자연어 처리 기반의 투자분석 및 예측시스템 개발

Turnaround

About

Object	주가를 예측하는 로보 어드바이저 개발
Mentor	정좌연 PE
Mentee	이지훈, 이문형, 강민재, 구병진, 김서정
Team Name	Turnaround

Index

Part 1

- 프로젝트 목적 및 방법
- 일정계획
- 개발환경 및 라이브러리
- 시스템 기술명세서
- 관련지표

Part 2

- 자연어처리
- 주가예측
- 평가방법

Part 3

- 결과
- 한계점
- 향후연구과제
- Q&A

Schedule

STEP.1 분석 & 기획	STEP.2 설계	STEP.3 구현	STEP.4 평가
2020.08.03~	2020.09.25~	2020.10.15~	2020.11.3
도메인 관련 지식 습득	데이터 수집 및 전처리	알고리즘 구현	알고리즘 최적화 및 평가
종목별 주가 분석	NLP 및 머신러닝, 딥러닝 기반의 주가 예측 알고리즘 설계	프로토 타입 구현	

기술적 분석

주식 시장에 나타난 과거의 데이터를 기초로 활용하여 시세를 예측하는 방법

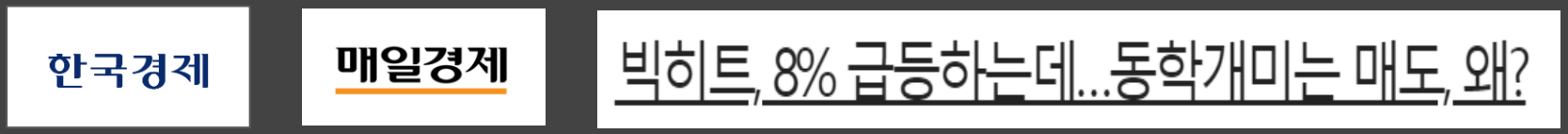
주가, 거래량, 보조지표(이동평균선 등)들을 활용

기술적 분석의 한계

- 투자자의 주관적인 심리 작용
- 작전 세력, 기사 등에 의해 예측하지 못하는 변수 발생 가능성

뉴스기사 분석

관련 종목의 뉴스 기사에서 데이터를 수집해 시장을 예측하는 방법



뉴스 기사 분석의 **한계**

- 모든 뉴스를 분석하는 것이 **불가능**
- 기사의 내용이 주식 시장에 긍정적인지 부정적인지 파악하기 **어려움**
- 뉴스 분석을 하는 **투자자의 심리**가 크게 작용

머신러닝, 딥러닝 활용

기술적 분석의 한계

뉴스 기사 분석의 한계

인사이트 도출

Hidden Pattern

Objectivity

기존 분석 방법의 한계를 넘어서

- 기술적 분석과 기사 분석을 주식 투자 의사결정에 활용하는 것이 적합한지 확인
- 실제 주식 투자에 활용하기 위함

Process

1. 데이터 수집 (정형 데이터, 비정형 데이터)

