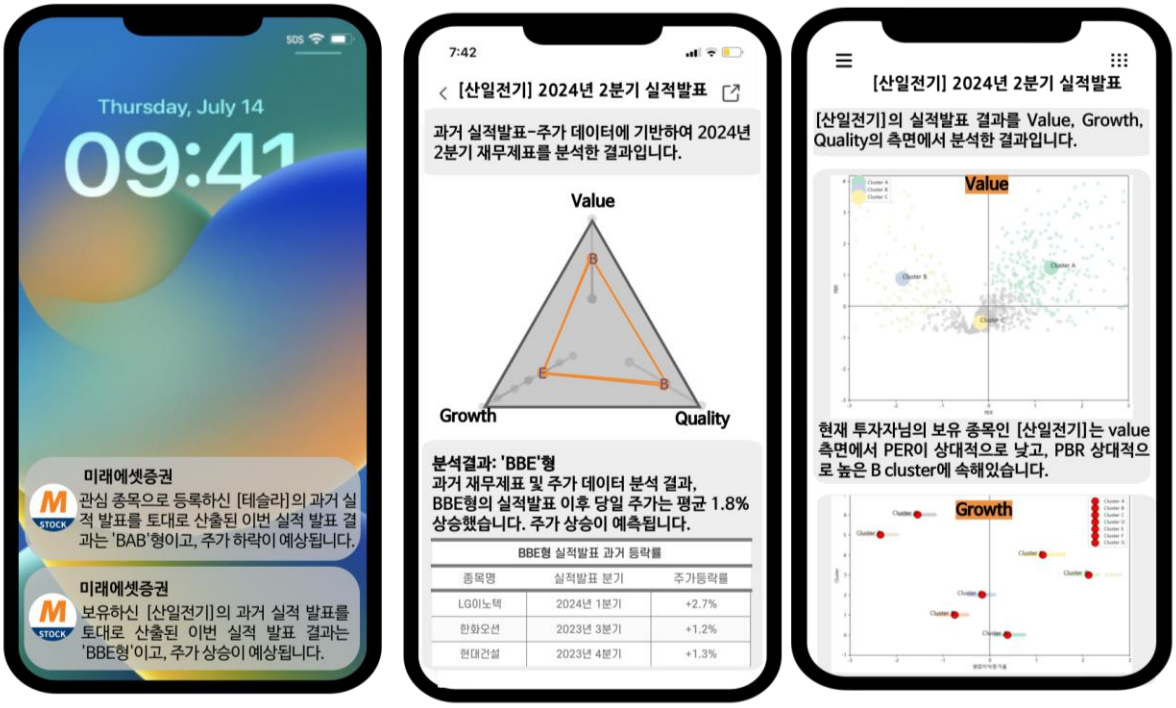
이를 통해 투자자는 실적이 주가에 미치는 영향을 보다 직관적으로 이해할 수 있어, 보다 합리적인 투자 결정을 내릴 수 있을 것이다.

2. 서비스 필요성

기업의 실적은 기업의 본질적 가치를 직접적으로 나타내며, 주가를 결정짓는 중요한 요소이다. 실적 발표 결과는 PER, PBR, 영업이익 증가율, ROE, 부채비율 등 다양한 지표를 통해 주가에 영향을 미친다. 따라서 주식 투자 시 분기별로 발표되는 실적 결과를 확인하는 것은 필수적이다.

과거 실적을 이해하기 쉽게 시각화하고, 미래 실적 예상치에 따른 예상 주가 변동률을 제공한다면, 투자자들은 보다 쉽게 기업 실적의 중요성을 인식하고 효과적인 투자 결정을 내릴 수 있을 것이다. 이러한 정보 제공은 주식 투자에 있어 중요한 참고자료가 될 것이다.

3. 프로토타입



4. 모델링 과정

본 연구의 목적은 기업의 실적 관련 지표가 주가에 미치는 영향을 분석하고, 이를 기반으로 투자자에게 유용한 예측 정보를 제공하는 것이다. 이를 위해 다음과 같은 분석 과정을 진행하였다.

4-1. 데이터 수집 및 전처리

데이터 수집: 주가와 관련된 중요 지표로 PER, PBR, 영업이익 증가율, ROE, 부채비율을 선정하여 클러스터링을 진행하였다. 이 5 가지의 지표를 다음과 같이 3 가지 factor 로 분류하였다.

- Value factor: PER, PBR
- Growth factor: 영업이익 증가율
- Quality factor: ROE, 부채비율

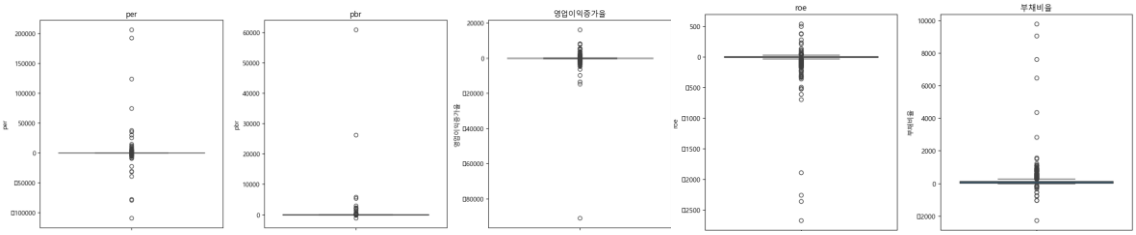
상장 종목의 재무제표와 주가 정보를 스크래핑한 후, 동일한 종목코드에 대해 데이터를 병합하였다. 이 과정에서 주가 등락률은 2023 년 12 월 28 일의 시가와 종가의 차이를 사용하여 계산하였다.

