특정 시간대의 주가 변동 패턴을 이용한 실시간 주가 예측

빅데이터 연합동아리 보아즈 ADV 프로젝트

15 기 분 석 김 상 휘
15 기 분 석 김 성 용
15 기 분 석 김 해 준

BOAZ

INDEX

BOAZ

01

프로젝트 배경

주식과 가상화폐 투자 열풍

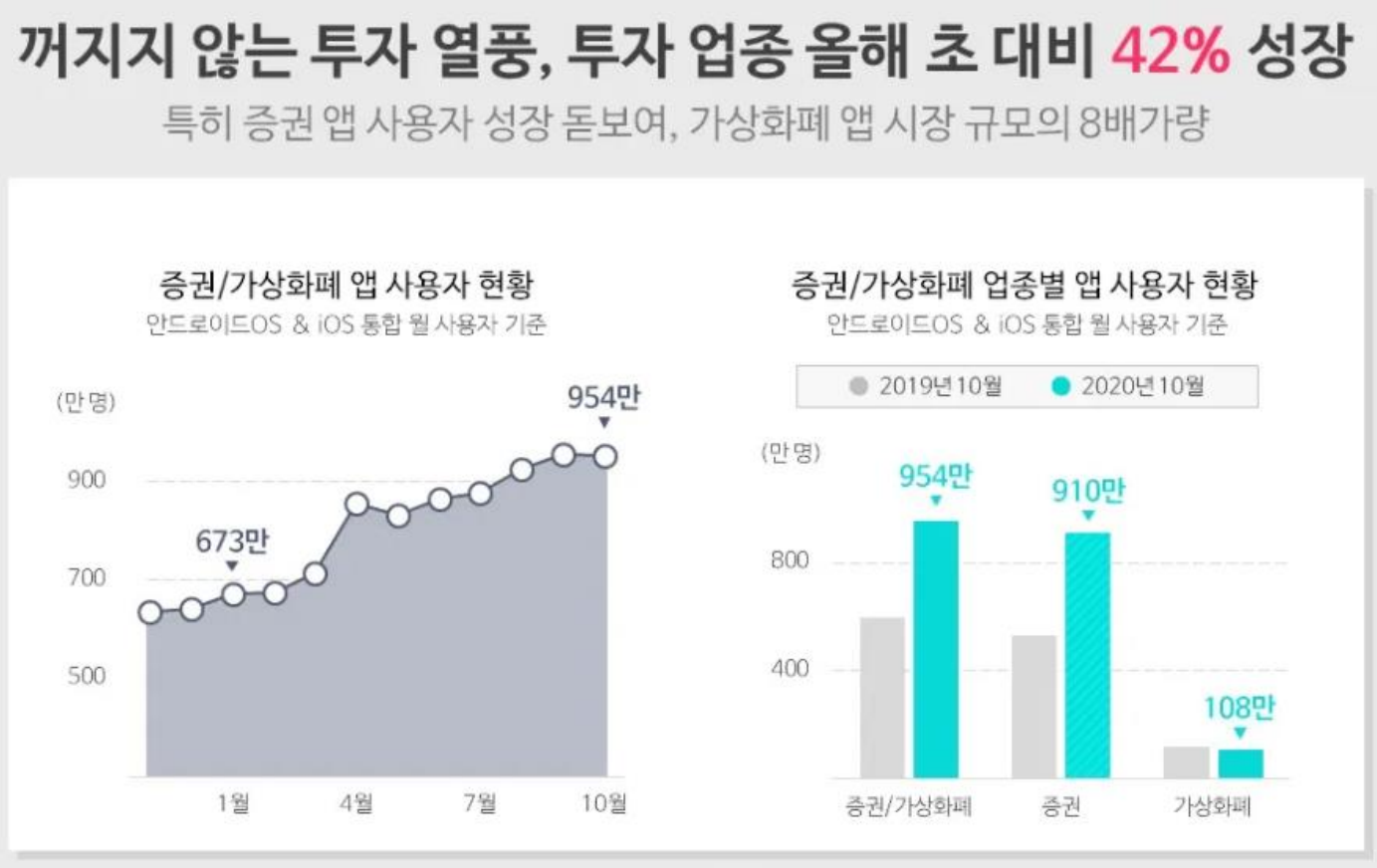


사진 출처 : https://wowtale.net/2020/12/05/mobile-securities-and-crypto-currency-app-market-in-korea/

