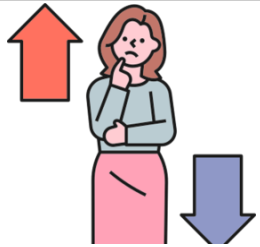
주가 데이터를 활용한
일, 주, 월 모델 통합 인공지능 모델링 분석



팀명 : 5인조

이승연 임소연 임형섭 곽병찬 김한호

목차

분석 개요

1. 주제 선정 배경
2. 분석 목표

분석 개요

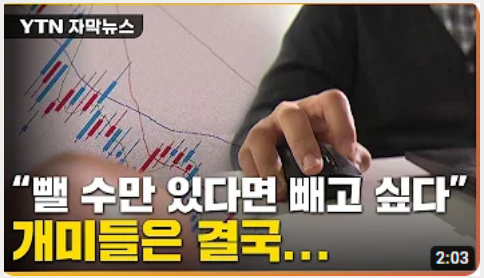
주제 선정 배경

1

주식 투자 열풍

2

개인투자자 정보 부족



3

주식 시장의 변동성

분석 개요

분석의 필요성

- 주식 투자에 대한 관심이 높아지면서 주가 예측 정확도를 높이는 분석의 필요성 부각
- 기존 연구는 시간 / 일 단위의 기술적 지표 활용하여 진행

분석 목표

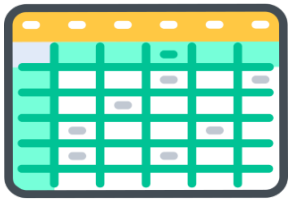
10일 / 10주 / 10월 단위의 기술적 지표를 활용한 다음날 등락률 상승 여부 예측

분석 과정

1. 데이터 수집
2. EDA
3. 전처리
4. 모델링

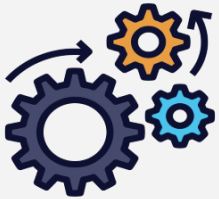
분석 과정

STEP1. 데이터 수집



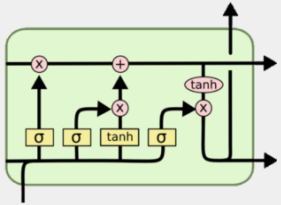
- 일 단위 주가 데이터 수집
- KOSPI 200
- 데이터 확인

STEP2. 전처리



- 주가 스케일링
- 거래정지 데이터 제거
- 파생변수 추가
- 종속변수 이진화

STEP3. 모델링



- 모델 학습 환경 구축

STEP4. 평가

- 성능 비교

STEP5. 결론

- 기대 효과
- 서비스 활용방안
- 의의 및 한계

데이터 수집

데이터 수집

- pykrx 라이브러리 활용
- 네이버 / 한국거래소에서 주가정보로 스크래핑한 데이터

데이터 정보

- 기간 : 2013-01-01 ~ 2023-06-30
- 특징 : KOSPI 200 일별 주가 데이터
- 주요 정보 : 시가, 고가, 저가, 종가, 거래량, 등락률, 거래대금

	날짜	종목코드	시가	고가	저가	종가	거래량	등락률	거래대금
1	2013-01-02	005930	30660	31519	30540	31520	11293461	3.547963	355968994
2	2013-01-03	005930	31680	31680	30860	30860	14358005	-2.093909	443088035
3	2013-01-04	005930	30840	30840	30199	30500	12959143	-1.166559	395253862
...	...	...	...	...	...	...	...	...	
359175	2023-06-30	057050	45600	46100	45200	45800	7976	1.103754	365329