	종목코드	당기순이익_당기금액	부채총계_당기금액	영업이익_당기금액	자산총계_당기금액	당기순이익_전기금액	부채총계_전기금액	영업이익_전기금액	자산총계_전기금액	시가	종가	등락률	상장주식수
0	10	3.070000e+12	4.750000e+14	NaN	5.080000e+14	3.050000e+12	4.610000e+14	NaN	4.920000e+14	2415	NaN	NaN	NaN
1	20	2.823705e+10	1.580000e+11	1.877588e+10	5.650000e+11	2.159121e+10	8.246321e+10	2.991508e+10	4.610000e+11	4860	10270.0	3.53	27931470.0
2	30	2.520000e+12	4.310000e+14	NaN	4.580000e+14	2.900000e+12	4.180000e+14	NaN	4.430000e+14	17170	NaN	NaN	NaN
3	40	-2.102667e+10	1.160000e+11	-1.785685e+10	1.390000e+11	-1.300128e+10	1.090000e+11	-6.234393e+09	1.490000e+11	6600	358.0	3.47	96138465.0
4	50	-1.345222e+10	4.690000e+11	1.596883e+10	1.210000e+12	4.415250e+09	4.930000e+11	3.189531e+10	1.250000e+12	22550	8840.0	0.23	27415270.0

결측치를 제거하고 각 지표의 계산식을 바탕으로 값을 산출한 후, 종목코드와 클러스터링에 필요한 지표만을 최종적으로 남겼다.

	종목코드	per	pbr	영업이익증가율	roe	부채비율
1	20	4.807406	0.333531	-37.236065	6.937849	38.820639
3	40	-30.176623	27.587560	186.424874	-91.420300	504.347826
4	50	-45.956309	0.834297	-49.933616	-1.815414	63.292848
6	70	0.142801	0.011577	-28.179752	8.107143	83.571429
7	80	255.364402	8.096675	-35.078534	3.170636	200.000000

본 연구에서는 PER, PBR, 영업이익 증가율, ROE, 부채비율 등 5 가지 주요 지표에 대한 기본적인 통계량을 확인하였다. 데이터의 이상치를 제거하기 위해 Box plot 을 활용하여 분포를 확인한 결과, Q1 과 Q3 범위 외에 데이터가 많이 분포함을 확인하였다.



box plot 에서 Q1, Q3 사이 범위 외에 데이터들이 많이 분포하는 것을 한 눈에 확인할 수 있었고, 이에 이상치 제거는 필수적인 과정이라 판단하였다. IQR(Interquartile Range)을 기반으로 1.5*IQR 범위를 벗어난 데이터를 이상치로 간주하고 제거하였다. 이후 모든 특성을 동일 중요도로 고려하기 위해 스케일링 작업을 수행하였다.

3 가지 클러스터링 기법인 K-Means 클러스터링, Hierarchical 클러스터링, DBSCAN 을 이용하여 3 가지 factor 에 대해 클러스터링을 진행하였다. 군집화 타당성 평가 지표로는 Dunn Index 와 Silhouette Index 2 가지 평가지표를 사용하되, Silhouette Index 를 우선적으로 고려하였다.

4-2. K-Means 클러스터링

K-Means 클러스터링을 수행하기 전, 데이터셋 크기가 1129 개라는 점과 Rule of thumb 을 고려하여 $\sqrt{\frac{1129}{2}}$=24 까지의 군집 수 탐색 범위를 설정하였다. 2 가지 타당성 평가 지표를 종합적으로 고려한 결과, Value factor 는 3 개, Growth factor 는 7 개, Quality factor 는 3 개가 최적의 군집 수로 확인되었다.

최적의 군집 수를 바탕으로 K-Means 클러스터링을 진행하였고, 클러스터링 결과에 따라 52 개의 군집(final_label)과 각 군집의 평균 등락률을 산출하였다.

	final_label	label별 평균등락률
0	AAA	1.57
1	AAB	1.10
2	ABA	0.56

	종목코드	per	pbr	영업이익증가율	roe	부채비율	value_cluster	growth_cluster	quality_cluster	value_label	growth_label	quality_label	final_label
0	20	-0.086563	-0.458811	-0.303391	0.138777	-0.547890	2	2	0	C	C	A	CCA
1	50	-3.342965	0.123095	-0.482676	-0.953300	-0.086514	1	1	2	B	B	C	BBC
2	70	-0.385790	-0.832931	-0.175519	0.284661	0.295799	2	2	0	C	C	A	CCA
3	100	-0.226551	-0.651547	1.034806	0.070056	-0.639960	2	4	0	C	E	A	CEA
4	120	-0.352313	-0.800043	0.455413	0.021769	1.192068	2	0	1	C	A	B	CAB

