

정형/비정형 데이터 활용 저평가 우량주 예측

예상일 | 안아련 | 장재석 | 최종완 | 한성용



Slido Members

프로젝트 소개

프로젝트 조직도

Leader



예상일

데이터 크롤링
정형 데이터 전처리
저평가주 판단 모듈 생성
정형 데이터 변수 산출
프로젝트 총괄

Member 1



안아련

데이터 크롤링
정형 데이터 분석
Machine learning Modelig
PPT 및 디자인 총괄
프로젝트 발표

Member 2



장재석

데이터 크롤링
비정형 데이터 전처리
뉴스기사 NLP 분석
뉴스기사 변수 산출
Machine Learning Modeling

Member 3



최종완

데이터 크롤링
비정형 데이터 전처리
종목 토론방 NLP 분석
종목 토론방 변수 산출
웹 시각화 페이지 제작

Member 4















한성용

기획 배경 조사
데이터 전처리
말뭉치 사전 작성

프로젝트 소개

프로젝트 사용 기술

사용언어	개발 툴	API	운영체제
 Python	 Jupyter	 OpenDart	 Windows10
 JavaScript	 VSCode		

라이브러리					
 NumPy	 Pandas	 ScikitLearn	 Pykrx	 LightGBM	 CanvasJS
 TA Lib	 TensorFlow	 Keras	 Bs4	 Selenium	 ChartJS



CONTENTS

01 기존 서비스 및 수요 분석

02 저평가 우량주 선정 모델 구축

03 데모 시현 및 개선 방안

01

기존 서비스 및 수요 분석



시장 환경 분석

주식 거래 열풍

<https://news.einfomax.co.kr/news/articleView> ▼

'주식 열풍'에 한국거래소, 작년 영업익 1천979억원...전년비 2배 ...

2021. 3. 31. — (서울=연합인포맥스) 정선영 기자 = 주식투자 열풍에 한국거래소의 지난해 별도 기준 영업이익이 두 배 이상 급증했다. 31일 한국거래소에 따르면 ...

<https://imnews.imbc.com> > 뉴스투데이

[뉴스터치] 주식 열풍에 주식 소유자 300만 명 급증 - MBC 뉴스

다음 소식 볼까요? "저도 주주예요" 1천만 시대" 제 주변에도 적은 금액으로 주식을 처음 시작하신 분 정말 ...

2021. 3. 31.

투자 정보 수요 급증

<http://m.journalist.or.kr> > m_article ▼

주식투자 열풍, 주식콘텐츠 훈풍 - 한국기자협회

2021. 3. 23. — 서점가에선 주식투자 비법을 알려주는 책이 베스트셀러 상위권을 차지하고 있고, 유튜브엔 '주린이'(주식 초보자를 어린이에 빗댄 말)를 겨냥한 강의 ...

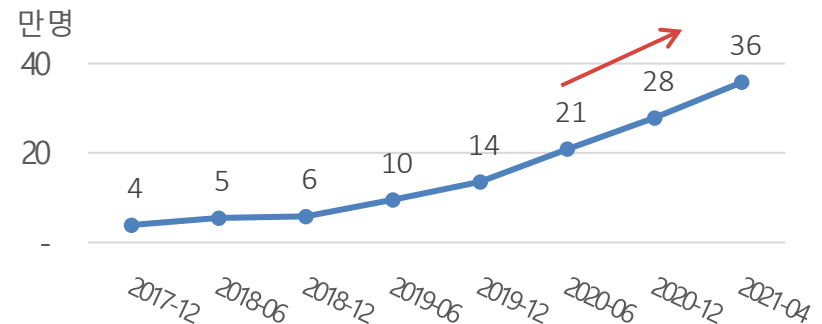
<https://news.mt.co.kr> > mtview ▼

2030세대 주식열풍...그러나 대부분 실전투자 지식·경험 부족 ...