2. 데이터 전처리

3. 예측 모델 생성

4. 예측 모델 비교 및 평가

Target

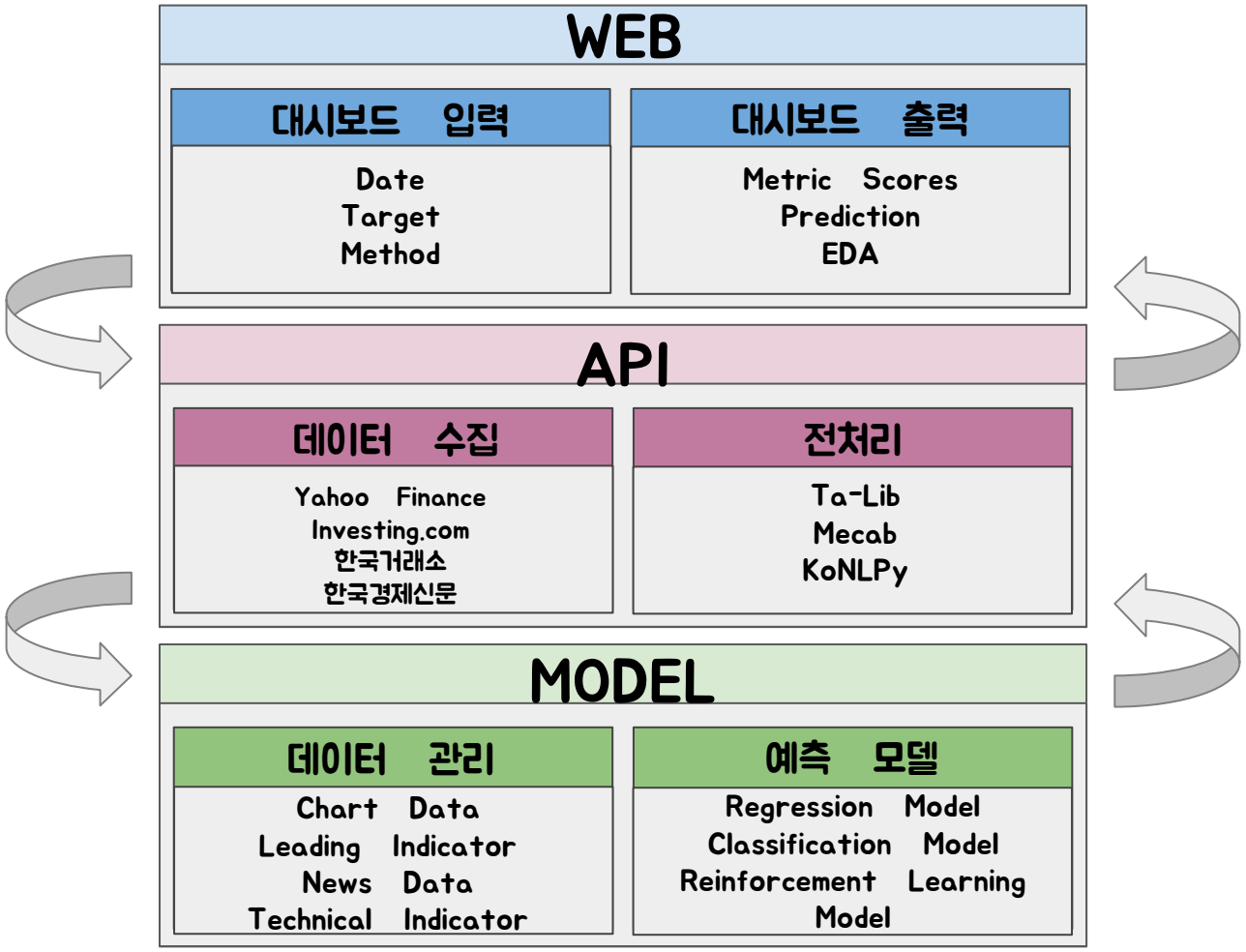
데이터 수집 기간	2018. 01.01 ~ 2020. 10.26 (API 구축, 데이터 수집)
Train Data	2018.01.01~2020.02.29 (70%)
Test Data	2020.03.01~2020.10.26 (30%)
Feature Description	Chart data Investment index data Stock price index data Exchange rate data Raw material data Treasury data Global investment index data Stock technical indicator News data Black pink album release date

개발환경 및 라이브러리

Major Libraries & Frameworks

Windows 10
Google Drive
GitHub
PyCaret
Streamlit
TensorFlow
PyTorch
Keras
Python
Pandas
ANACONDA
jupyter
MeCab
User Dictionary
Google BERT
KoNLPy
Excel
LightGBM
colab
PyCharm
zoom
NVIDIA CUDA
YAHOO! FINANCE for python
investpy
PYSTOCK
PROPHET

시스템 기능명세서


WEB
대시보드 입력
Date
Target
Method
대시보드 출력
Metric Scores
Prediction
EDA
API
데이터 수집
Yahoo Finance
Investing.com
한국거래소
한국경제신문
전처리
Ta-Lib
Mecab
KoNLPy
MODEL
데이터 관리
Chart Data
Leading Indicator
News Data
Technical Indicator
예측 모델
Regression Model
Classification Model
Reinforcement Learning
Model