4-3. 계층적 클러스터링

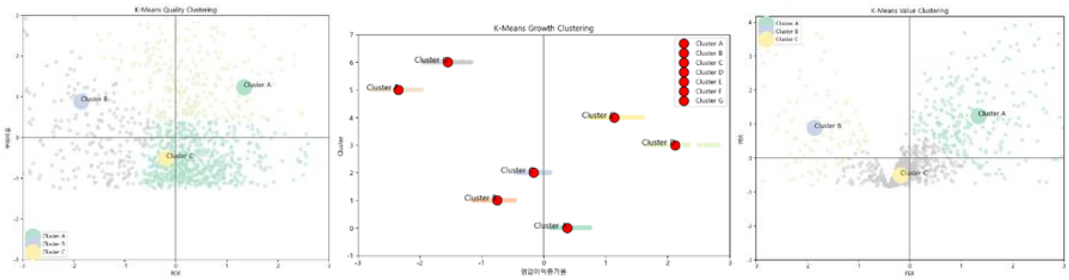
K-Means 클러스터링과 동일한 방식으로 계층적 클러스터링의 최적 군집 수 탐색을 진행하였다. 이때, 군집 수 변화에 따른 Single Linkage, Complex Linkage 의 군집화 타당성 지표 값을 비교하여 연결방식에 따른 최적의 군집 수 탐색을 추가적으로 진행하였다. 탐색 결과 complete linkage 방식을 채택하였고, K-Means 와는 달리 3 가지 factor 에 대한 최적의 군집 수가 모두 3 이라는 결론을 얻었다. 이를 바탕으로 최종 군집 별 평균 등락률을 산출하였다.

4-4. DBSCAN

DBSCAN 클러스터링을 수행하기 전, 하이퍼파라미터 eps 와 minPts 의 값을 조정하여 최적의 하이퍼파라미터를 탐색하였다. Value 클러스터의 경우, 최적 eps 값은 0.5, minPts 값은 4, Silhouette Score 는 0.57677 로 나타났으며, Quality 클러스터는 eps 0.5, minPts 2, Silhouette Score 0.48419 로 나타났다. Growth factor 에서는 유효한 군집이 형성되지 않아 하이퍼파라미터를 도출할 수 없었다. 이에 따라 2 가지 factor 에 대해서만 DBSCAN 클러스터링을 진행하였고, 각 방식의 Silhouette Index 는 다음과 같다.

클러스터링 기법	Value factor	Growth factor	Quality factor
K-Means	0.59329	0.53729	0.43995
Hierarchical	0.56016	0.51239	0.37011
DBSCAN	0.57677	-	0.48419

3 가지 clustering 방식 중 DBSCAN 이 Quality factor 에 대한 Silhouette index 가 가장 우수하게 나왔다. 그러나, 분류한 2 개의 클러스터에 대해 전체 데이터 1129 개 중 1122 개가 하나의 클러스터에 속하는 지나치게 편중된 결과가 나타났다. 이 점을 감안하여 모든 factor 에 대해 Silhouette index 가 가장 높은 K-Means 방식으로 3 가지 factor 에 대해 클러스터링을 진행하였다. K-Means 방식으로 각 데이터 포인트에 대해 클러스터링 결과를 시각화한 결과는 다음과 같다.



4-5. 서비스 기능: 새로운 실적 발표 시 예측 주가등락률 제공

클러스터링 결과를 바탕으로, 새로운 실적 발표가 있을 때마다 다음 두 가지 방법을 통해 예측 등락률을 제공할 수 있다.

1)기업 실적 발표 지표와 클러스터링 결과 나온 중심점과의 거리를 비교하여 중심점과 거리가 가장 가까운 클러스터에 속한다고 가정하여 예상 등락률을 산출 및 시각화하여 알림 서비스와 함께 투자자에게 제공

2)기업 실적 발표 지표와 클러스터링 결과 각 클러스터의 5 가지 경제지표의 값과 직접적으로 비교하여 값이 가장 유사한 클러스터에 속한다고 가정하여 예상 등락률을 산출 및 시각화하여 알림 서비스와 함께 투자자에게 제공

2 가지 방식을 종합적으로 이용한다면 보다 더 나은 서비스를 제공할 수 있을 것이라 생각된다. 특별히 1)방식의 경우 계산이 간단하다는 장점이 있으나 클러스터의 분포가 다양할 경우 중심점이 대표성이 떨어진다는 단점이 있다. 이 단점을 보완하고자 2)방법을 1)방법과 함께 사용함으로써 다양한 지표들을 다각도로 비교 가능하여 1)보다 정밀한 결과를 산출할 수 있을 것으로 기대된다.

2) 주요거시경제변수 등락에 따른 주가변동 예측

1. 현재 미래에셋증권 '투자캘린더' 서비스의 문제점

미래에셋증권의 '투자캘린더' 페이지에는 경제지표의 발표 일정이 정리되어 있다. 그러나 현재 제공되는 정보는 '발표예정일'과 '경제지표명'뿐이며, 이는 주식 입문자에게 부족한 정보이다. 투자자들은 경제지표의 중요성, 의미, 그리고 주가에 미치는 영향에 대해 충분히 이해하지 못하기 때문에 단순한 지표명과 발표예정일 정보로는 투자에 도움을 얻기 어렵다. 따라서 '투자캘린더' 페이지에서 특정 지표를 클릭할 때 '해당 지표의 정의', '해당 지표가 국내 주식시장에 미치는 영향', 그리고 '새로운 경제지표 발표 당일의 코스피지수 등락률 예측' 정보를 제공하여 투자자에게 실질적인 정보를 제공하는 서비스를 제안한다.