데이터 수집

시계열 특성의
대표적인 주식 지수

자동 매매 활발

KOSPI 200

예측 용이성

연구 및 리포트 활용

데이터 수집

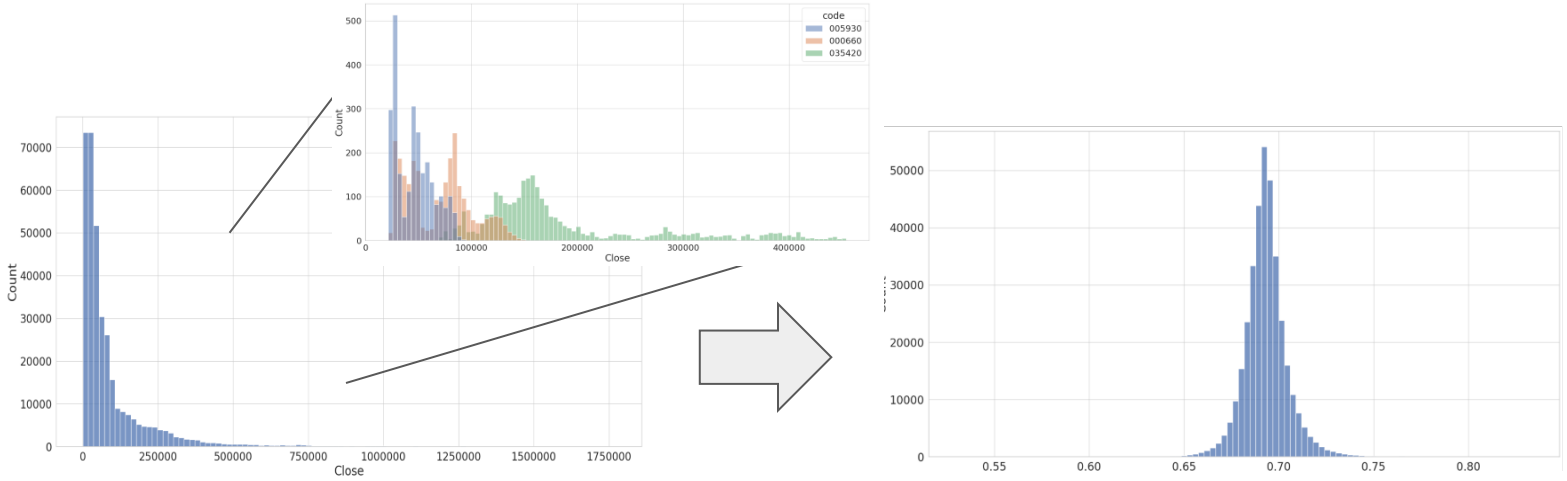
데이터 확인

	Open	High	Low	Close	Volume	Change	Trading	captalization
count	359176.00	359176.00	359176.00	359176.00	359176.00	359176.00	359176.00	359176.00
mean	106543.30	108155.75	104941.24	106696.03	692252.63	0.00	30280200081.27	7876443393189.45
std	179173.96	181785.77	176640.03	179119.33	2190141.91	0.02	92928692990.80	26864419606418.76
min	0.00	0.00	0.00	895.00	0.00	-0.30	0.00	36008276820.00
25%	22550.00	22950.00	22200.00	22700.00	77799.00	-0.01	4015280650.00	1362504346500.00
50%	45600.00	46300.00	44947.00	45700.00	206445.50	0.00	10183023100.00	2721699545600.00
75%	105000.00	106500.00	103000.00	105000.00	571042.75	0.01	26161083437.50	6578745098800.00
max	1770000.00	1784000.00	1756000.00	1770000.00	175665905.00	0.30	8379237727064.00	543250212050000.00

- ☑ 종목 간 주가 표준편차가 크게 나타남
  -> 주가 분포 확인 필요
- ☑ 주가, 거래량, 거래대금 최솟값 0 존재
  -> 해당 종목 확인 필요
- ☑ 등락률 -30% ~ 30% 존재
  -> 주가가 급변동한 종목 X