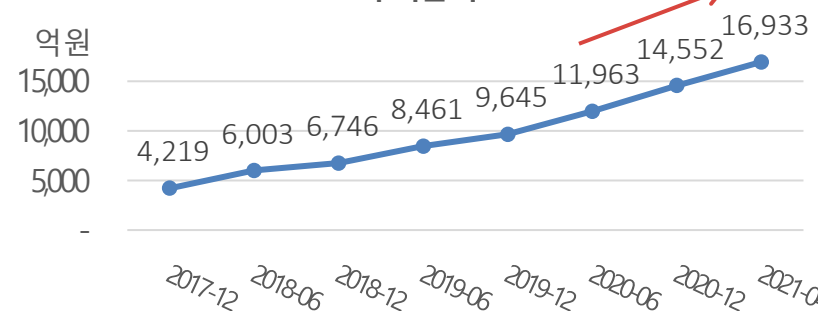
2021. 1. 29. — 특히 지금까지 국내 지수 상승을 보면 하락에 대한 대비가 필요하며, 주식초보자와 예비투자자들은 모의투자를 통해 정확한 프로그램 사용법과 자신 ...

로보어드바이저 시장 규모

가입고객수



투자금액



자료 : 로보어드바이저 테스트베드 센터

기존 서비스 분석

경쟁 상품 분석

■ MK 라씨로

차트를 통한 실시간 매매신호 포착

국내 27개 증권사 리서치 리포트 분석

관련 뉴스/공시 등 수집 분석

■ 금융社 로보 어드바이저

포트폴리오 자동구성/자동 매매 및 리밸런싱

주식 외 펀드/채권 등의 상품 추천

관련 뉴스/공시 등 수집 분석

■ 쿼트킹

종목 주가/실적 등의 쿼트 분석

종목비교 및 현황



Pain Point

■ 비정형 데이터 분석의 부재

기존 방식

정량적 접근



AI 활용