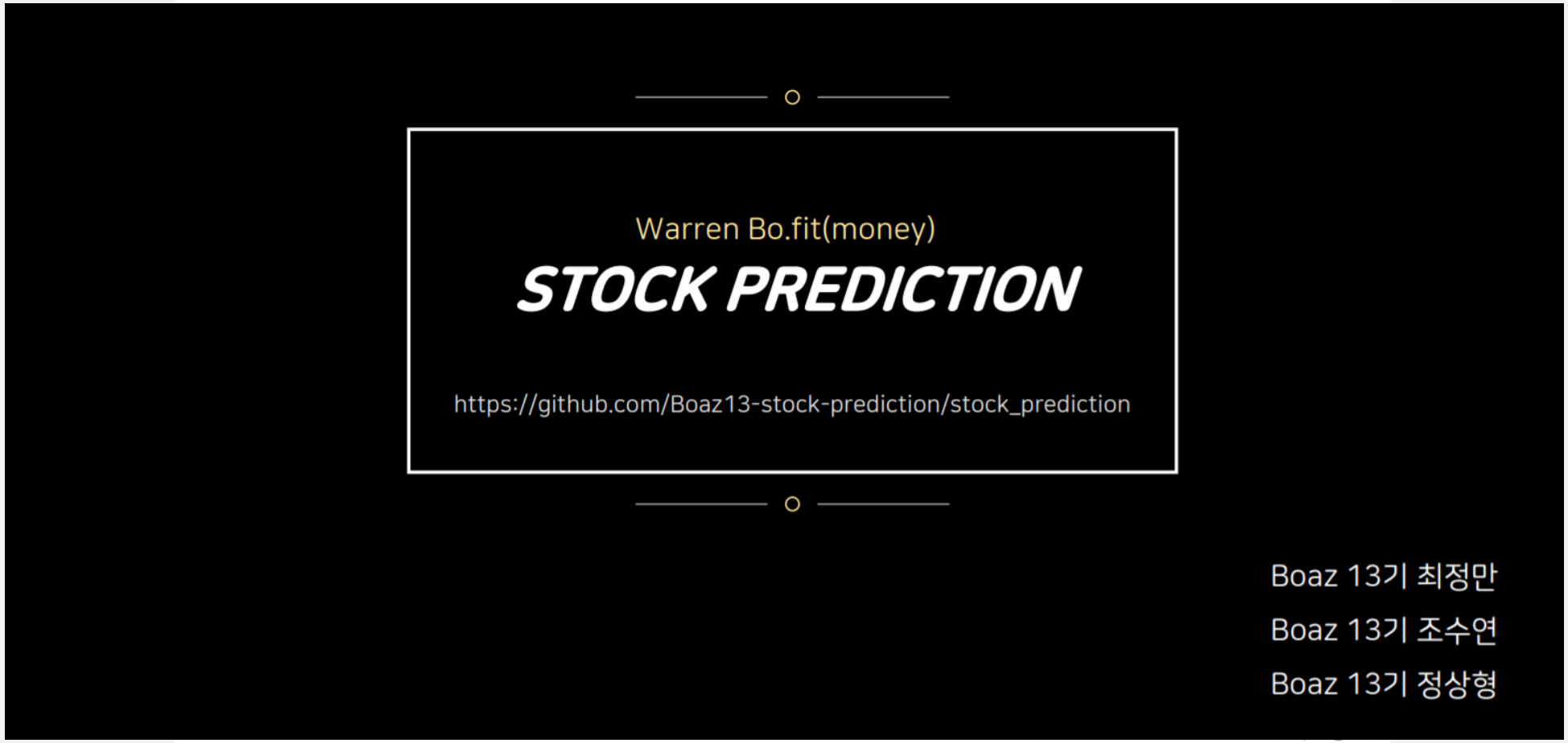
"쥐꼬리만 한 월급 받아서 월세와 생활비 내고 나면 남는게 없지 않나. **노동에만 의존하면 영원히 가난에서 탈출할 수 없을 것 같아** 마이너스 통장을 만들고, 투자처를 알아보고 있다"

회사원 정 모씨(36)

"**은행 저축으로 돈을 모으거나 청약통장으로 집을 사는 일은 현실적으로 너무 어렵기 때문**에 직장을 그만두고 전업 주식 투자자로 활동을 시작했다"

대학원생 이 씨(30)

수많은 기존의 주가 예측 프로젝트



수많은 기존의 주가 예측 프로젝트

"주식시장은 **불치의 감정적 문제**를 가지고 있어서
정신 분열적이고 비논리적인 행위를 매일 일삼는다"

"주식시장의 변동성은 합리적인 기대를 가지고 있는 것이 아니라, 이러한 **감정적인 변화**에 따라 수시로 변동을 부린다"

[Benjamin Graham]

우리가 가지는 차별성

" 주가 변동 패턴 "

기존 프로젝트 방향

- 주가에 영향을 미치는 최대한 많은 **변수 생성**
- 실시간 이슈를 반영하기 위한 **자연어 처리**
- RNN, LSTM 등의 **인공신경망**을 이용한 모델링
- 주가 변동의 기준이 **일(day)** 이 됨