EDA 및 전처리

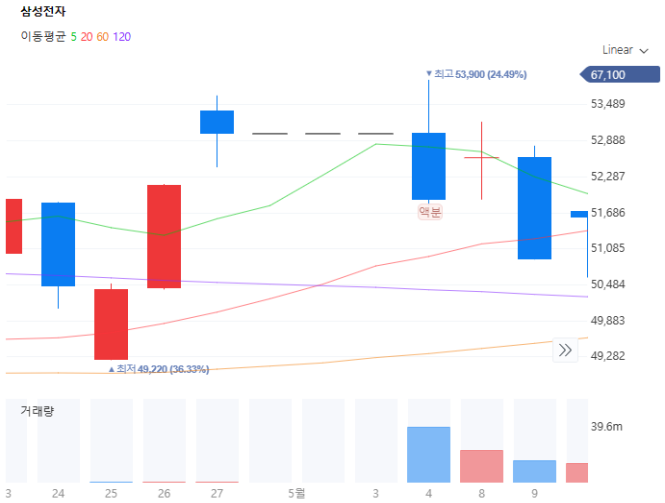
종가 분포 확인



- ☑ 종가의 분포를 보면 800원 170만원대 사이에 존재함
- -> 전날 종가로 나누어 스케일을 조정한 후 로그 변환을 적용한 데이터의 분포

EDA 및 전처리

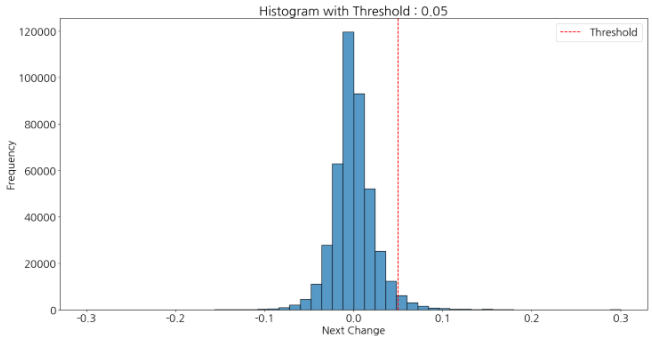
삼성전자 거래정지기간 주가 시각화



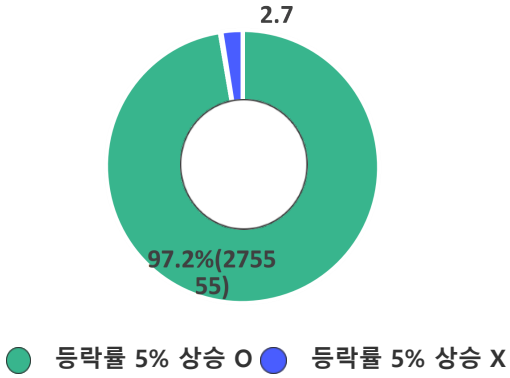
- 주가가 0으로 나타나는 종목 중 삼성전자를 선정하여 시각화
- 액면분할로 인해 4월 30일부터 5월 2일까지 거래정지
- 거래정지기간의 주가는 실제 주식 시장의 움직임을 반영하지 않음

-> 거래정지 기간 데이터 제거

EDA 및 전처리



다음날 등락률 분포



타겟 변수 (다음날 등락률 5% 상승 여부)

- 다음날 등락률 분포를 고려하여 임계값 설정
- 클래스 비율 불균형

-> 가중치를 조정하여 모델 학습

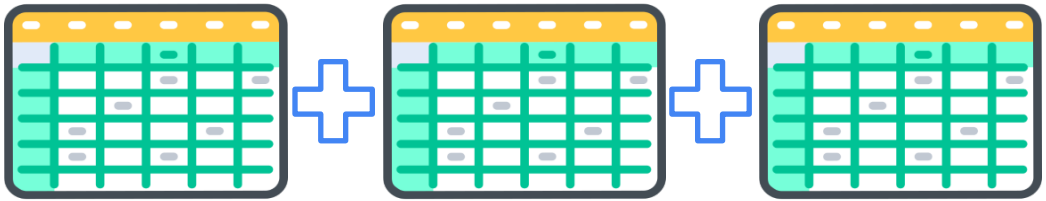
EDA 및 전처리

파생변수 추가

일봉 데이터(+보조지표)

주봉 데이터(+보조지표)

월봉 데이터(+보조지표)



최종 데이터

- ☑ 주식 시장의 복잡성을 고려하여 파생변수 추가
- -> raw 데이터에 보조지표 / 주 / 월봉 데이터 추가

EDA 및 전처리

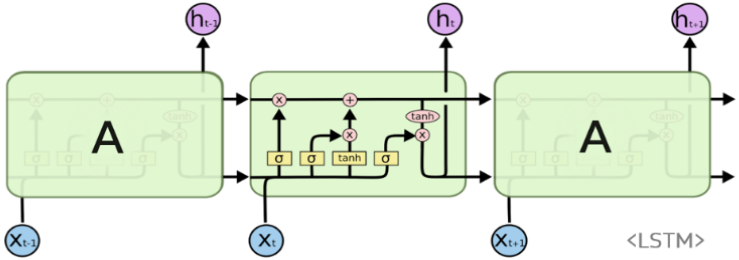
최종 전처리 데이터

- 날짜를 기준으로 10일 (D-9~D-0)/ 10주(W-9~W-0) /10달(M-9~M-0) 간의 주가(시가,고가,저가,종가,거래대금)와 48개의 보조지표가 추가된 데이터
- 약 1600개의 열이 생성됨(일 : 530개, 주 : 530개, 월 : 530개)
- 스케일링 및 로그 변환 / 거래정지 데이터 제거 / 타겟 변수 이진화 처리를 모두 마친 최종 데이터