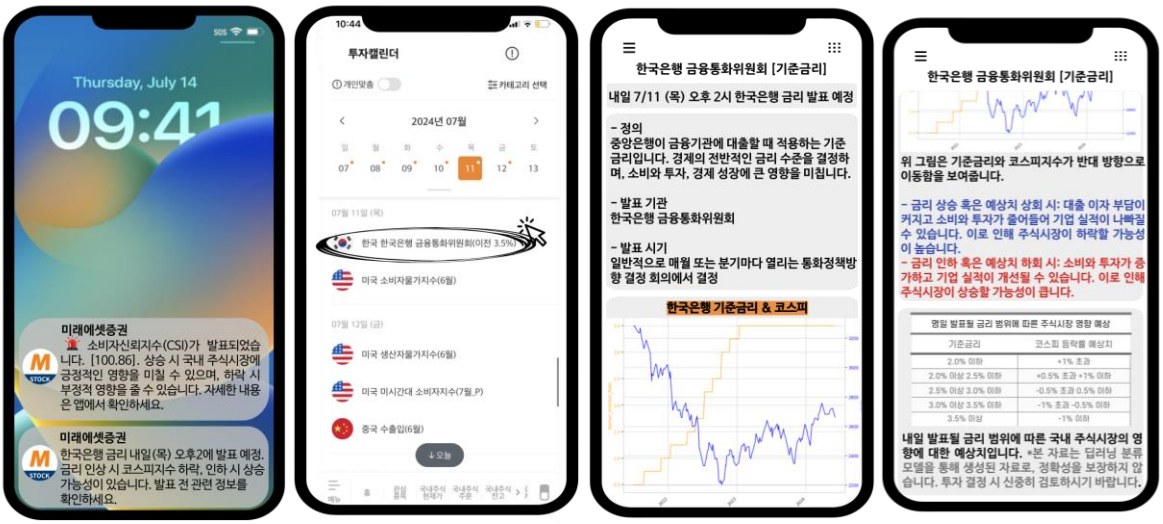
2. 서비스 필요성

주가는 기업의 본질적 가치인 '기업 내부 요인'과 기업 외부의 경제 환경을 나타내는 '기업 외부 요인'에 의해 결정된다. 기업 외부 요인에는 기준금리, 고용지표, 인플레이션, 중앙은행 통화정책 등 다양한 경제지표가 포함된다. 경제 상황과 같은 외부 조건들이 주가에 미치는 영향 때문에 경제지표를 올바르게 이해하고 주기적으로 확인하는 것은 투자에 필수적이다. 특히 미국, 중국 등 경제 강대국의 고용지표나 실업자 수는 전 세계 증시에 영향을 줄 수 있다. 경제지표는 투자의 중요한 지표로서, 투자상품과 연관된 지표를 활용하면 성공적인 투자 가능성을 높일 수 있다.

경제지표 리스트	
고용지표	**실업률**, 비농업고용지수, 실업수당 청구건수 등
물가지표	**소비자물가지수**, 개인소비지출 물가지수 등
경기지표	**경상수지, 소비자신뢰지수, 금리(한/미)**, 경기선행지수 등

따라서 국내 종목의 주가에 유의미한 영향을 미칠 것으로 예상되는 경제지표 6 가지를 선정하고, 지난 3 년간의 지수 변동과 주식시장의 수익성과의 관계를 분석하였다. 이를 통해 (1) 특정 지표가 코스피지수에 미치는 영향을 설명하고, (2) 특정 지표가 발표된 당일 코스피지수의 등락률을 예측하는 자료를 제공하고자 하였다.

3. 프로토타입



4. '경제지표에 따른 코스피지수 등락률' 분류모델 모델링 과정

본 연구는 주식시장에 영향을 주는 다양한 요인들 중 거시 경제적 새로운 정보의 유입(지표 발표)이 한국 주식시장에 어떠한 영향을 미치는지 분석하였다. 따라서 지난 3 년간의 소비자물가지수, 경상수지, 소비자신뢰지수, 실업률, 한국 금리 및 미국 금리의 변동, 그리고 한국 주식시장의 대표적 지수인 코스피지수(KOSPI)를 표본으로 하였다. 이를 위해 2021 년 6 월부터 2024 년 6 월까지 실제 지수 발표가 발생한 날의 지표 변동 및 코스피지수를 수집하였다. 금리를 제외한 모든 지수는 월별 자료이며 분석기간은 2021 년 6 월부터 2024 년 6 월까지로 설정하여 각 147 개의 데이터를 분석대상으로 하였다.

제공하고자 하는 주요 정보는 '경제지표의 정의', '경제지표 발표에 따른 코스피지수 변동 양상', 그리고 '분류모델을 통해 예측한 경제지표에 따른 코스피 등락률 예측'이다. '경제지표 정의'의 경우 복잡한 모델링이 필요하지 않기에, 이를 제외한 나머지 두 가지 주요 정보에 대한 모델링 과정을 다음과 같이 정리하였다.