Background Knowledge

Background Knowledge

RSI

가격의 상승압력과
하락 압력의
상대적인 강도를
나타내는 지표

MACD

단기와 장기
이동평균선의 간격
차이를 통해 주가
흐름을 파악하는
보조지표

Background Knowledge

Stochastic

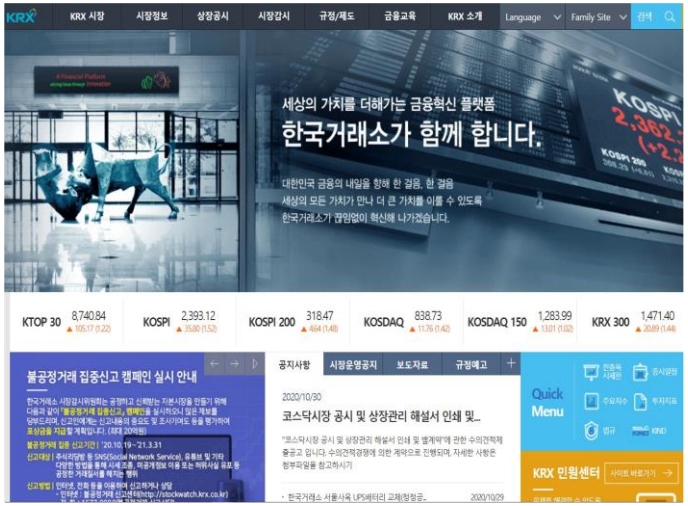
최근 N일간의 최고가와 최저가 사이 현재 주가의 위치를 알려주는 지표

CCI

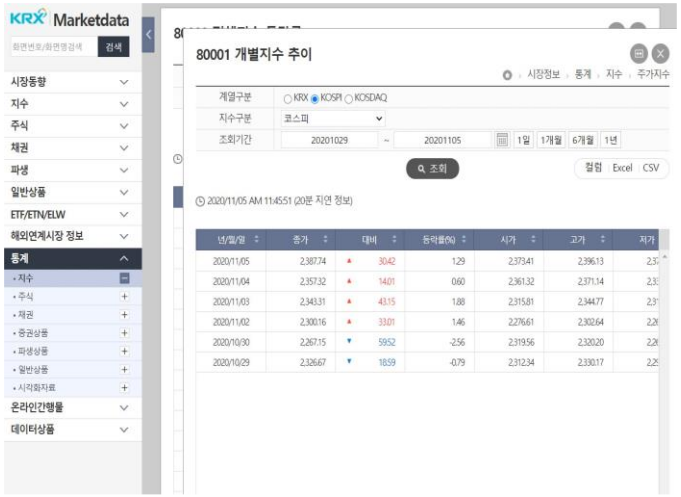
이동평균선과 현재주가의 차이로 이격도와 비슷하지만 값이 평균으로 수식이 다른 지표

데이터 수집

- Source : KRX, Yahoo Finance, Investing.com



크롤링



데이터 수집 기간 : 2018. 01. 01 ~ 2020. 10. 26

※ 주가 등락률을 확인하여 전일 기사의 라벨링에 반영

데이터 수집

- Source : 한국경제신문

크롤링

데이터 수집 기간 : 2019. 01. 01 ~ 2020. 10. 26
한국경제신문의 **경제, 국제** 면에서 기사 제목 추출
＊ 1일당 최대 100개의 기사 활용

형태소 분석기 : Mecab

"아버지가방에들어가신다"

Hannanum	Kkma	Komoran	Mecab	OKT
아버지가방에들어가 / N	아버지 / NNG	아버지가방에들어가신다 / NNP	아버지/NNG	아버지/Noun
이 / J	가방 / NNG		가/JKS	가방/Noun
시ㄴ다 / E	에 / JKM		방/NNG	에/Josa
	들어가 / VV		에/JKB	들어가신다/Verb
	시 / EPH		들어가/VV	다 / Eomi
	ㄴ다 / EFN		신다/EP+EC	

데이터 전처리

1. 불용어 및 단일 글자 처리

2. 복합 명사

정보 ➔ 정보 통신, 정보 시스템, 정보 기술, 정보 통신부

Unigram, Bigram 으로 사용자 사전 등록

3. 한자 처리

美 ➔ 미국　北 ➔ 북한　韓 ➔ 한국　中 ➔ 중국　日 ➔ 일본

4. 일반 명사

TF Top1000 기준, 약 165개의 단어를 15개 카테고리로 나누어 변환

5. 고유명사 처리

유명 인물: 트럼프, 오바마, 클린턴 ➔ 미국대통령

기업 이름: 고려화학, 금강제화 etc ➔ 한국기업

데이터 전처리

TF(Text Frequency) : Uni-gram

Word	TF
미국	214
금융	202
한국	186
경제	177
기업	152
달러	146
LG	127
일본	123
정보	123
구조	113

TF(Text Frequency) : Bi-gram

Word	TF
구조조정	95
정보통신	51
LG전자	38
시스템개발	36
금융기관	34
천만달러	31
벤처기업	29
한국경제	28
금리인하	28
금융위기	26