우리 프로젝트 방향

- 변수에 집중하기 보다, **주가 변동 모양**에 집중
- 예측의 근거는 **비슷한 변동성**을 보인 과거 데이터
- 주가 변동의 기준이 **분(min)** 이 됨

02

프로젝트 목표

프로젝트 목표

9시부터 9시 30분까지

특정 시간대의 주가 변동 패턴을 이용한 실시간 주가 예측

9시 35분

→ 일일 매매 종목 제안

03

프로젝트 진행

프로젝트 개요 및 방법

프로젝트 개요

- 분별 주가 데이터 수집
- 전처리
- 비슷한 주가 변동 패턴 클러스터링
- 새로운 데이터가 들어왔을 때, 해당 데이터가 속하는 클러스터로 분류
- 분류된 클러스터의 데이터들을 이용하여 특정 시간의 주가 예측
- 최적의 투자 종목 추천

방법

- 크롤링 (Selenium)
- Standard Scale
- 시계열 클러스터링 (GMM)
- 분류 모델 (GMM+CNN)
- 예측 모델 (가중치 모델)
- 수익률 한계선

데이터 수집

[9:00, 9:01, 9:02, ..., 9:35]

[종목명_날짜]

	A	B	C	D	E	F	G	H
1		9:00	9:01	9:02	9:03	9:04	9:05	9:06
2	AK홀딩스_20210201	28500	28650	28450	28300	28150	27950	27800
3	AK홀딩스_20210202	29650	29700	29450	29700	29550	29600	30200
4	AK홀딩스_20210203	30150	29950	29900	29850	29900	29900	29900
5	AK홀딩스_20210204	29900	29800	29800	29800	29750	29650	29800
6	AK홀딩스_20210205	32450	32050	32050	32250	32400	32200	32150
7	AK홀딩스_20210208	32050	31650	31600	31500	31600	31600	31750
8	AK홀딩스_20210209	33300	33250	33200	33300	33250	33050	32800
9	AK홀딩스_20210210	32450	32500	32350	32400	32400	32350	32350
10	AK홀딩스_20210215	32550	32300	32550	32300	32250	32150	32200
11	AK홀딩스_20210216	31550	31500	31500	31200	31400	31400	31350
12	AK홀딩스_20210217	30750	30900	30850	30700	30700	30700	30750
13	AK홀딩스_20210218	30950	30750	30800	30600	30500	30600	30550
14	AK홀딩스_20210219	30300	30300	30300	30300	30200	30200	30200

2월 1일 ~ 7월 1일까지 수집된 데이터 13,202 개

전처리

[Scaled Data]

	A	B	C	D	E	F	G	H
1		9:00	9:01	9:02	9:03	9:04	9:05	9:06
2	AK홀딩스_20210201	0.701614	1.322716	0.494581	-0.12652	-0.74762	-1.57576	-2.19686
3	AK홀딩스_20210202	-1.41478	-1.24095	-2.11009	-1.24095	-1.76243	-1.5886	0.497344
4	AK홀딩스_20210203	0.127651	-0.95363	-1.22395	-1.49427	-1.22395	-1.22395	-1.22395
5	AK홀딩스_20210204	-0.67764	-1.02123	-1.02123	-1.02123	-1.19302	-1.53662	-1.02123
6	AK홀딩스_20210205	2.744333	-0.46854	-0.46854	1.137894	2.342723	0.736284	0.334675
7	AK홀딩스_20210208	0.815074	-1.33852	-1.60771	-2.14611	-1.60771	-1.60771	-0.80012
8	AK홀딩스_20210209	2.183657	2.014963	1.846268	2.183657	2.014963	1.340185	0.496712
9	AK홀딩스_20210210	1.149953	1.563936	0.321987	0.73597	0.73597	0.321987	0.321987
10	AK홀딩스_20210215	3.077748	0.922127	3.077748	0.922127	0.491003	-0.37125	0.059878
11	AK홀딩스_20210216	0.521182	0.145931	0.145931	-2.10558	-0.60457	-0.60457	-0.97982
12	AK홀딩스_20210217	-1.33919	-0.56987	-0.82631	-1.59563	-1.59563	-1.59563	-1.33919
13	AK홀딩스_20210218	0.873894	-0.27011	0.015889	-1.12812	-1.70012	-1.12812	-1.41412
14	AK홀딩스_20210219	1.085218	1.085218	1.085218	1.085218	0.295968	0.295968	0.295968

2월 1일 ~ 7월 1일까지 수집된 데이터 13,202 개

시계열 클러스터링

ALGORITHMS

비교 기준

1.MAE(실제값-예측값)

2.CNN과의 일치도

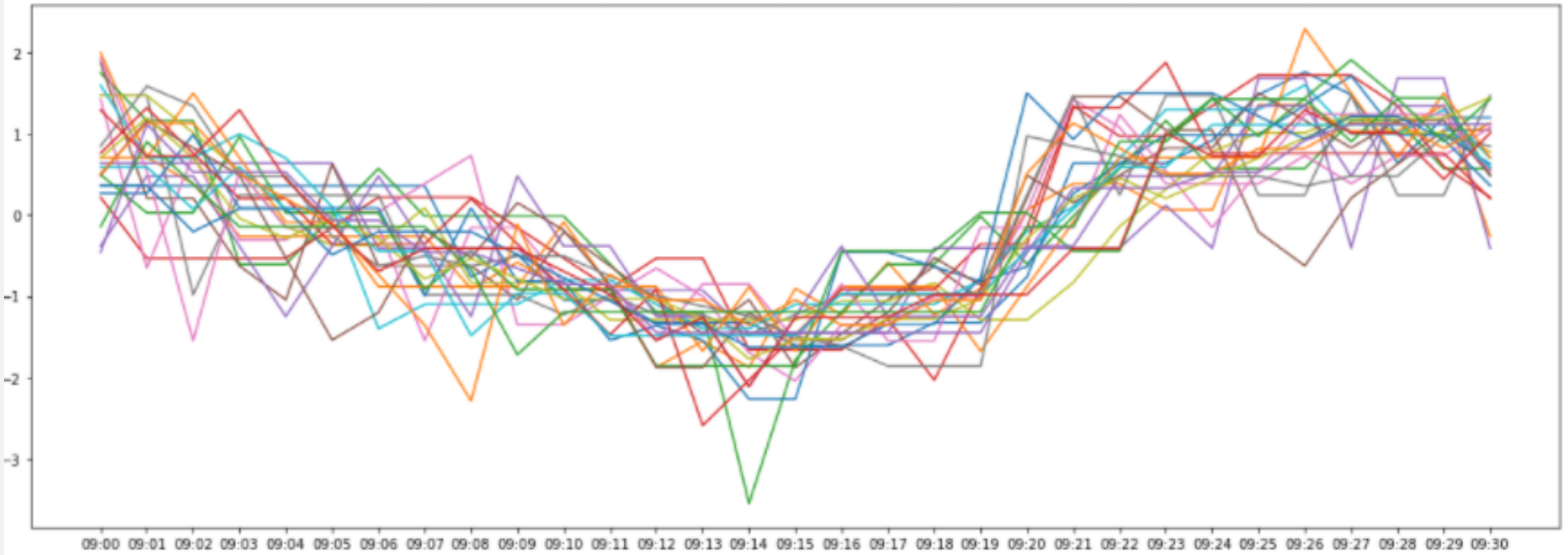
알고리즘 / 클러스터 수	GMM	Time KNN	mean shift	그 외 DTW, DBSCAN등
300	**MAE:** 0.45498 **Number:** 514	**MAE:**0.53136 **Number:** 1554	Quantile:0.01 **MAE:**0.57159 **Number:**947	1.비교할 필요없이 성능이 매우 낮은 경우
400	**MAE:** 0.428578 **Number:** 412	**MAE:**0.52393 **Number:** 1393		
500	**MAE:** 0.40147 **Number:** 296	**MAE:** 0.50799 **Number:** 1107	Quantile:0.5 **MAE:**0.61258 **Number:**725	2.알고리즘이 매우 복잡하고 데이터 수가 많아 계산에만 하루이상 소요되는 경우 비교 대상 제외

비교결과 GMM모델, 클러스터 수 400의 성능이 가장 좋음

BOAZ

시계열 클러스터링

```
cluster_0 :  25 개
['롯데케미칼_20210507' '넥센타이어_20210331' '화승엔터프라이즈_20210601' '티웨이항공_20210324'
 'NPC_20210604' '대한제당_20210601' '윌비스_20210601' '후성_20210601'
 '웅진씽크빅_20210610' '삼성중공업_20210419' '두산밥캣_20210610' '롯데케미칼_20210402'
 '삼성생명_20210322' '솔루스첨단소재_20210601' '지누스_20210601' '대영포장_20210202'
 '동양철관_20210604' '에코프로비엠_20210219' '삼성증권_20210601' '삼성전자우_20210601'
 '쌍방울_20210607' '대영포장_20210419' '한섬_20210604' '태경산업_20210610'
 'TIGER 미디어컨텐츠_20210610']
```



분류 모델

Gaussian Mixture Model + Convolutional Neural Network (보조)

문제점

GMM model의 predict만을 활용해서
새로운 데이터를 알맞는 그룹에 분류하는 정확성이 낮다.

개선 방향

GMM model외에 다른 분류 모델을 같이 활용해서,
두 모델 모두 같은 그룹을 예측하는 것만 활용하자.