4-1. 데이터 수집

	날짜	소비자물가지수_절대	소비자물가지수_변동	경상수지_절대	경상수지_변동	소비자신뢰지수_절대	소비자신뢰지수_변동	실업률_절대	실업률_변동	실질GDP_절대	한국금리_절대	한국금리_변동	미국금리_절대	미국금리_변동	종가	등락률
23	2021-06-24	2.6	0.0	1.800	0.0	110.3	0.0	3.8	0.0	1.34	0.5	0.0	0.25	0.0	3,286.10	0.30%
24	2021-06-25	2.6	0.0	1.800	0.0	110.3	0.0	3.8	0.0	1.34	0.5	0.0	0.25	0.0	3,302.84	0.51%
27	2021-06-28	2.6	0.0	1.800	0.0	110.3	0.0	3.8	0.0	1.34	0.5	0.0	0.25	0.0	3,301.89	-0.03%
28	2021-06-29	2.6	0.0	1.800	0.0	110.3	0.0	3.8	0.0	1.34	0.5	0.0	0.25	0.0	3,286.68	-0.46%
29	2021-06-30	2.6	0.0	1.800	0.0	110.3	0.0	3.8	0.0	1.34	0.5	0.0	0.25	0.0	3,296.68	0.30%
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1094	2024-05-30	2.9	0.0	69.314	0.0	98.4	0.0	2.8	0.0	1.30	3.5	0.0	5.50	0.0	2,635.44	-1.56%
1095	2024-05-31	2.9	0.0	69.314	0.0	98.4	0.0	2.8	0.0	1.30	3.5	0.0	5.50	0.0	2,636.52	0.04%
1098	2024-06-03	2.9	0.0	69.314	0.0	98.4	0.0	2.8	0.0	1.30	3.5	0.0	5.50	0.0	2,682.52	1.74%
1099	2024-06-04	2.7	-0.2	69.314	0.0	98.4	0.0	2.8	0.0	1.30	3.5	0.0	5.50	0.0	2,662.10	-0.76%
1100	2024-06-05	2.7	0.0	69.314	0.0	98.4	0.0	2.8	0.0	1.30	3.5	0.0	5.50	0.0	2,689.50	1.03%

727 rows × 16 columns

먼저, 6 가지 경제지표와 코스피 종가, 코스피 등락률 데이터를 수집한 후, 경제지표 발표가 없는 행은 제거하였다.

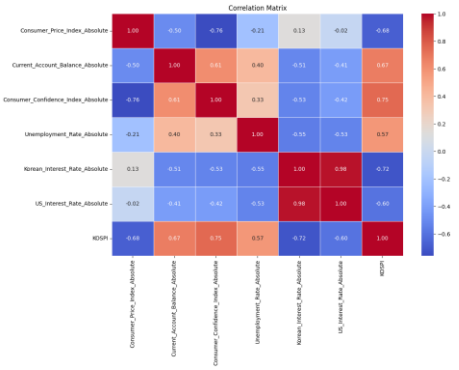
4-2. 경제지표와 코스피 지수의 상관관계

OLS Regression Results

Dep. Variable:	KOSPI	R-squared:	0.889
Model:	OLS	Adj. R-squared:	0.884
Method:	Least Squares	F-statistic:	186.6
Date:	Wed, 31 Jul 2024	Prob (F-statistic):	3.29e-64
Time:	04:15:54	Log-Likelihood:	-863.31
No. Observations:	147	AIC:	1741.
Df Residuals:	140	BIC:	1762.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3750.2288	310.094	12.094	0.000	3137.157	4363.301
Consumer_Price_Index_Absolute	-126.3478	14.345	-8.808	0.000	-154.709	-97.987
Current_Account_Balance_Absolute	0.4980	0.271	1.837	0.068	-0.038	1.034
Consumer_Confidence_Index_Absolute	-4.9849	2.393	-2.083	0.039	-9.715	-0.254
Unemployment_Rate_Absolute	91.6354	25.371	3.612	0.000	41.476	141.794
Korean_Interest_Rate_Absolute	-231.8141	47.126	-4.919	0.000	-324.986	-138.643
US_Interest_Rate_Absolute	49.5096	24.096	2.055	0.042	1.870	97.149

Omnibus:	8.205	Durbin-Watson:	0.568
Prob(Omnibus):	0.045	Jarque-Bera (JB):	3.764
Skew:	-0.205	Prob(JB):	0.152
Kurtosis:	2.332	Cond. No.	4.62e+03



OLS 분석과 상관관계 분석을 통해 각 경제지표가 코스피 지수에 미치는 영향을 평가하였다. 분석 결과, 각 경제지표는 코스피 지수와 유의미한 상관관계를 가지며, 새로운 경제지표 발표 시 코스피 지수의 등락을 예측하는 모델링이 중요하다는 결론에 도달하였다.

4-1-3. 경제지표와 코스피 지수의 상관관계

각 경제지표 수준에 따른 코스피 지수의 변동 양상을 시각화하였다. 소비자물가지수, 한국 금리, 미국 금리는 코스피 지수와 음의 상관관계를 가지며, 경상수지, 소비자신뢰지수, 실업률은 양의 상관관계를 가진다. 이를 통해 각 경제지표가 주식에 미치는 영향을 주식 입문자에게 쉽게 설명하고자 하였다.