Opinion 분석 필요

■ 소액 투자자를 위한 서비스 부재

기존 서비스

포트폴리오
관리



포트폴리오 형성이 힘든
소액 투자자를 위한
서비스의 필요

02

저평가 우량주 선정 모델





Ideas

미디어 데이터 기반 모멘텀 분석 + 접근성 강화

정형 데이터

Point of Parity

- 경제지표, 재무제표 등을 이용한 재무 분석
- 차트를 이용한 추세, 모멘텀 분석
- 주가정보 관련 분석



비정형 데이터

Point of Difference 1

- 뉴스기사 감정/빈도 분석
- 뉴스기사 키워드 분석
- 주식커뮤니티 유저 게시글 감정/빈도 분석



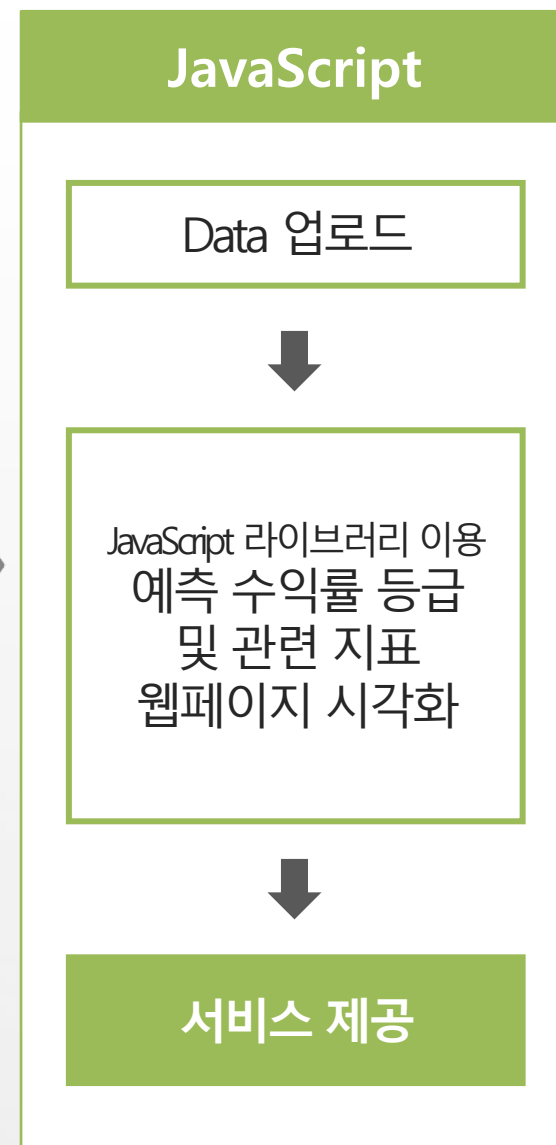
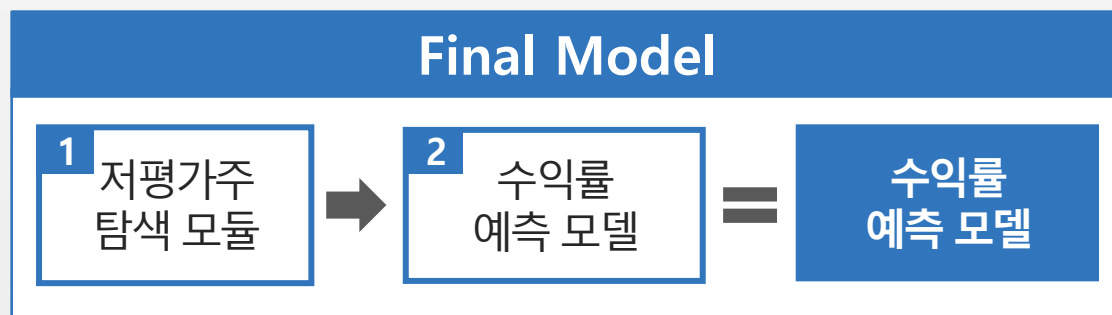
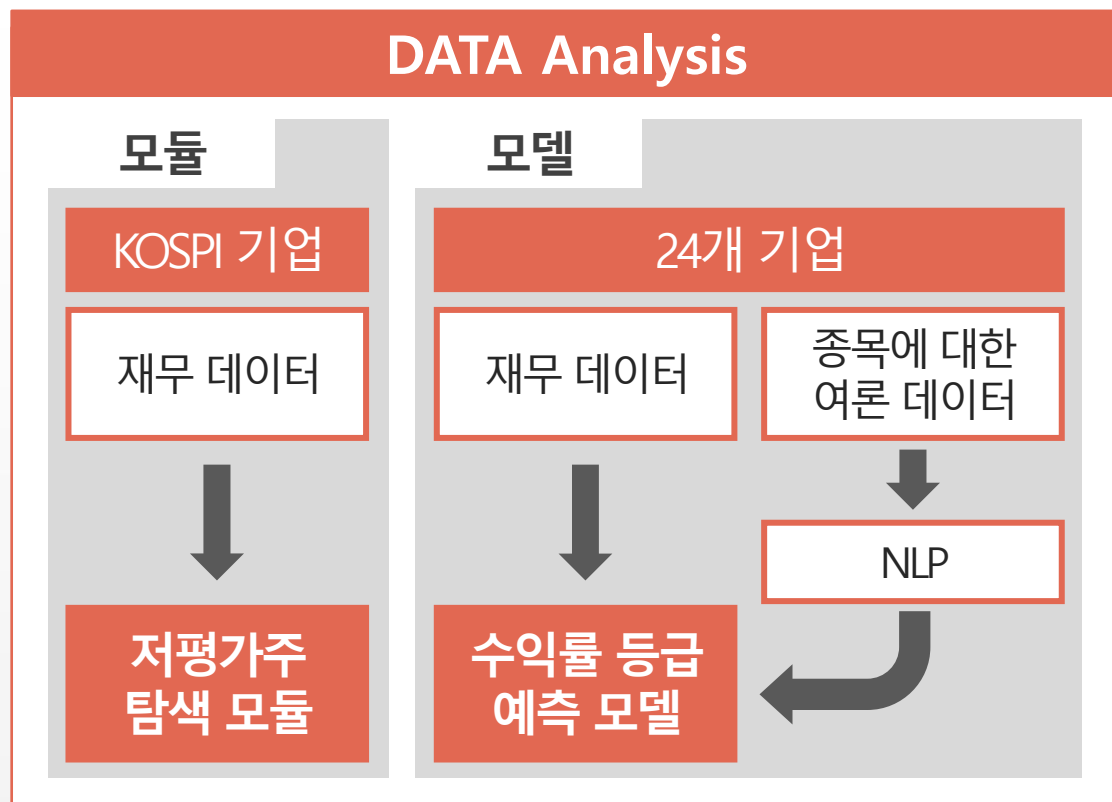
이용 고객 확대