데이터 전처리

Sample

국제	처리 전		처리 후
	"美 상·하원 외교위원장, 하원 군사위원장 교체"	→	미국 상하 외교 위원장 하원 군사 위원장 교체
	트럼프 "모든 일 승리할 것…4년 더 있다"	→	미국대통령 모든 승리
	아베 "일본의 내일 여는 선두에 설 것"	→	아베 일본 내일 선두

경제			
	LG화학, 전기차 배터리용 탄소나노튜브 1200% 증설	→	한국기업 전기차 배터리 탄소 나노 튜브 증설
	현대차 'N'에 자극받은 토요타, 고성능 소형 확대	→	현대차 자극 토요타 성능 소형 확대
	8월 은행 예금금리 0.81%·대출금리 2.63%…또 최저	→	은행 예금 금리 대출 금리 최저

Model Description

Target	Regression Model	Classification Model	Reinforcement Learning
KOSPI	Prophet AutoML (TheilSen, Linear Regression, Ridge) -> Ensemble	AutoML (Gradient Boosting, Linear Discriminant, Ridge) -> Ensemble NLP (Bert, LSTM, AutoML)	
YG	Prophet AutoML (Linear Regression, RANSACR) -> Ensemble	AutoML (Gradient Boosting, Decision Tree, LightGBM) -> Ensemble NLP (Bert, LSTM, AutoML)	A2C (value & policy network LSTM)

Feature Description1

Chart Data

Date	Close	Start	High	Low	Volume
2020-10-05	2,358.00	2,330.55	2,364.73	2,327.83	763,618
2020-10-06	2,365.90	2,369.17	2,378.28	2,355.39	861,755
2020-10-07	2,387.94	2,350.82	2,387.45	2,347.82	737,994
2020-10-08	2,391.96	2,408.49	2,409.01	2,383.82	961,726
2020-10-12	2,403.73	2,404.18	2,409.42	2,393.74	843,574

Feature Description2

Technical Indicator

Bollinger Bands

Moving Average

Exponential Moving Average

Weighted Moving Average

Parabolic SAR

Three Line Strike

Evening Star

Abandoned Baby

On Balance Volume (OBV)

Average Directional Movement Index (ADX)

Commodity Channel Index (CCI)

Plus Directional Indicator (PlusDI)

Plus Directional Movement (PlusDM)

Relative Strength Index (RSI)

Stochastic

Williams' %R

Average True Range (ATR)

Variance

Feature Description3

Leading Indicator : KOSPI

Exchage Rate

- 달러/원 환율 종가
- 유로/원 환율 종가
- 엔/원 환율 종가
- 위안/원 환율 종가

Raw Material

- 금/은/동/플래티넘/팔라듐 시세 종가
- WTI유 시세 종가
- 가솔린 RBOB 시세 종가
- 천연가스,난방유 시세 종가

Treasury

- 미국 국채수익률 종가 (13주/5년/10년/30년)
- 한국 채권수익률 (3년/5년/10년/20년/30년)

Investment Index

- 투자자별 공매도 거래대금/거래량
- 코스피 대표지수 종가
- 코스피 섹터지수 종가

Global Stock Index

- Vix 종가
- KOSPI Volatility 종가
- Bitcoin USD 종가
- S&P 500 종가
- Dow Jones 종가
- NASDAQ 종가
- NYSE 종가
- EURONEXT 100 종가
- AMEX 종가
- Russell 2000 종가
- DAX 종가
- Nikkei 225 종가
- HANG SENG 종가
- SSE 종가
- ESTX 50 종가

Feature Description4

Leading Indicator : YG

Chart Data

- YG 차트 데이터
- YG 시가총액
- YG 거래대금
- YG 상장주식수

Investment Index

- 코스닥 종합지수
- 코스닥 대표지수
- 코스닥 섹터지수
- 코스닥 산업별 지수
- 코스닥 시가총액 규모별 지수
- 코스닥 소속부 지수

Fundamental Data

- YG 배당수익률 (DIV)
- YG 주당순자산가치 (BPS)
- YG 주가수익비율 (PER)
- YG 주당순이익 (EPS)
- YG 주가순자산비율 (PBR)
- YG 공매도 비중/금액/거래대금/거래량
- YG 잔고/잔고금액