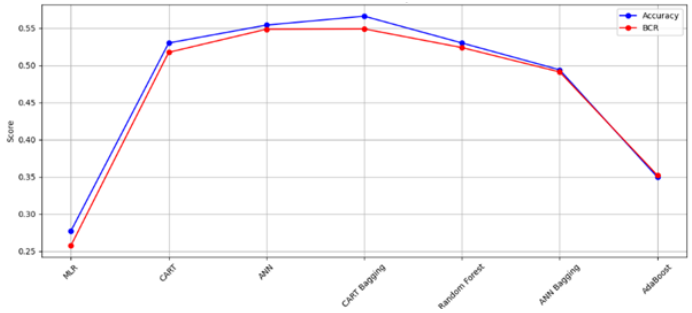
4-3. 코스피 등락 예측 분류모델 생성

코스피 등락률	분류 class	Oversampling 전 데이터 수	Oversampling 후 데이터 수
1 초과	상승(uptrend)	32 개	55 개
0.5 초과 1 이하	약한상승(weak uptrend)	12 개	55 개
-0.5 초과 0.5 이하	유지(stable trend)	55 개	55 개
-1 초과 -0.5 이하	약한하락(weak downtrend)	25 개	55 개
-1 이하	하락(downtrend)	23 개	55 개

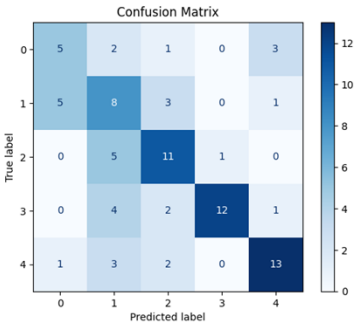
새로운 경제지표 발표 시 해당 값을 바탕으로 당일의 코스피 지수 등락을 예측하는 분류 모델을 생성하였다. 데이터는 5 가지 클래스로 라벨링하였으며, 클래스 불균형 문제를 해결하기 위해 SMOTE(Synthetic Minority Over-sampling Technique)기법을 이용하여 Oversampling 을 적용하였다.

MLR, CART, ANN, CART Bagging, Random forest, ANN Bagging, Adaboost 의 7 가지 모델을 생성하고 그 성능을 비교한 결과는 다음과 같다.

	Accuracy	BCR
MLR	0.277108	0.257416
CART	0.530120	0.517534
ANN	0.554217	0.548719
CART Bagging	0.566265	0.549113
Random Forest	0.530120	0.523952
ANN Bagging	0.493976	0.491134
AdaBoost	0.349398	0.351984



Accuracy 와 BCR 모두를 고려하여 CART Bagging 모델의 성능이 가장 우수하다고 판단하였고, 해당 모델의 성능을 개선하기 위해 최적의 하이퍼파라미터 조합을 탐색하여 분류 모델을 완성하였다. 최종 분류모델은 ['n_estimators': 100, 'max_samples': 1.0, 'max_features': 0.8, ''max_depth': None, 'min_samples_split': 5, 'min_samples_leaf': 1]이 하이퍼파라미터 조합을 이용한 CART Bagging 모델이다. 해당 모델이 테스트 데이터셋을 분류한 결과는 아래와 같다. 이때 Accuracy 0.590361, BCR 0.577596 의 성능을 보였다.



이렇게 만든 분류모델을 이용하여, 새로운 경제지표의 수치가 발표되었을 때 해당 수치가 한국 주식시장에 전반적으로 어떠한 영향을 나타낼지 5 가지의 경우로 분류하는 기능을 구현하고자 하였다. 이를 통해 사용자는 경제지표 발표에 따른 주가 변동 예측 정보를 제공받을 수 있다.

3) 소비자 반응 분석에 따른 주가 예측

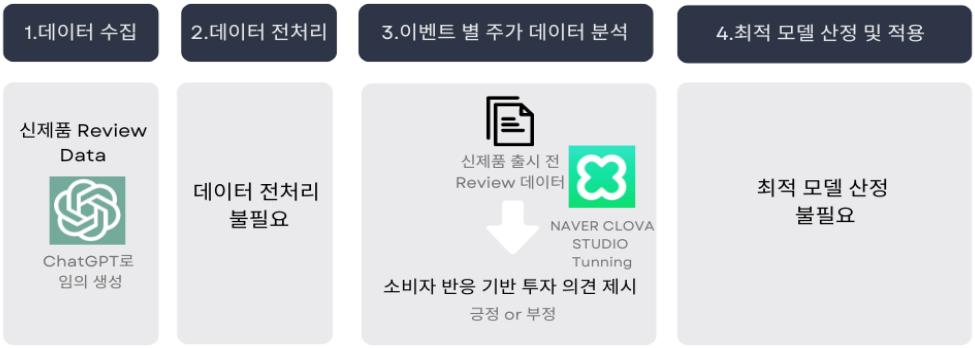
1. 서비스 필요성

신제품 출시 전 신제품 관련 소비자들의 반응은 기업의 주가에 중요한 영향을 미친다. 이는 제품의 성능, 품질, 사용자 경험 등을 반영하며, 이러한 리뷰의 긍정적 또는 부정적 경향은 투자자들에게 중요한 신호가 된다. 하지만, 수많은 리뷰를 수작업으로 분석하는 것은 매우 비효율적이며 시간과 노력이 많이 소모된다. 자동화된 시스템을 통해 소비자 리뷰를 분석하고, 이를 바탕으로 주가 변동을 예측할 수 있다면 투자자들에게 큰 도움이 될 것이다. 나아가 이러한 자동화된 분석 시스템은 투자자들이 더 나은 결정을 내리는 데 기여할 수 있으며, 주가 변동을 사전에 예측하여 투자 위험을 최소화할 수 있다.