Point of Difference 2

- 자산 관리가 아닌 주식 거래에만 초점을 맞춰 개별 종목에 대한 정보 제공
- 사용자 편의를 위한 다양한 정형/비정형 데이터 시각화 제공



저평가 우량주 선정 모델 구축



저평가주 탐색 모듈 개발 저평가주 판단 기준

분석대상종목기준

- KOSPI 종목
- 상장 이후 3년 경과 종목
- 우선주/지주사 제외
- 금융사 제외

저평가선정기준

- 순이익 기반 : PER 사용
 $0 < \text{PER} < 10$
종목 PER < 동일 업종 평균 PER
- 순자산 기반 : PBR 사용
 $0 < \text{PBR} < 1$
종목 PBR < 동일 업종 평균 PBR
- 기업 성장성 기반
매출액 증가율 > 0
 $5\% < \text{ROE} < 20\%$

저평가 우량주 선정 모델 구축

저평가주 탐색 모듈 개발 판단 기준 설정 및 결과

티커	NAME	SECTOR	PER	PER(SECTOR)	PBR	PBR(SECTOR)	ROE
095570	AJ네트웍스	도로와철도운송	4.039062	37.250000	0.580078	1.019531	6.964844
155660	DSR	비철금속	9.906250	19.515625	0.560059	1.157227	17.687500
006360	GS건설	건설	6.621094	18.703125	0.750000	1.370117	8.828125
025000	KPX케미칼	화학	7.328125	24.984375	0.649902	1.859375	11.273438
058860	KTis	상업서비스와공급품	9.781250	17.781250	0.549805	1.702148	17.796875
...
020000	한섬	섬유,의류,신발,호화품	9.921875	30.046875	0.910156	1.067383	10.898438
213500	한솔제지	종이와목재	7.808594	8.937500	0.529785	1.327148	14.742188
004960	한신공영	건설	2.910156	18.703125	0.409912	1.370117	7.097656
013520	화승코퍼레이션	자동차부품	9.710938	49.031250	0.899902	0.945312	10.789062
032560	황금에스티	비철금속	4.949219	19.515625	0.479980	1.157227	10.312500

- 1 Pykrx 라이브러리 및 네이버 증권 크롤링 활용
- 2 입력된 날짜 기준 가장 최근 영업일 데이터를 산출
- 3 종목 이름, 코드 및 선정 데이터 기반 DataFrame 반환



선정된 저평가주를
수익률 등급 예측 모델에 적용하여
우량주 여부 판단

저평가 우량주 선정 모델 구축



수익률 등급 예측 모델 개발 분석 대상 기업

■ 분석 대상 기업 선정 기준

KOSPI 종목

상장일 3년 이상 종목

우선주/지주사/금융사 제외

우량 섹터별 대/중/소형 종목으로 선정
(총 19개 섹터, 회사 크기별 2~3개의 종목)

	산업군	기업명(45개)
1	건설	GS건설, 현대건설, 삼부토건
2	화학	LG화학, 진양화학
3	철강	POSCO, 현대제철, 조선선재, 하이스틸
4	전기제품	삼성SDI, 삼화전기, KH필룩스

	산업군	기업명
5	조선	한국조선해양, 대우조선해양
6	석유와가스	SK, 극동유화, 미창석유, 대성산업
7	백화점과 일반상점	롯데쇼핑, 세이브존 I&C
8	가스유틸리티	한국가스공사, 서울가스
9	양방향미디어와 서비스	NAVER, 카카오
10	제약	삼성바이오로직스, 한미약품, 동화약품
11	항공사	대한항공
12	반도체와 장비	삼성전자, 유니퀘스트
13	호텔, 레스토랑, 레저	호텔신라, 이월드
14	비철금속	고려아연, 삼아알미늄, 조선내화
15	전자장비와 기기	LG이노텍, 일진머티리얼즈, 씨니전자
16	항공화물운송과 물류	현대글로벌, 세방, CJ 제일제당
17	자동차 부품	현대위아, 계양전기
18	전기유틸리티	한국전력
19	다각화된 통신서비스	KT

수익률 등급 예측 모델 개발 정형 데이터 독립변수

주가관련지표	기술적분석 지표	재무분석 지표
주가정보기반 OHLCV data (종가, 거래량 등) PER, PBR, EPS, BPS, DVI, DPS 등	차트분석기반 추세(MA) 모멘텀(ADX 등)	재무분석기반 성장성 지표 수익성 지표 안정성 지표 활동성 지표

분석 목표

최적합 분석지표 설정



상관분석, Tree plot, 변수 중요도 활용

수익률 등급 예측 모델 개발 정형 데이터 독립변수

■ 정형 데이터 관련 독립 변수 (24개)

기반 정보	지표	컬럼명	설명	활용 데이터
주가 기반 정보		SECTOR	종목별 섹터 구분(Label Encoding, 범위 0~17)	WCS 업종 분류 활용
	종가	total_Variable_list	종가의 1개월 기준 분산	Pykrx 라이브러리 활용
	거래량	VOLUME	일일 거래량	
	배당 관련	DIV	배당률	
		DPS	주당 배당금	
	주가 적정성	PER(SECTOR)	종목 섹터별 일일 기준 주가수익비율	
		PBR(SECTOR)	종목 섹터별 일일 주가순자산비율	
		PBR(MARKET)	KOSPI 전체 기업 평균 일일 주가순자산비율	
	투자자별 거래금액	INSTITUTION(NP)	기관 일일 순매수 거래금액	
		CORP(NP)	기타법인 일일 순매수 기준 거래금액	
		FOREIGN(NP)	외국인 일일 순매수 거래금액	

SECTOR, total_Variable_list 컬럼 제외 Standard Scaler 적용
total_Variable_list 컬럼 MinMaxScaler 적용

02 저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 정형 데이터 독립변수

■ 정형 데이터 관련 독립 변수 (24개)

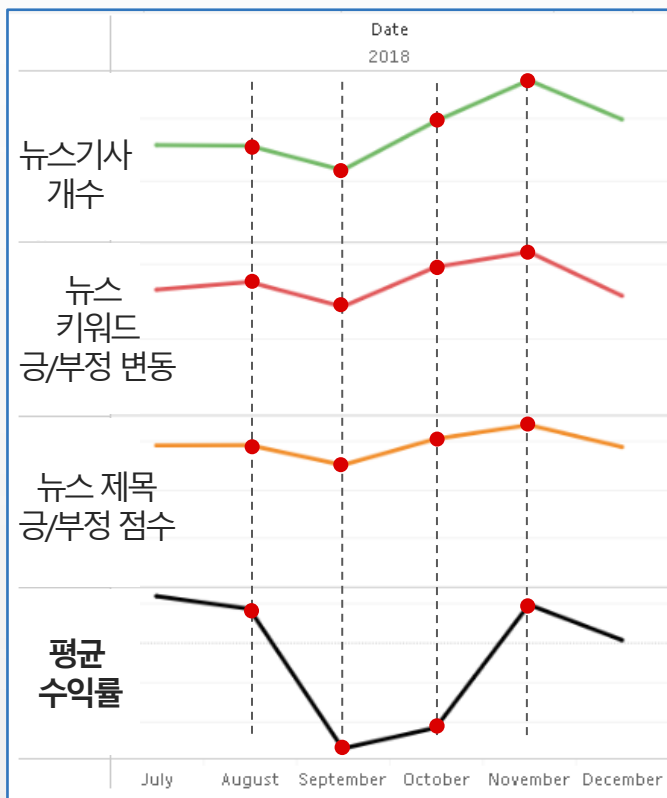
전체 칼럼 Standard Scaler 적용

기반 정보	지표	컬럼명	설명	활용 데이터
재무제표 기반 정보	성장성	ASST_INC	총자산 증가율	Dart API 활용
		REV_INC	매출액 증가율	
		S_ASST_INC	자기자본 증가율	
	수익성	REV_BPR	매출액 대비 영업이익률	
		RA_BPR	경영자산 대비 영업이익률	
	안정성	R_RATIO	유동비율	
		F_RATIO	부채비율	
	활동성	ASST_TO	총 자본 회전율	
		SA_C_TO	매출채권 회전율	
		ST_TO	재고자산 회전율	
차트 기반 기술적 정보	추세	MA10	10일 이동평균지수	Talib 라이브러리 활용
	모멘텀	ADX	평균 방향성 운동 지수	
		WILLR	WILLIAMS%R	