Black Pink Data

- 블랙핑크 앨범 발매일자

Metric Scores 도출

Regression Model

MAE = Mean absolute error
MSE = Mean squared error
RMSE = Root mean Squared error
R^2 = Coefficient of Determination

Classification Model

Accuracy = (TP+TN)/(TP+TN+FP+FN)
ROC AUC = root(Sensitivity^2+Specificity^2)/root(2)
Recall = TP/(TP+FN)
Precision = TP/(TP+FP)
F1 = (2*Precision*Recall)/(Precision+Recall)

AutoML Process

Target 지정

1.Model Selection

2.Feature Selection

3.Hyper Parameter Tuning

4.Model Ensemble

Training Result

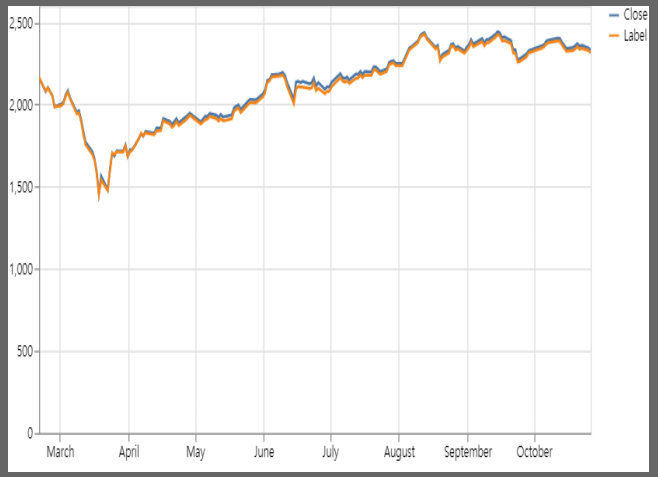
KOSPI Model

Regression Model

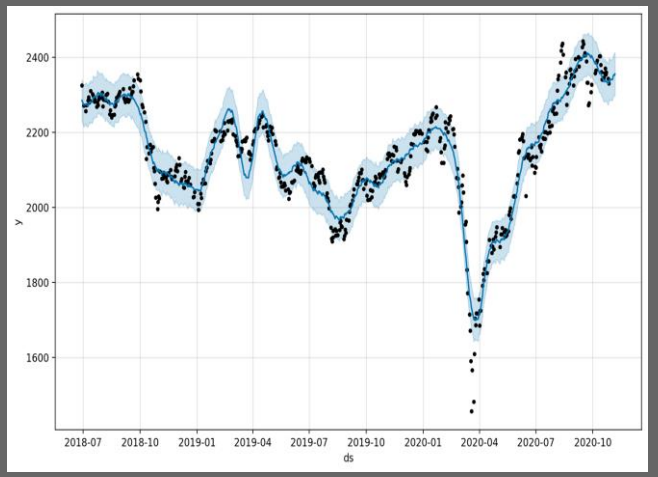
Test Metric Scores

	MAE	MSE	RMSE	R^2
Prophet	85.415	15,005	122.4969	-5.5035
AutoML	16.1176	311.3388	17.6446	0.9942

Actual & Prediction Graph



AutoML



Prophet

KOSPI Model

Classification Model

Test Metric Scores

	Accuracy	ROC AUC	Recall	Precision	F1
AutoML	0.9152	0.9081	0.8676	0.9219	0.8939
NLP Bert	0.8008	0.8205	0.7429	0.9246	0.8238
NLP LSTM	0.8005	0.8069	0.7645	0.8729	0.8151
NLP AutoML	0.8088	0	0.859	0.8171	0.6068