2. 서비스 구현

2-1. 기술 개요

이 서비스는 Streamlit 을 이용한 웹 애플리케이션으로, 사용자가 업로드한 CSV 파일의 리뷰 데이터를 처리하여 사전에 '플레이그라운드' 기능으로 훈련시킨 네이버 하이퍼클로바 모델을 API 를 통해서 리뷰의 긍/부정 여부를 분석하고, 그 결과를 바탕으로 주가 변동 예측과 투자 의견을 함께 제공한다.



2-2. 구현 방안

이 서비스는 크게 3 단계로 나뉘어져 있다.

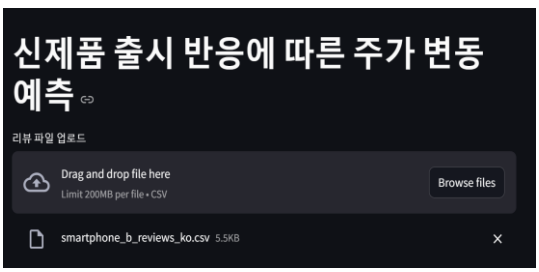
1. 사용자 리뷰 파일 업로드 - 사용자가 리뷰 파일을 업로드하면, 파일을 읽고 리뷰 데이터를 확인할 수 있도록 화면에 표시한다.
2. 하이퍼클로바 API 호출 - API 와 통신하기 위한 CompletionExecutor 클래스를 정의하여 API 호출을 처리했다.Naver Clova Studio 활용하여 프롬프트 및 파라미터 조정. 모델의 안정성 향상을 위해 Base model 에서 Temperature 을 낮게 조정하였다.
3. 리뷰 분석 및 결과 출력 - 업로드된 리뷰를 하나씩 클로바 API 에 보내 분석하고, 그 결과를 바탕으로 긍정/부정 중 하나의 결과값을 리턴한 후, 전체 리뷰 결과를 분석하여 긍정/부정 중 다중 클래스로 종합하여 이를 바탕으로 투자 의견을 제공하였다.

3. 프로토타입

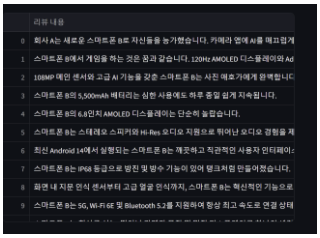
본 연구에서는 실제 데이터셋을 가져오려고 하였으나, 적절한 실제 기업의 신제품 Review 데이터를 충분히 확보하지 못해 가상의 데이터를 사용하였다. 실제 서비스로 구현할 경우, 블로그, 뉴스 기사 등 충분한 데이터를 확보해야 할 것으로 예상된다.

웹 프로토타입은 앞서 언급한 Streamlit 을 활용하여 제작하였고, 실행 과정은 아래와 같다.

<Review Dataset 업로드>



<리뷰 데이터 시각화>



<Review Dataset 기반 투자 의견 제시>

투자 의견 : 신제품에 대한 소비자들의 긍정적인 반응이 주를 이루고 있습니다. 이러한 반응을 토대로 볼 때, 향후 주가 상승이 기대되므로 이 기업의 주식을 매수하는 것을 권장합니다.

1) 종목 이슈에 따른 주가변동률 예측

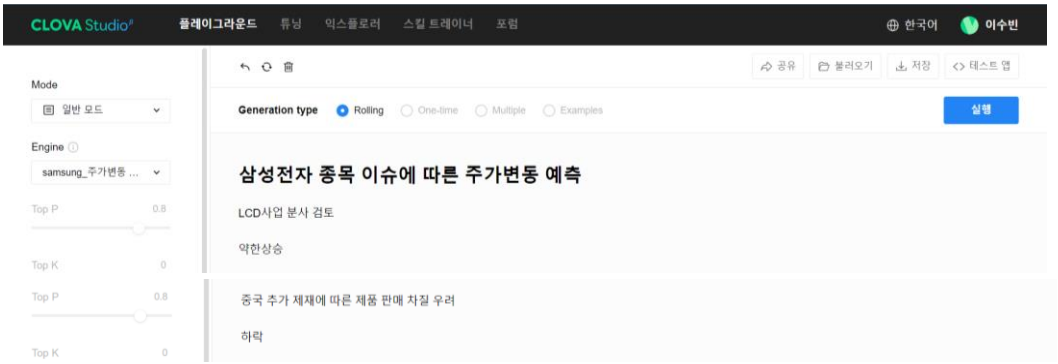
현대 투자자들은 금융 시장의 변동성과 불확실성 속에서 빠르고 정확한 정보를 바탕으로 한 투자 결정을 필요로 한다. 이에 따라 미래에셋증권 어플의 '투자 NOW' 섹션에서 제공하는 뉴스 기사와 함께, 해당 뉴스가 종목 주가에 미칠 영향을 분석하여 제공하는 새로운 서비스를 제안한다. 본 서비스는 인공지능 모델을 활용하여 종목 이슈에 따른 주가 변동률을 예측하고 이를 사용자에게 직관적으로 제공함으로써, 보다 스마트한 투자 결정을 지원하는 것을 목표로 한다.