모델 선정

Random Forest,CNN,RNN,LSTM과 같은 다양한 모델을 활용해
각각의 성능을 비교.(MAE, 수익률, GMM과의 일치성)

비교결과

보조 모델	MAE	GMM과의 일치 수
CNN	0.42857	412개
LSTM	0.44943	414개
그 외 (RNN,RF etc)	x>0.5	x<400

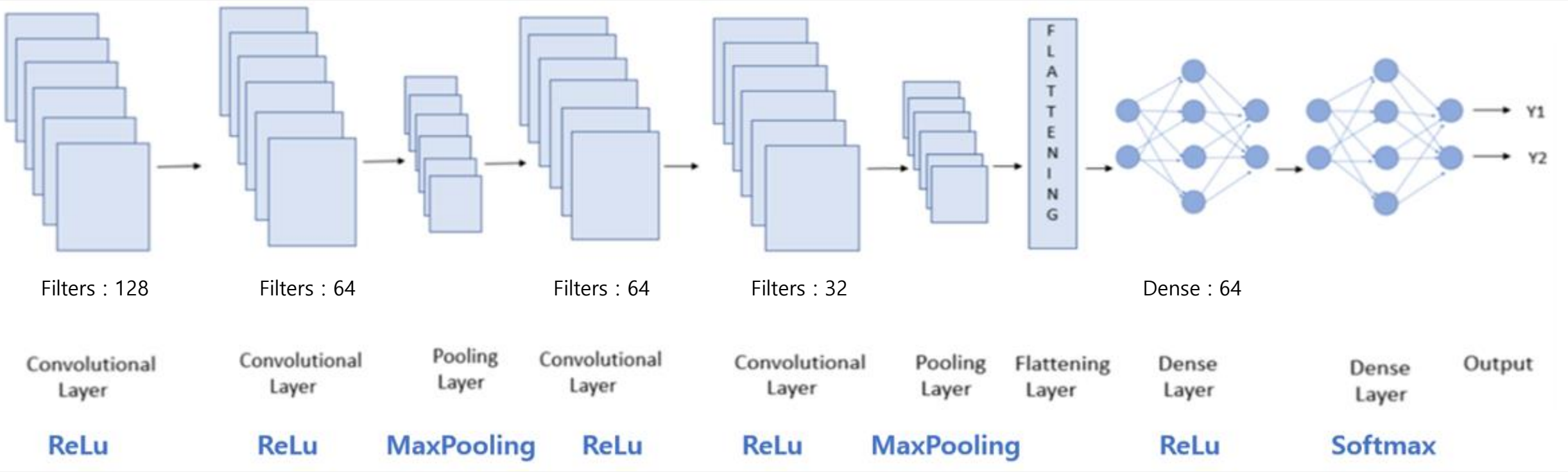
결과

CNN 모델을 활용하는 것이 GMM과의 일치성, 수익률 모두 가장 높게 나왔다.
CNN 모델을 GMM 보조로 결정.

BOAZ

분류 모델

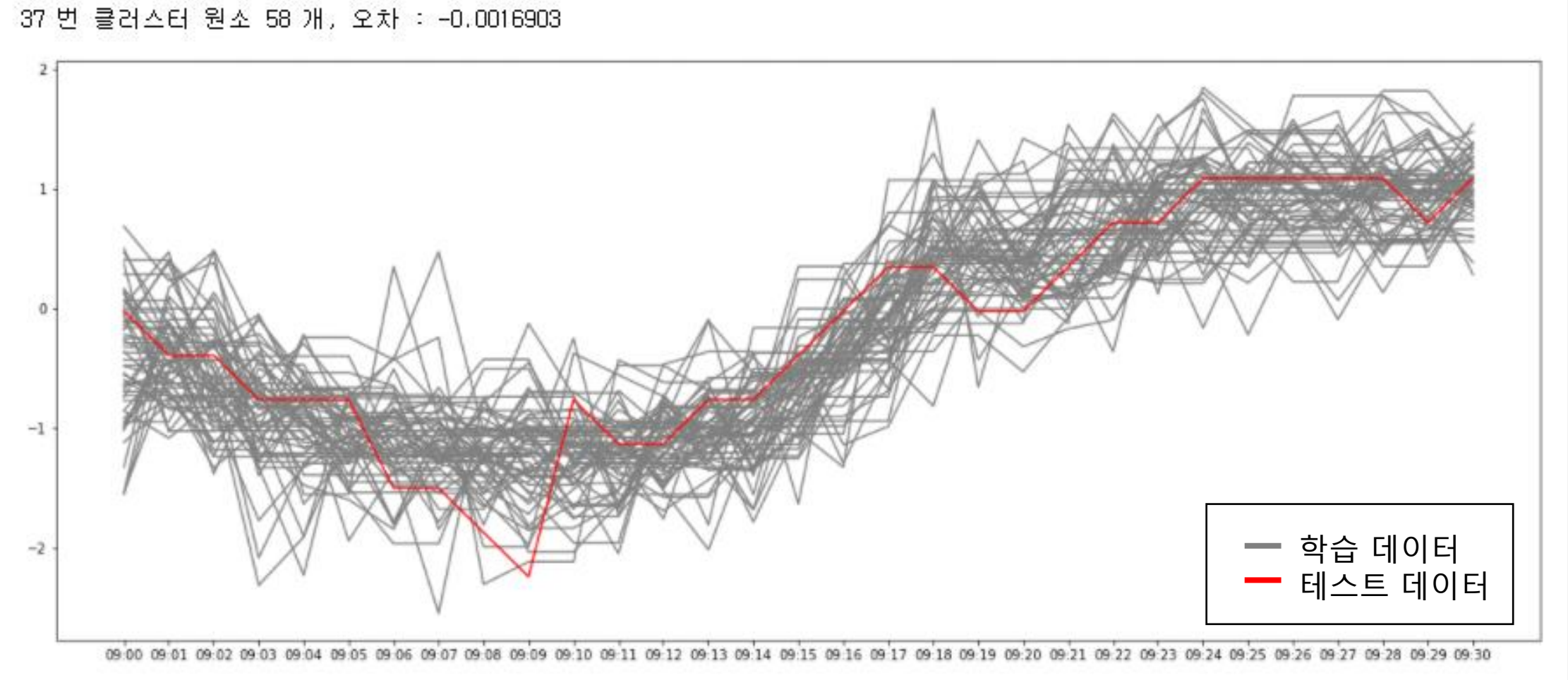
CNN layer 구조



분류 모델 활용 예시

09:06	09:07	09:08	09:09	...	09:23	09:24	09:25	09:26	09:27	09:28	09:29	09:30	clst_CNN	clst_GMM
-1.305917	-1.599748	-1.452833	-1.012086	...	0.163240	0.163240	0.457071	0.603987	0.457071	0.603987	1.338565	0.163240	49	371
-0.087409	-0.284078	0.109261	0.699268	...	-0.480747	-0.480747	-0.677416	-0.677416	-0.480747	-0.480747	-0.677416	-0.874085	92	92
0.017374	1.268266	0.330097	1.268266	...	0.017374	0.330097	-0.608073	0.330097	0.017374	0.017374	0.330097	0.017374	331	146
0.731925	1.707825	1.707825	1.707825	...	-0.243975	0.731925	-0.243975	-0.243975	0.731925	-0.243975	0.731925	0.731925	44	101
1.039606	1.301325	1.563044	1.824764	...	-1.054146	-1.054146	-0.792427	-0.792427	-1.054146	-0.792427	-1.054146	-0.530708	369	352

예측 모델


37 번 클러스터 원소 58 개, 오차 : -0.0016903
2
1
0
-1
-2
09:00 09:01 09:02 09:03 09:04 09:05 09:06 09:07 09:08 09:09 09:10 09:11 09:12 09:13 09:14 09:15 09:16 09:17 09:18 09:19 09:20 09:21 09:22 09:23 09:24 09:25 09:26 09:27 09:28 09:29 09:30
학습 데이터
테스트 데이터

예측 모델

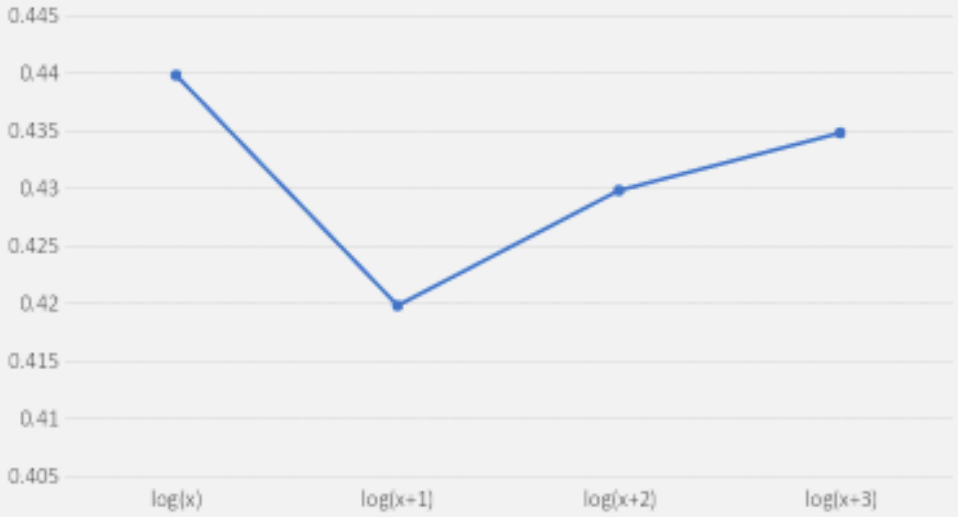
가중치 평균 함수

$$\alpha = \frac{\sum(\log(D+1)^{-1} * a)}{N}$$

D = dissimilarity(Euclidean Distance)
N = 데이터 수
a = 35분 주가
α = 35분 가중 평균값

→ 유사도가 높을수록 높은 가중치 적용

가중치 조정에 따른 MAE



→ 가중치 +1 조정

투자 종목 추천

다음 조건을 만족할 시에 투자 종목으로 추천

1. 9시부터 9:30분 사이에 거래가 25분 이상 발생

2. GMM과 CNN의 predict가 일치하는 종목

3. pred_35가 일정 값을 넘을 때

실제 조건을 만족시켜 추천해준 종목들

	pred_35	clst_CNN	clst_GMM
DL건설_20210624	0.841556	186	186
DL이앤씨_20210625	0.851929	186	186
HANARO e커머스_20210623	0.685019	378	378
NPC_20210624	0.879338	379	379
TIGER 차이나항셍테크_20210625	1.296374	258	258
WISCOM_20210624	0.804605	372	372

=

04

최종 결과

전체 프로세스 예시

1. 921개 데이터 중, GMM과 CNN의 결과가 일치한 412개의 종목 list를 반환.

```
Index(['제주항공_20210209', '아시아나항공_20210329', 'TIGER 미국테크TOP10 INDXX_20210617',
       '티와이홀딩스우_20210604', '효성_20210304', '한신기계_20210531', '에코프로비엠_20210209',
       '케이탑리츠_20210617', '디와이_20210601', '한국조선해양_20210311',
       ...
       '삼성전자_20210330', '넥센타이어_20210322', '일양약품_20210607', '혜인_20210611',
       '현대차우_20210601', '태영건설_20210607', '대영포장_20210614', '세아베스틸_20210528',
       '포스코케미칼_20210310', '삼성바이오로직스_20210225'],
      dtype='object', length=412)
```

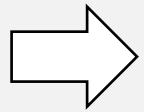
2. 412개의 종목 중에서 이윤이 예상되는 조건을 만족한 종목 11개를 추출

```
추천 종목 :  Index(['샘표식품_20210527', 'SK이노베이션_20210302', '현대위아_20210304', '현대백화점_20210428',
       '아모레퍼시픽_20210415', '동방_20210318', '에이디테크놀로지_20210419', '한신기계_20210607',
       '한국타이어앤테크놀로지_20210601', '동원금속_20210611', '제이콘텐트리_20210526'],
      dtype='object')
```

3. 투자한 11개의 종목에서 얻어낸 이윤을 계산 후 반환

   -> 11개의 종목에서 총 수익률 **2.2278** 달성

	rev_percent
현대차2우B_20210622	0.945946
DL건설_20210624	0.677966
까뮤이앤씨_20210625	0.375940
현대차우_20210624	0.267023
케이비아이동국실업_20210624	0.114811



수익률 : 2.2278

2021년 7월 19일 ~ 7월23일 모의 투자 결과

	7월 19일	7월 20일	7월 21일	7월 22일	7월 23일
추천된 종목 수	1	4	2	1	1
추천된 종목	에이엔피	KC코트렐 TIGER Fn 신재생 광전자 넥센타이어	DL 대림산업	DB금융투자	KTB투자증권
수익률	+11.1392%	+0.6876%	+0.1248%	+0.2894%	-0.9889%
코스피 등락률	-1.00%	-0.35%	-0.52%	+1.07%	+0.13%

일주일 모델 수익률 : 2.25%

일주일 코스피 수익률 : -0.69%

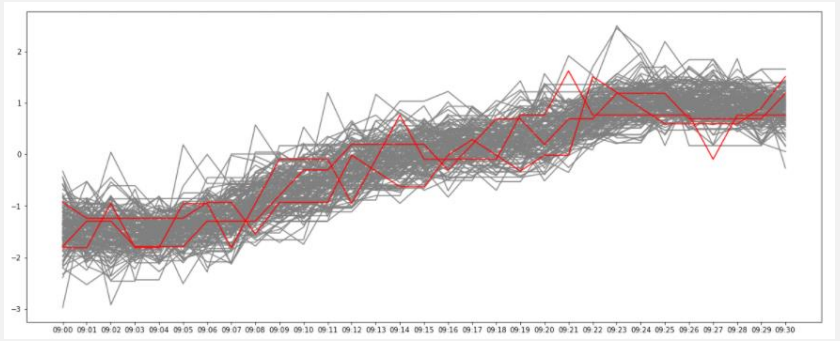
BOAZ

전체 프로세스 예시

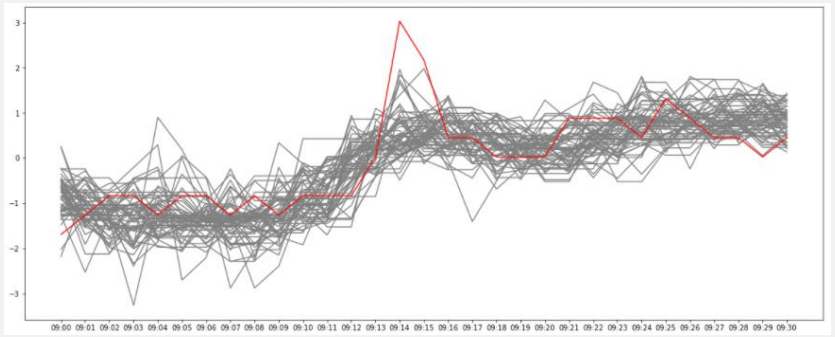
추천된 종목들이 분류된 클러스터 그래프

	clst_num
현대차2우B_20210622	2
DL건설_20210624	186
까뮤이앤씨_20210625	2
현대차우_20210624	2
케이비아이동국실업_20210624	108

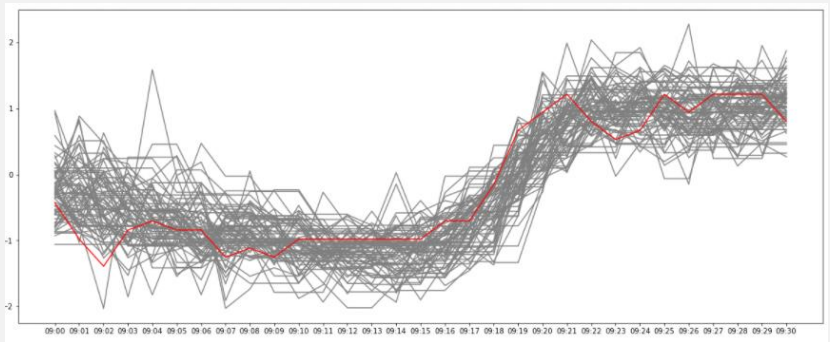
Clst=2



Clst=108



Clst=186



분류가 잘 된 것을 확인할 수 있다.

05

투자 구현

데이터 자동 크롤링

실시간성 특징 구현

AWS lambda의 trigger를 활용하여 09:30에 자동 크롤링

데이터 자동 크롤링

실시간성 특징 구현

Lambda

코드 소스 Info

에서 업로드 ▼

File Edit Find View Go Tools Window Test Deploy Changes deployed

Go to Anything (Ctrl-P)

lambda_function × Execution results ×

Environment

myaws - /

lambda_function.py

```python
import pandas as pd
import datetime
import boto3
from selenium.webdriver.chrome.options import Options
from selenium import webdriver
from io import StringIO

def get_driver():
    chrome_options = Options()
    chrome_options.add_argument('--headless')
    chrome_options.add_argument('--no-sandbox')
    chrome_options.add_argument('--disable-gpu')
    chrome_options.add_argument('--window-size=1280x1696')
    chrome_options.add_argument('--user-data-dir=/tmp/user-data')
    chrome_options.add_argument('--hide-scrollbars')
    chrome_options.add_argument('--enable-logging')
    chrome_options.add_argument('--log-level=0')
    chrome_options.add_argument('--v=99')
    chrome_options.add_argument('--single-process')
    chrome_options.add_argument('--data-path=/tmp/data-path')
    chrome_options.add_argument('--ignore-certificate-errors')
    chrome_options.add_argument('--homedir=/tmp')
    chrome_options.add_argument('--disk-cache-dir=/tmp/cache-dir')
    chrome_options.add_argument('user-agent=Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36')
    chrome_options.binary_location = "/opt/python/bin/headless-chromium"

    driver = webdriver.Chrome('/opt/python/bin/chromedriver', chrome_options=chrome_options)
    return driver

def lambda_handler(event, context):
    stock_dict = {
    '삼성전자' : '005930'

    }

    stock_codes = {k:v for v,k in stock_dict.items()}
```

88:37 Python Spaces: 4

S3

객체 (4)

객체는 Amazon S3에 저장되어 있는 기본 엔터티입니다. Amazon S3 인벤토리를 사용하여 버킷에 있는 모든 객체의 목록을 얻을 수 있습니다. 다른 사용자가 객체에 액세스할 수 있게 하려면 명시

S3 URI 복사 | URL 복사 | 다운로드 | 열기 | 삭제 | 작업 ▼ | 폴더 만들기 | 업로드

접두사로 객체 찾기

이름	유형	마지막 수정
삼성전자_0728_0728_tri.csv	csv	2021. 7. 30. pm 3:00:35 PM KST
삼성전자_0728_07282.csv	csv	2021. 7. 29. pm 6:28:59 PM KST
삼성전자_0728_07283.csv	csv	2021. 7. 29. pm 7:12:08 PM KST
삼성전자_0728_07284.csv	csv	2021. 7. 29. pm 8:04:20 PM KST

BOAZ

감사합니다