	날짜	종목코드	시가(D-9)	...	PVO(D-0)	시가(W-9)	...	PVO(W-0)	시가(M-9)	...	PVO(M-0)	다음날 5% 상승여부
1	2013-01-02	005930	0.687587	...	-0.795291	0.616712	...	-13.194155	0.60822	...	-16.504565	0
2	2013-01-03	005930	0.688516	...	-0.736062	0.612919	...	-11.308896	0.59278	...	-16.115162	1
3	2013-01-04	005930	0.693147	...	-1.463722	0.622672	...	-9.6571	0.602299	...	-15.765787	0
...	...	...	...	...	...	...	...	...	...	...	...	
331999	2023-06-30	057050	0.691133	...	-16.190021	0.726764	...	-7.472236	0.730888	...	-27.460826	0

모델링

LSTM



- RNN이 출력이 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장/단기 기억을 가능하게 설계한 신경망의 구조
- 시간에 따른 데이터의 패턴과 추세를 파악하는데 뛰어난 성능을 보임
- 다양한 형태의 입력 데이터를 처리할 수 있어 주/월봉 데이터와 결합하여 사용하기 용이

모델링

모델 학습 환경 구축

2013년 ~ 2022년 1월 2023년

데이터 shape

일별 : (341999,10,53)
주별 : (341999,10,53)
월별 : (341999,10,53)

최종 : (341999,30,53)

모델 구축

LSTM(50)
LSTM(50)
LSTM(50)
LSTM(50)
DENSE(1)

주요 파라미터 정보

Batch_size : 32
Class 0 weight : 0.52
Class 1 weight : 17.84

모델 컴파일

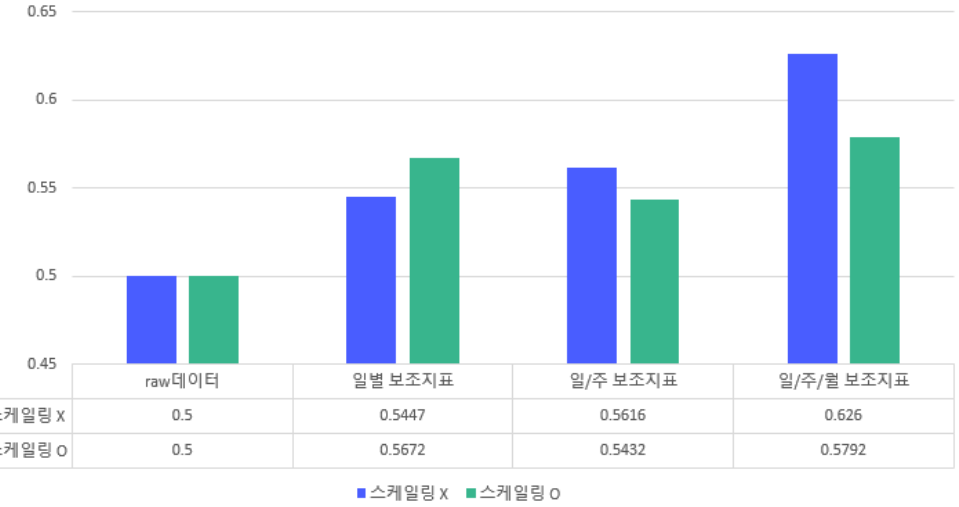
Loss : binary_crossentropy
Optimizer : adam
Metrics : AUC

결론

1. 결론
2. 기대효과

결론

모델 성능 비교



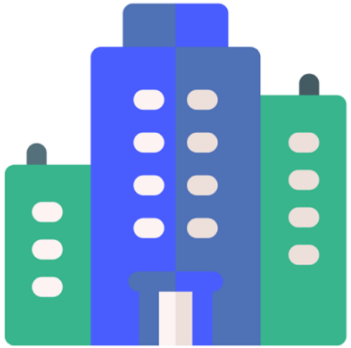
	raw데이터	일별 보조지표	일/주 보조지표	일/주/월 보조지표
스케일링 X	0.5	0.5447	0.5616	0.626
스케일링 O	0.5	0.5672	0.5432	0.5792

- 일 / 주 / 월별 데이터를 추가하면서 성능이 향상됨을 확인함
- 주별과 월별 데이터의 경우 스케일링 처리를 하지 않은 데이터의 모델 성능이 더 좋게 나타남