이 서비스를 구현하기 위해 먼저 인포스탁에서 삼성전자의 최근 15 년치 종목 이슈 데이터를 수집하였다. 수집된 데이터는 해당 이슈가 발표된 날의 주가 변동률을 분석하는 데 사용되었다. 주가 변동률은 -1.5% 이하는 '하락', -1.5%에서 -0.5%는 '약한 하락', -0.5%에서 0.5%는 '유지', 0.5%에서 1.5%는 '약한 상승', 1.5% 이상은 '상승'으로 분류했다. 수집한 데이터셋은 다음과 같다.

	Text	Completion
2	2분기 세계 스마트폰 시장 점유율 1위 소식에	상승
3	52주 신고가 - 5G(5세대 이동통신), 자율주행차 테마 상승 속 2분기 잠정실적 어닝 서프라이즈 모멘텀 지속 및 美 지역통신사 목표 5G 네트워크 장비 시장 공략 강화 소식	유지
4	52주 신고가 - 일부 반도체 관련주 상승 속 2분기 잠정실적 어닝 서프라이즈	상승
5	2분기 잠정실적 어닝 서프라이즈에	상승
6	엔비디아 HBM3E 퀄테스트 통과 논란 속 HBM 개발팀 신설 소식 및 실적 기대감 부각 등에	상승
7	HBM 엔비디아 인증 통과 가능성 크다는 전망 등에	상승
8	삼성전자 노조, 창사 이래 첫 파업 선언 소식에	하락
9	엔비디아향 HBM3E 8단 공급 지연설 속	약한하락
10	美 AMD에 4조원대 HBM 물량 공급 소식에	상승
11	1분기 어닝 서프라이즈 발표에도	약한하락
12	52주 신고가 - 반도체 관련주 상승 속 1분기 호실적 기대감	약한상승
13	52주 신고가 - 일부 반도체 관련주 상승 속 1분기 호실적 기대감 지속	상승
14	52주 신고가 - 반도체 관련주 상승 및 일부 OLED(유기 발광 다이오드) 테마 상승 속 실적 턴어라운드 기대감 지속	상승

이후 네이버 클로바 스튜디오의 '튜닝 - 문서 분류 (Multi-class Classification)' 기능을 이용하여, 수집한 데이터셋을 기반으로 예측 모델을 학습시켰다. 이 모델은 다중 분류 기능을 활용하여 주가 변동률을 5 가지 범주로 예측한다.

예를 들어, '중국 추가 제재에 따른 제품 판매 차질 우려'라는 이슈가 입력되면 결과값이 '하락'으로 도출되는 방식이다. 학습시킨 모델에 임의의 테스트 값을 입력한 결과는 아래와 같다.



해당 예측 모델을 바탕으로 '투자 NOW' 섹션에서 뉴스 기사와 함께 해당 이슈가 주가에 미칠 영향을 분석하여 제공한다. 사용자들은 뉴스 기사와 더불어 해당 이슈가 주가에 어떤 영향을 미칠지를 직관적으로 확인할 수 있다. 이를 통해 투자자들은 보다 신뢰성 있는 정보를 바탕으로 투자 결정을 내릴 수 있다. 본 서비스의 도입으로 투자자들의 의사 결정 지원이 강화될 것으로 기대된다. 차별화된 정보 제공을 통해 미래에셋증권 어플의 경쟁력을 높이고, 사용자들에게 더 나은 경험을 제공할 수 있다.

III. 결론

1. 기대 효과

1-1. 신속하고 정확한 맞춤형 투자 정보 제공

인공지능 모델을 통해 실시간으로 중요한 금융 이벤트와 경제 지표를 분석하여 투자자의 포트폴리오와 관심사를 고려하여 정확한 정보를 알림 서비스를 통해 신속하게 제공함으로써 투자 결정을 지원하고 잠재적인 투자 기회를 놓치지 않도록 돕는다. 뉴스 지수 실적 신제품

1-2. 정보 해석력 및 설명력 강화

주가 관련 긍정/부정 판단에서 나아가 실적 발표, 지수 변화, 뉴스, 신제품 출시와 같은 관련 이벤트가 주가에 미치는 영향을 분석하고 설명함으로써 투자자들의 이해를 돕는다.

2. 의의

2-1. 투자자의 투자결정 지원

복잡하고 방대한 금융 데이터를 쉽게 이해하고 활용할 수 있도록 도와줌으로써 투자자의 의사결정을 효과적으로 지원하고 금융 시장에서의 경쟁력을 높인다.

2-2.지속 가능한 투자 문화 조성:

보다 신뢰성 있는 정보와 분석을 통해 투자자들이 장기적이고 지속 가능한 투자 전략을 수립할 수 있도록 지원함으로써 건전한 투자 문화를 조성한다.

본 서비스는 금융 시장의 변동성과 불확실성 속에서 투자자들에게 중요한 정보를 제공하고, 이를 바탕으로신속하고 정확한 투자 결정을 내릴 수 있도록 돕는 데 큰 의의를 두고 있다. 이를 통해 미래에셋증권은 고객에게 더욱 높은 가치를 제공하여 시장에서의 경쟁력 강화를 통한 금융 서비스 시장에서의 리더십을 확립할 수 있을 것이다.