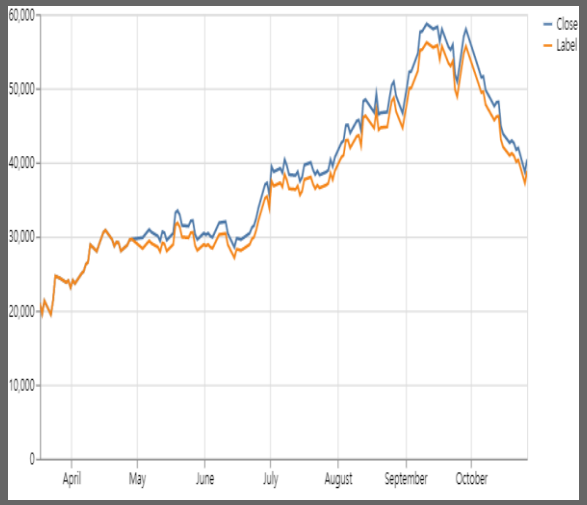
YG Model

Regression Model

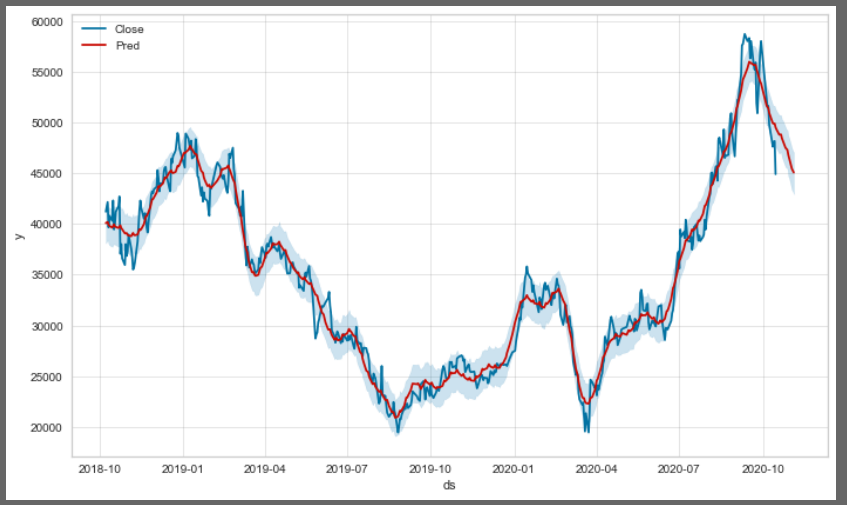
Test Metric Scores

	MAE	MSE	RMSE	R^2
Prophet	1686.042	4,849,041	2202.053	0.9422
AutoML	1,491.8399	2,835,922.830	1,684.3167	0.9724

Actual & Prediction Graph



AutoML



Prophet

YG Model

Classification Model

Test Metric Scores

	Accuracy	ROC AUC	Recall	Precision	F1
AutoML	0.8844	0.8782	0.8387	0.8814	0.8595

UI 구현

Dashboard Input

KOSPI 지수 및 YG 종목 주가 예측 모델

Date

20201027 - +

Target

YG

Method

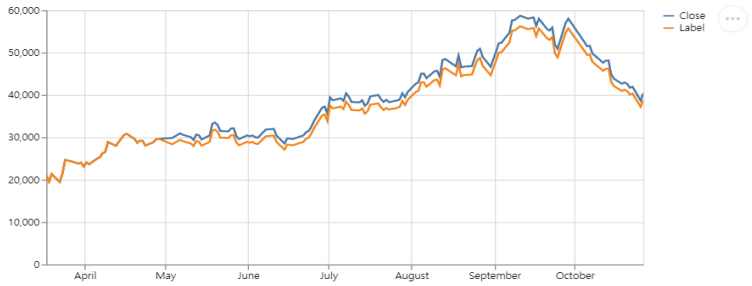
AutoML_REG

Dashboard Input

Test Data Metrics Score

	MAE	MSE	RMSE	R2
0	1,491.8399	2,836,922.8303	1,684.3167	0.9724

Forecast Data (Test Data)