기대효과

등락률 예측은 전반적인 시장의 전망을 제시, 투자자의 의사결정을 지원

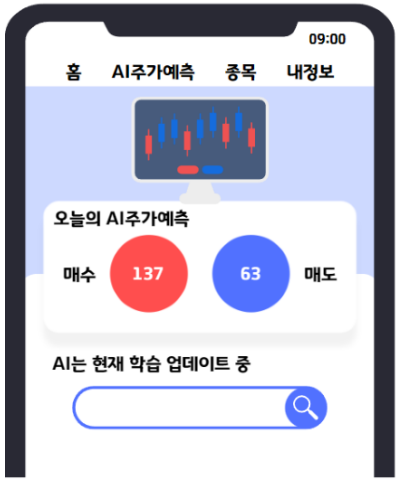
효율적인 투자 전략 구축 및 조정

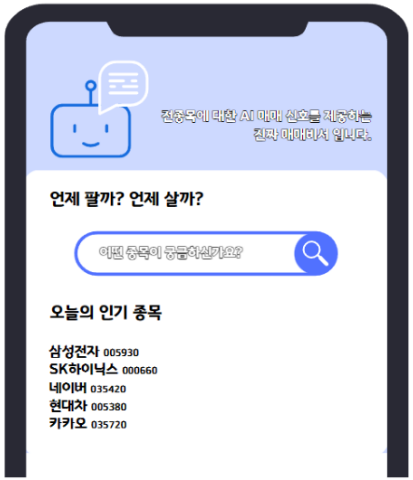


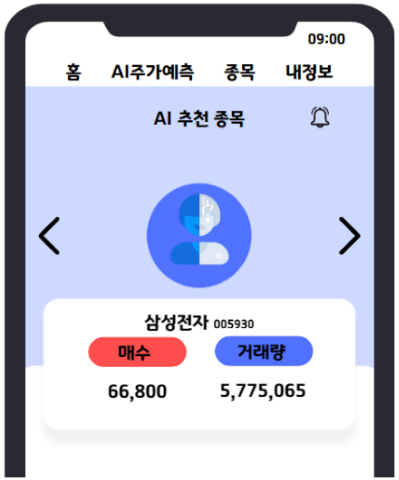
리스크를 최소화 하기 위한 안전장치로써의 역할

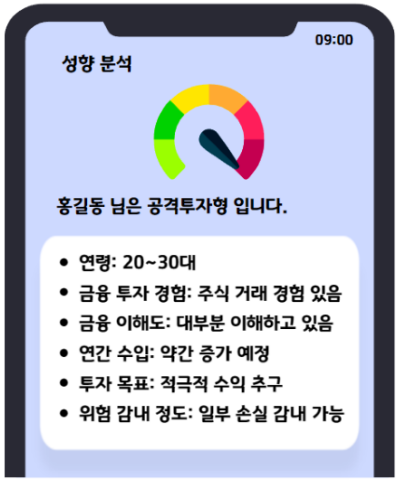
서비스 활용방안

AI 매매비서 서비스를 제공하는 주식 투자 플랫폼으로 확장 가능









의의 및 한계

분석의 의의

- 다양한 시간 단위의 데이터(주/월봉)를 활용한 주가 예측
- 주 / 월봉 데이터의 보조지표 사용이 성능 향상에 긍정적인 영향을 미침

분석 한계

- 외부 변수 미반영 (인수합병, 사고, 정책 발표 , 정치적 이슈 등과 같은 요소)
- 코로나와 같은 비정상적인 사건의 예측 불가능성, 급격한 시장 변동성이 있을 때의 어려움

부록

1. 참고문헌 및 분석도구

참고문헌 및 분석도구

논문 출처

- 주봉 데이터를 활용한 CNN-LSTM 기반 주가예측 모델 연구
- 주가 예측 모델에서의 분할 예측을 통한 성능향상 탐구
- 주식시세 예측을 위한 딥러닝 최적화 방법 연구
- LSTM, SVM 기계 학습 기반 KOSPI200 시계열 학습 값에 따른 미래 값 예측력 비교
- 통합 종목 주가 예측을 위한 시계열 스케일러 비교
- LSTM분류의 KOSPI 지수 예측에의 적용
- 시계열데이터 예측을 통한 주가예측 정확도 향상

기사 출처

- 한국 주식시장 수익률 2번째로 낮고 변동성 2번째로 높아
  https://biz.sbs.co.kr/article/20000104134
- 소신 투자그룹 개인투자자 애로사항
  http://sjbnews.com/news/news.php?number=785988
- 주식 투자자 43% "코로나 이후 시작"... 92% "계속할 것"
  https://www.hankookilbo.com/News/Read/A2021050316140000896

분석도구

python

jupyter