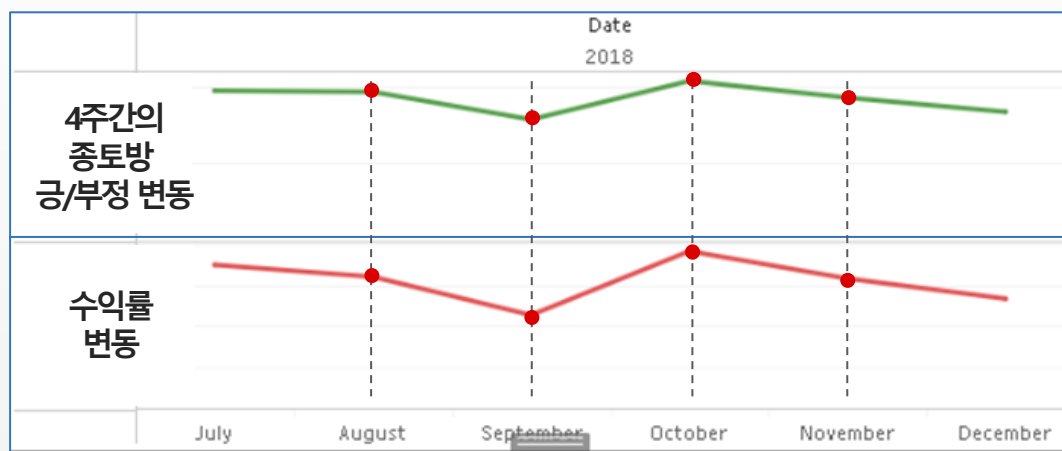
저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 비정형 데이터 독립변수

1주차 뉴스기사 데이터와 평균 수익률 비교



종목토론방 금/부정 변동과 주가 변동 비교



주가 수익률 추세와의 유사성 발견

저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 비정형 데이터 독립변수

종목 토론방 분석

감정 분석

■ 4주간 금/부정 변동

4주 간의 종목 토론방 유저 게시글의
금/부정 변동 지수 반영

빈도 분석

■ 4주간 글 수 변동

4주 간의 종목 토론방 유저 게시글의
개수 변동 반영

뉴스기사 분석

감정 분석

■ 주차별 기사 본문 금/부정 키워드 개수 변동

자체 키워드 단어사전을 제작하여
뉴스기사 본문 추출 후
금/부정 키워드 개수 주차별 반영

■ 주차별 기사 제목 금/부정 점수

주차별 뉴스기사 제목 금/부정 점수 반영

빈도 분석

■ 주차별 뉴스기사 개수 합계

주차별 뉴스기사 개수 합계 반영

각 주별(1~4주) 칼럼
PCA 차원 축소 시행
[12개 칼럼 → 4개 칼럼]

생성 변수
(14개)

4주간 종목토론방 금/부정 변동
4주간 종목토론방 글 수 변동

2개

최종 변수
(6개)

4주간 종목토론방 금/부정 변동
4주간 종목토론방 글 수 변동

2개

Week1 금/부정 키워드 개수	Week1 제목 금/부정 점수	Week1 기사 개수
Week2 금/부정 키워드 개수	Week2 제목 금/부정 점수	Week2 기사 개수
Week3 금/부정 키워드 개수	Week3 제목 금/부정 점수	Week3 기사 개수
Week4 금/부정 키워드 개수	Week4 제목 금/부정 점수	Week4 기사 개수

12개

df_pca1

df_pca2

df_pca3

df_pca4

4개

저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 비정형 데이터 독립변수

■ 비정형 데이터 관련 독립 변수 (6개)

네이버 종목별 종목 토론방 게시글 활용

4주간 종목토론방 금/부정 변동

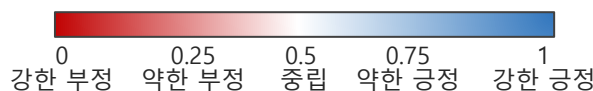
MinMaxScaler 적용

$$\text{Sentiment_moving} = (\sum_{i=1}^3 |(i+1)\text{주 前 평균 긍정도} - i\text{주 前 평균 긍정도}|)$$

금/부정 평가
수기 라벨링 진행



여론 긍정도 도출 : 0~1 사이 지수



금/부정 라벨링용
LSTM 지도학습 진행

LSTM 훈련 데이터 : 9천개
LSTM 평가 결과 : 73%
(테스트 데이터 : 3천개)

4주간 종목토론방 글 수 변동

MinMaxScaler 적용

$$\text{Post_count_moving} = (\sum_{i=1}^3 |(i+1)\text{주 前 총 게시글 수} - i\text{주 前 총 게시글 수}|)$$

종목토론방 1~4주차별 게시글 합계 변동 개수 반영

저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 비정형 데이터 독립변수

■ 비정형 데이터 관련 독립 변수 (6개)

빅카인즈 뉴스기사 검색 조건 내 전체 신문사 뉴스기사 활용(스포츠 기사/중복 기사 제외)

주차별 뉴스기사 긍/부정 키워드 개수 변동

Keyword_count_diff = 1~4주별 긍정 키워드 수 - 1~4주별 부정 키워드 수

경제 용어 관련
자체제작 단어사전 구축



뉴스기사 본문
키워드 추출



추출 키워드 단어사전 매칭

단어사전

■ 515개 키워드로 구성

긍정 키워드 : 215

부정 키워드 : 300

예시

긍정	부정
강세	하락
유입	낙폭
호조	제한
증가	실망

저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 비정형 데이터 독립변수

■ 비정형 데이터 관련 독립 변수 (6개)

빅카인즈 뉴스기사 검색 조건 내 전체 신문사 뉴스기사 활용(스포츠 기사/중복 기사 제외)

주차별 뉴스기사 개수 합계

News_count = 1~4주별 뉴스기사 개수 합계

1~4주차별 뉴스기사 개수 합계 반영

주차별 뉴스기사 제목 긍정/부정 점수

NewsTitle_Sentiment_Score = 1~4주별 긍정 + 중립 + 부정점수

긍/부정 평가
수기 라벨링 진행



긍/부정 점수 계산
긍정(1) / 중립(0) / 부정(-1)



긍/부정 라벨링용
LSTM 지도학습 진행
LSTM 훈련 데이터 : 약 1만개
(Test_set : 1,500개 별도)
LSTM 평가 결과 : 90%

02 저평가 우량주 선정 모델 구축

수익률 등급 예측 모델 개발 모델 종속변수 및 독립변수

1개 컬럼 종속변수

1개월 이후 수익률 기준

구분	기준
0	주가 하락
1	주가 0~10% 상승
2	주가 10% 초과 상승

30개 컬럼 독립변수

각 컬럼 간의 상관관계 지수를 0.5 이하 기준으로 독립변수 선정

	SECTOR	VOLUME	DIV	DPS	PER(SECTOR)		df_pca1	df_pca2	df_pca3	df_pca4
0	1	-0.111912	-0.289877	-0.580378	-0.480647		-0.318137	-0.267250	0.252172	1.277612
1	1	-0.161502	-0.273122	-0.580378	-0.482474		-0.360404	-0.211943	0.453968	0.442214
2	1	0.129258	-0.306632	-0.580378	-0.479002		-0.236006	-0.309300	0.618335	0.025837
3	1	-0.109927	-0.312217	-0.580378	-0.478819		-0.107428	-0.271747	0.817332	0.140215
4	1	-0.133340	-0.317802	-0.580378	-0.478179	...	-0.013037	-0.331180	0.890547	-0.129883
...
18445	11	-0.269615	-1.043854	-0.713182	3.265064		-1.447825	0.325826	0.342073	-1.012987
18446	11	-0.271321	-1.043854	-0.713182	3.228326		-1.605257	0.462321	0.643844	-0.204876
18447	11	-0.248468	-1.043854	-0.713182	3.156221		-2.247791	0.935961	-0.202337	1.125298
18448	11	-0.238792	-1.043854	-0.713182	3.391727		-2.247791	0.935961	-0.202337	1.125298
18449	11	-0.260413	-1.043854	-0.713182	3.397484		-2.032317	0.658017	-0.391213	0.437278

수익률 등급 예측 모델 개발 모델 검증 및 튜닝

■ GridSearchCV 이용 하이퍼 파라미터 튜닝 기준

Model	Accuracy	Hyperparameter
LGBM Classifier	0.856	<ul style="list-style-type: none">- n_estimators : 300- min_data_in_leaf : 7- max_depth : 40- num_leaves : 1,000
DecisionTree Classifier	0.771	<ul style="list-style-type: none">- criterion : gini- max_depth : None
RandomForest Classifier	0.839	<ul style="list-style-type: none">- n_estimators : 450- min_samples_leaf : 5- max_depth : 25- min_samples_split : 15

03

데모 시현 및 개선 방안





JavaScript 데모 시현

삼성전자

회사 이름 (대소문자 구분)

종목 변경

종목 수익률 등급

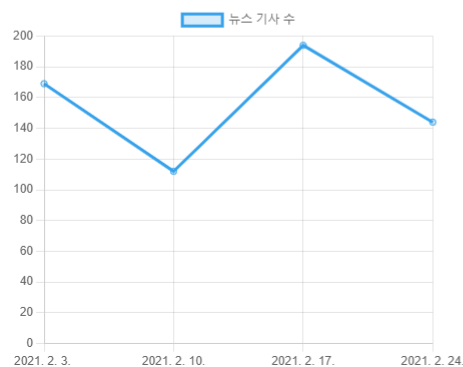
회사 정보 및 분석

C

등급



뉴스 기사 수 (4주간)



주요사업 반도체와반도체장비

시가총액 501조4620억원 (2021-03-03 기준)

액면가 100

주식 수 5969782550

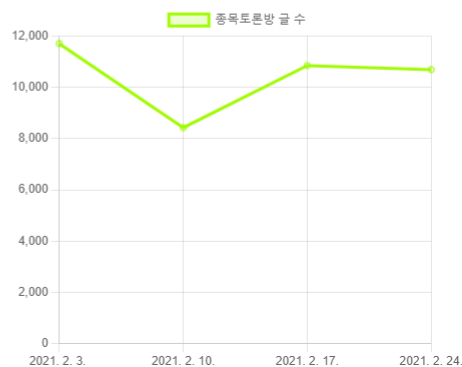
뉴스
긍정도 56.6% (2021-2-3 ~ 2021-03-03)



종토방
긍정도 41.5% (2021-2-3 ~ 2021-03-03)



종목토론방 글 수 (4주간)





서비스 지향점 및 개선방안

■ 훈련이 되지 않은 종목의 경우 예측력 저조

모델의 한계 I

전 종목 대상 학습용 셋을 만들기에 **시간적 제약 존재**
주제에 적합하다고 생각되는 **일부 종목을 샘플링하여 분석 및 예측 진행**

개선방안

시장 전 종목에 대한 훈련세트 구축

■ 비정형 데이터가 부족한 종목의 경우 예측력 저조

모델의 한계 II












일부 종목의 비정형 데이터 수집의 한계 존재

개선방안

비정형 데이터가 부족한 종목에 대해
정형 데이터와 LSTM을 활용한 시계열 예측을 시행하여
데이터 부족을 극복할 수 있는 대안 마련



Git Hub Link

 lignas12015 project3		a21f08a yesterday	🕒 1 commit
	.vscode	project3	yesterday
	dataset	project3	yesterday
	fsdata	project3	yesterday
	종토방모델제작및금부정평가	project3	yesterday
	종토방크롤링	project3	yesterday
	QuantativeAnalysis.py	project3	yesterday
	Quantative_Analysis.ipynb	project3	yesterday
	find_undervalued_stock.ipynb	project3	yesterday
	undervaluedstock.py	project3	yesterday
	금부정평가 및 label, 분류모델.ipynb	project3	yesterday

<https://github.com/lignas12015/project3>

THANK YOU