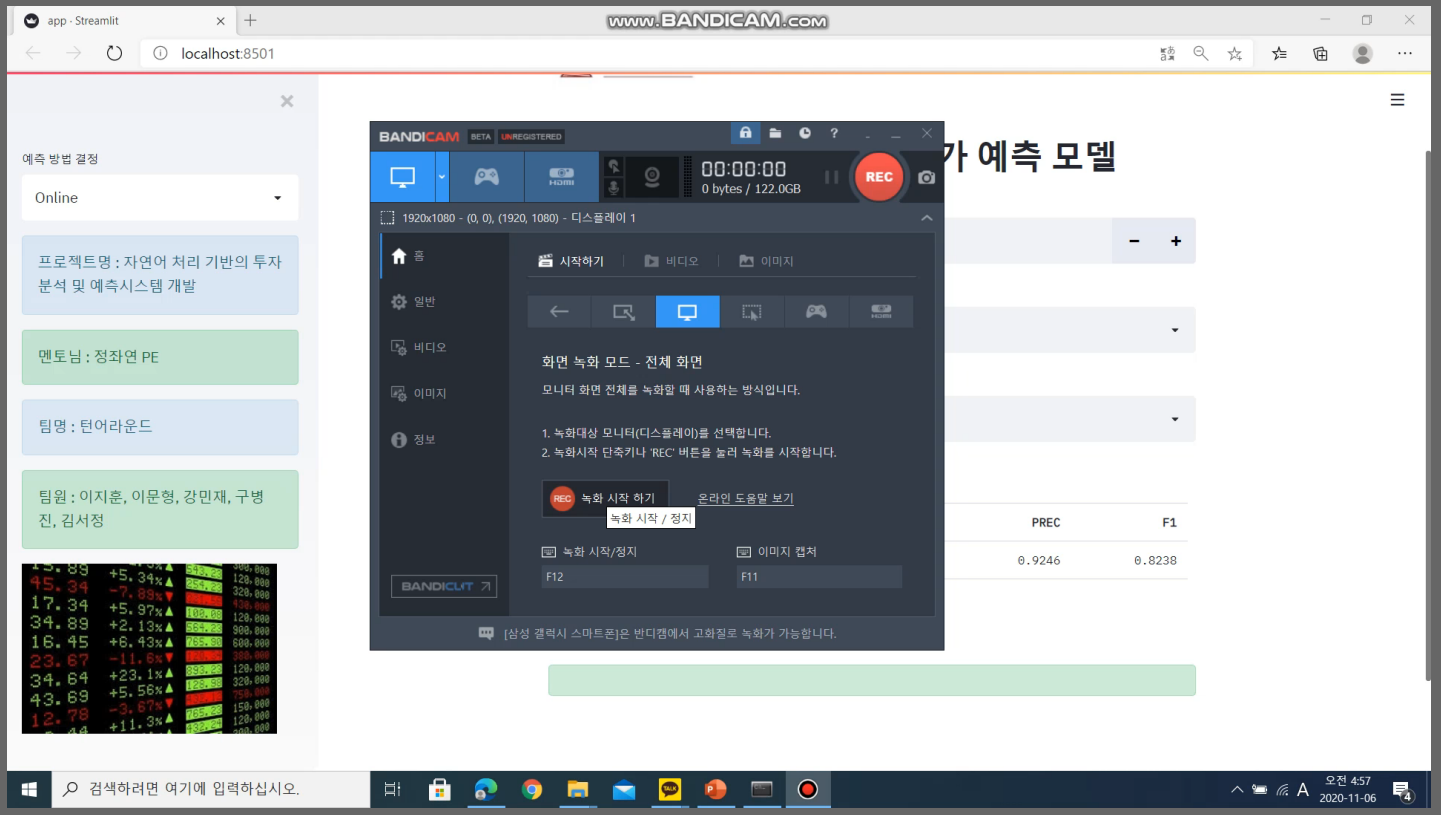
주가 예측

20201027주가 상승 예상 -> 매매 어드바이스 : 매수

WEB 시연 영상

Prediction Model

https://youtu.be/aDm0r-_bh3I



결론

선행지표, 보조지표 및 뉴스 데이터를 활용

↓

예측 정확도 상승

↓

선행지표, 기술적 분석 및 뉴스 분석을
주식투자 의사결정에 활용

한계점 및 향후계획

같은 단어가 긍정 의미일 수도 부정 의미일 수도 있다.

라벨링 방식의 한계점

Mecab의 동사, 복합 명사 추출 문제

모델 업데이트를 자주 해야 성능이 좋음
수익률 계산이 필요

머신러닝을 이용한
감성분석

주식시장에 특화된
감성 사전 구축 필요

감성사전의
지속적인 확장 필요

모델 업데이트 및
완전 자동화 필요

Turnaround

Reference

재무비율과 기술적 분석을 통한 AI 주식 트레이딩 알고리즘 모델링 (2019, 정해성, 김용현, 임한준, 정기백, 정진태, 최원화)

Introduction of Reinforcement Learning (2016, 곽동현)

뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형 (2012, 김유신, 김남규, 정승렬)

Deep Learning for financial applications : A survey (2020, Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, Omer Berat Sezer)

Teaching Machines to Read and Comprehend (2015, Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom)

Bert:Pre-training of Deep Bidirectional Transformers for Language Understanding (2019, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova)