국내 Stock Selection 모델 매뉴얼

■ Intro

이 매뉴얼은 본 문서를 읽는 사람이면 누구나 국내 Stock Selection 모델을 자기의 로컬 PC에서 돌릴 수 있도록 하기 위해 작성되었습니다. 모델을 로컬 PC에서 돌리는 과정을 End-to-End로 상세하게 기술하고, 필요하다면 모델을 이해하는데 필요한 개념들에 대해서도 설명할 예정입니다. 모델을 학습하고 모니터링 리포트까지 산출하기 위해서는 아래의 5가지 절차가 필요합니다.

1. 기초 데이터 수집: 데이터가이드, QPMS로부터 기초데이터를 수집하고 업데이트합니다.
2. 데이터 가공: 수집한 기초 데이터를 전처리하고 모델 인풋으로 필요한 팩터를 연산합니다..
3. 모델 학습: 전 단계에서 계산된 팩터를 인풋으로 하여 모델을 학습합니다.
4. 포트폴리오 구성: 학습된 모델의 스코어를 산출하고 이를 기반으로 포트폴리오를 구성합니다.
5. 모니터링 리포트: 포트폴리오의 성과를 매일 모니터링할 수 있는 리포트를 작성합니다.

위 항목 중 하나라도 누락되게 되면 산출물은 최신의 정보를 반영하지 못하거나 산출 과정에서 에러가 발생할 수 있습니다. 지금부터 각 항목에 대해 더 자세하게 설명하겠습니다.

■ 기초 데이터 수집

모델 학습에 필요한 인풋은 기업 특성을 나타내는 팩터입니다. 해당 팩터를 계산하기 위해 먼저 기초 데이터 수집이 필요합니다. 기초 데이터에는 **재무데이터, 컨센서스 데이터, 가격 데이터, 섹터 데이터, 테마 데이터, 수급데이터**가 있습니다. 또한 동 모델은 코스피 조정 샤프지수를 Target 데이터로 하여 학습하게 되는데 이때 필요한 코스피 데이터, 그리고 이후 포트폴리오 구성 및 평가 시에 필요한 벤치마크 정보를 포함하고 있는 **벤치마크 데이터**가 필요합니다. 아래는 기초데이터 수집 과정에 대한 설명입니다.

1. 재무데이터 수집

   파일명: fin_df.pkl 과 fin_df_20240325.xlsx

   1) fin_df_20240325.xlsx 파일을 엽니다.

   2) DataGuide 엑셀 애드인 기능을 활성화하고, Refresh 버튼을 누릅니다.

   3) 저장을 하고 파일을 닫습니다.

   Fin_df.pkl은 1999.01.01~2023.12.31까지 재무데이터들이 담긴 피클 파일입니다. Fin_df_20240325와 동일한 재무데이터 항목에 대해 과거 기간을 조회한 파일이며, 대용량이기 때문에 코드에서 불러오는 시간을 단축시키기 위해 엑셀 파일이 아닌 pickle파일로 저장하였습니다. 이후 데이터 가공 파트에서 fin_df.pkl과 최신 날짜까지 업데이트한 fin_df_20240325.xlsx 파일을 이어 붙이는 전처리가 진행되고 최종적으로 '99년도부터 최근까지의 재무데이터 데이터프레임이 완성되게 됩니다.

2. 수급 데이터 및 거래 데이터

   파일명: daily_df_float.parquet과 daily_df_float_20240325.xlsx

   1) daily_df_float_20240325.xlsx 파일을 엽니다.

   2) DataGuide 엑셀 애드인 기능을 활성화하고, Refresh 버튼을 누릅니다.

   3) 저장을 하고 파일을 닫습니다.

   재무 데이터와 마찬가지로 '24년 이전의 과거 시계열에 해당하는 데이터는 코드 업로드를 빠르게 하기 위해 parquet파일로 압축한 daily_df_float.parquet 에 저장되어있고, daily_df_float_20240325.xlsx에서 업데이트 하여 전처리 코드에서 이 둘을 이어붙이는 방식으로 수급 데이터 및 거래 데이터의 데이터프레임을 생성합니다.

3. 컨센서스 데이터

   파일명: consen_dat.csv, consen_dat_update.xlsx

   1) consen_dat_update.xlsx 파일을 엽니다.

   2) G열의 날짜를 직전 날짜로 변경합니다.

   3) DataGuide 엑셀 애드인 기능을 활성화하고, Refresh 버튼을 누릅니다.

   4) consen_dat.csv 파일을 엽니다.

   5) consen_dat_update.xlsx의 G열을 복사해서 consen_dat.csv 마지막 열에 붙여넣습니다.

   6) consen_at.csv 저장하고 파일을 닫습니다.

   다른 데이터와는 다르게 컨센 데이터의 경우 전처리 코드에서 과거의 데이터와 이어 붙이는 작업이 없기 때문에 직접 consen_dat_update.xlsx에서 업데이트한 데이터를 복사해서 consen_dat.csv에 붙여넣기 하는 작업이 필요합니다.

4. 가격 데이터

   파일명: price_dat.csv

   1) DB관리프로그램을 엽니다.

   2) Market > MB에서 가격 데이터 추출해서 csv파일로 받습니다.

5. 테마 데이터

   파일명: thm_ETF_수정.xlsx

   1) 파일을 열고 최신 정보 반영 후 저장하고 닫습니다.

6. 섹터 데이터

   파일명: sector_dat.xlsx

   1) Sector_dat.xlsx 파일을 엽니다.

2) DataGuide 엑셀 애드인 기능을 활성화하고, Refresh 버튼을 누릅니다.

3) 저장을 하고 파일을 닫습니다.

7. 벤치마크 데이터

파일명: bm_dat.xlsx

1) bm_dat.xlsx 파일을 엽니다.

2) 최근 날짜까지 벤치마크 데이터를 업데이트합니다.

3) 저장을 누르고 파일을 닫습니다.

■ 데이터 가공

데이터 가공은 다음과 같은 두 py 파일에 의해 진행됩니다.

1. equity_data.py: 수집한 기초데이터를 바탕으로 팩터 연산에 필요한 컴포넌트를 생성합니다.

2. generate_factor.py: equity_data.py에서 계산된 컴포넌트와 기초데이터를 바탕으로 팩터를 계산합니다.

equity_data.py 파일에서는 주로 generate_factor.py에서 팩터 연산을 용이하게 하기 위해 연산 함수를 정의하거나, 기초 데이터 전처리, 또는 기초 데이터에는 없는 데이터 (예: ETF PDF정보)를 수집하고 가공합니다. 주된 전처리로는 한글로 기재된 데이터명을 영문으로 바꾸거나, 미래 참조 오류를 방지하기 위해 재무데이터 이용 가능 시점에 lag를 부여하는 함수, 데이터 항목별로 시계열 데이터프레임 생성 등이 있습니다.

모델에 인풋으로 들어가는 팩터는 generate_factor.py에서 계산됩니다. 재무 팩터와 컨센서스 팩터는 equity_data.py에서 생성된 재무 fin_dict.pickle과 daily_dict.pickle 파일의 컴포넌트들을 기반으로 계산됩니다. 추가로 가격데이터와 수익률 데이터를 통해 기술적 팩터들을 만들어냅니다. 최종적으로 종목별로 모든 팩터들을 나타내는 최종산출물은 FACTORS_FINAL.feather 파일입니다.

(실행방법)

팩터를 생성하기 위해서는

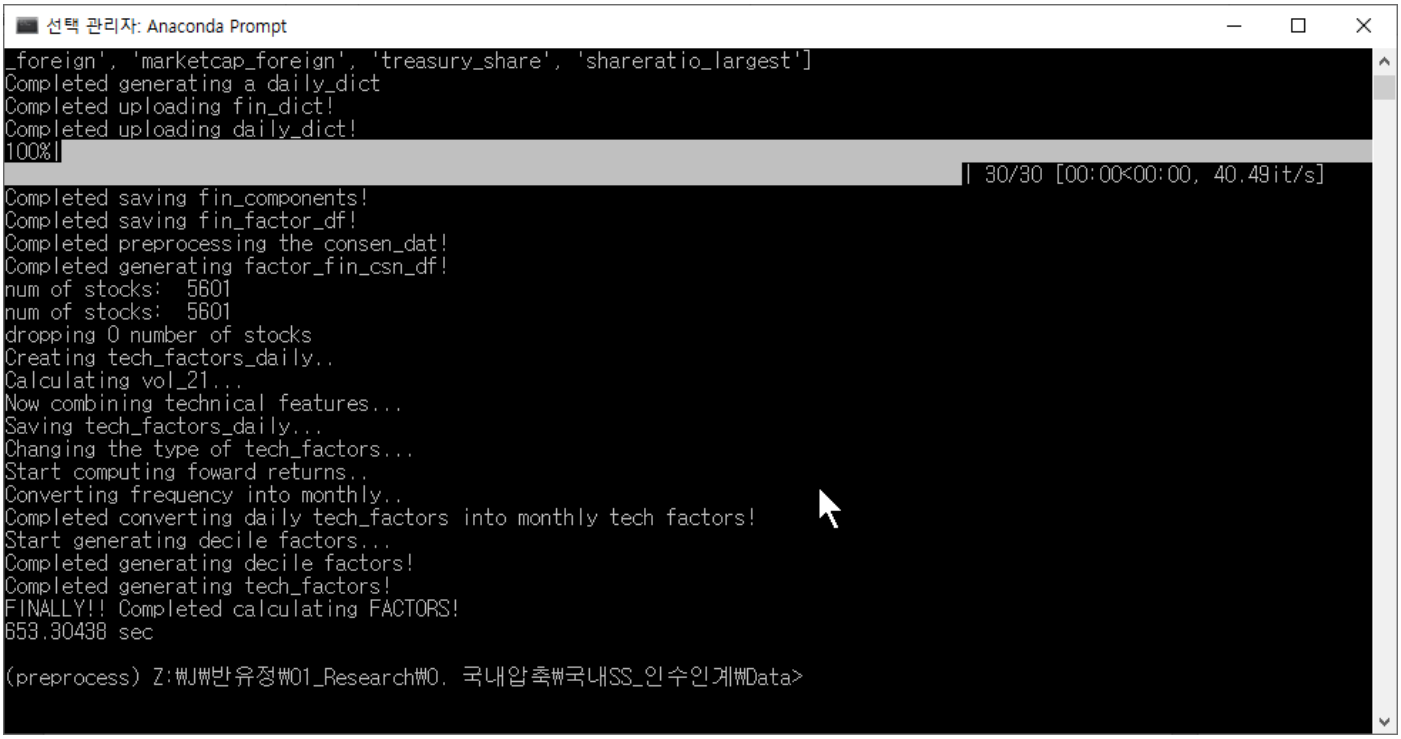
1) 터미널을 켭니다.

2) Data폴더 경로로 이동합니다.

3) python generate_factor.py를 실행합니다.

4) Data/ processed_data 폴더가 생성되고 가공된 데이터들의 파일이 잘 생성되었는지 확인합니다.

터미널에서 파일 실행시켰을 때 찍히는 로그는 다음과 같습니다.



■ 모델 학습

모델 학습은 Model/train.py에 의해 진행됩니다.

이전에 생성된 FACTORS_FINAL.feather 파일의 종목별 팩터를 인풋으로 받는 LightGBM 인스턴스를 생성한 뒤, 생성된 모델 파라미터 조합별로 학습을 시킵니다. 학습된 파라미터 조합별 모델 성능을 높은 순서대로 정렬시키고 이 중 5개를 뽑고, 모델별 스코어의 평균치를 모델 MSE 평균으로 나눈 값을 스코어로 하여 종목별 모델 스코어를 뽑아냅니다. 산출물은 Model/score/ind_stock_score_20240531.xlsx가 됩니다.

(실행방법)

모델을 학습시키기 위해서는

1) train.py를 열고 if __name__ == "__main__": 아래 부분에 start_date와 end_date를 원하는 날짜 입력합니다. 2024-05-31 기준으로 뽑고싶으면 start_date와 end_date 모두 "2024-05-31"로 입력하면 됩니다. 입력 후 스크립트 파일 저장합니다.

2) 터미널을 켭니다.

3) Model폴더 경로로 이동합니다.

4) train.py를 실행합니다.

5) Model/results/score 폴더가 생성되고 해당 월 모델 스코어가 잘 생성되었는지 확인합니다.

터미널에서 파일을 실행시켰을 때 찍히는 로그는 아래와 같습니다. Train/Test기간별로 CV 데이터셋이 만들어지고 각 파라미터 조합(cv_params)에 대해 학습이 이루어지는 것을 확인할 수 있습니다.

```
(preprocess) Z:\J\반유정\01_Research\0. 국내압축\국내SS_인수인계\Model>python train.py
  0%|                                                                                         | 0/1 [00:00<?, ?it/s]G
enerating 20240531
ticker  date
000010  2022-05-09    4331.5
        2022-05-10    4330.5
        2022-05-11    4329.5
        2022-05-12    4330.5
        2022-05-13    4333.5
                       ...
950220  2024-05-09     896.0
        2024-05-10     829.0
        2024-05-13     934.0
        2024-05-14     634.0
        2024-05-16    1030.0
Name: TRAN_AMT, Length: 2789298, dtype: float64
number of cv params :  [(0.5, 0.5, 250), (0.5, 0.5, 500), (0.5, 0.75, 250), (0.5, 0.75, 500), (0.75, 0.5, 250), (0.75, 0
.5, 500), (0.75, 0.75, 250), (0.75, 0.75, 500)]
number of test params :  5
Lookahead:  0 | Train:  60 | Test:  1 | Params:   8
Appending data for split 0, y_pred keys: ['25', '50', '75', '100', '125', '150', '175', '200', '225', '250', '275', '300
', '325', '350', '375', '400', '425', '450', '475', '500']
```

■ 포트폴리오 구성

모델 학습이 다 되면, 생성된 종목별 모델 스코어로 포트폴리오를 구성합니다. 포트폴리오 구성 방식은 다음과 같습니다.

1) 인덱스를 구성하는 종목들의 스코어 평균으로 인덱스 스코어를 산출합니다.

2) 인덱스 스코어가 가장 높은 인덱스 5개를 선택합니다.

3) 선택한 인덱스 내에서 모델 스코어가 가장 높은 종목 2개를 선택하는 경우 Model A, 선택한 인덱스 내에서 모델 스코어가 가장 높은 종목 1개, 시총 제일 큰 종목 1개를 총 2개를 선택하는 경우 Model B입니다.

추가로, 포트폴리오에 담을 수 있는 종목에 대한 조건을 부여하게 되는데, 포트폴리오 구성 시점에 시총 순위 1000위 이내, 거래대금 순위 500위 이내이면서 최근 2개년도 연속 적자 기업이 아닌 기업으로만 위의 로직으로 포트폴리오를 구성합니다.

(실행 방법)

1) 원하는 모델(A 또는 B)의 port_Model_{모델명}.py을 열고 self.date에 원하는 날짜를 입력합니다.

2) 파일을 저장합니다.

3) 터미널을 켜고 Portfolio/port_Model_{모델명}.py를 실행합니다.

4) Portfolio/port 폴더가 생성되었는지 확인하고 해당 폴더 안에 포트폴리오 구성정보가 담긴 산출물들이 잘 나왔는지 확인합니다.

터미널을 실행하였을 때 찍히는 로그는 다음과 같습니다.

```
(preprocess) Y:\J\반유정\01_Research\0. 국내압축\국내SS_인수인계\Portfolio>python port_Model_A.py
Connected QPMS DB successfully
Found pregenerated file pdf_df.feather
Start selecting index.....
20240430 is the first date..
Completed Saving selected_sec_df!
Completed saving selected index universe score data!
Let's start picking the stocks !
['AI반도체', 'IT_FICS', '방산', '원자력', '의료기기']
Completed picking stocks 20240430
186.20635 sec
```

■ 모니터링 리포트

마지막으로, 모델 결과물을 바탕으로 모니터링 리포트를 생성합니다. 국내 모델의 모니터링 파일을 작성하는 데 필요한 테이블을 생성하고 자동으로 엑셀파일을 업데이트 하는 작업입니다.

(실행방법)

1) portfolio-analysis/data 폴더에 bm_df.xlsx와 idv_stocks_price_df.xlsx 파일을 각각 엽니다.

2) DataGuide 엑셀 애드인 기능을 활성화하고, Refresh 버튼을 누릅니다.

3) 저장하고 파일을 닫습니다.

4) GenerateReport.ipynb 주피터 노트북을 열고 각 셀을 실행합니다.

5) domestic_port_monitoring.xlsx 파일이 잘 업데이트 되었는지 확인 후, pdf내보내기를 통해 리포트를 생성합니다.

Domestic_port_monitoring.xlsx 파일의 '포트폴리오 수익률', '종목별 수익률', '누적수익률', '종목별 수익률', '종목 기여도', '인덱스 기여도'에 들어가는 값들을 코드로 계산해서 xlwings 라이브러리를 이용해 엑셀을 열고, 시트 업데이트 후 파일을 닫는 과정을 자동화하였습니다.

생성된 리포트의 내용은 다음과 같습니다.

국내 모델 모니터링
2024-05-30

■ 포트폴리오 수익률

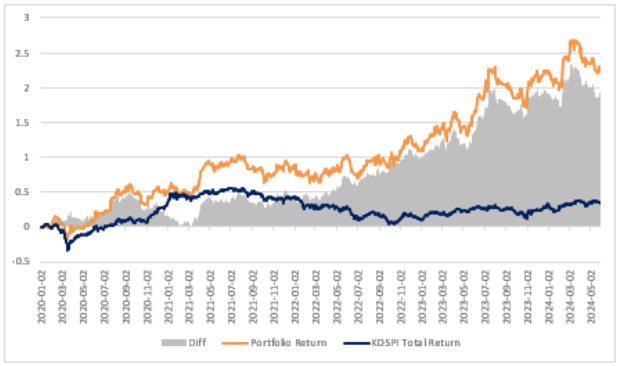
구분	1일	1주	1개월	3개월	MTD	YTD
포트폴리오	-0.96%	1.29%	-2.57%	-5.19%	-2.48%	0.69%
KOSPI	-1.67%	-1.69%	-0.37%	1.84%	-0.54%	1.56%

■ 포트폴리오 종목 분석

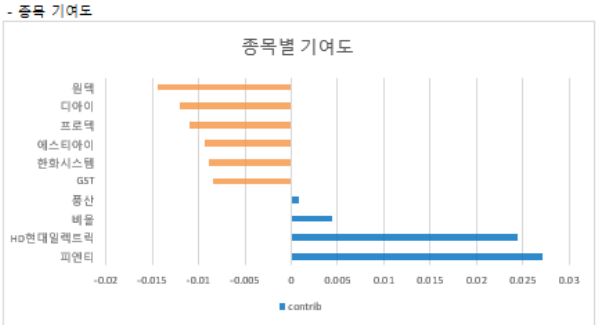
- 종목별 수익률

구분	인덱스명	종목명	비중	1일	1주	MTD	포지션
테마	AI반도체	에스티아이	9.21%	0.00%	0.00%	-10.21%	손절
테마	AI반도체	프로텍	9.01%	0.00%	-4.06%	-12.16%	손절
섹터	IT_FICS	디아이	8.87%	0.00%	0.00%	-13.50%	손절
섹터	IT_FICS	GST	9.33%	0.00%	0.00%	-8.98%	손절
테마	방산	한화시스템	9.26%	-3.47%	-0.28%	-9.65%	보유
테마	방산	풍산	10.33%	0.15%	-5.21%	0.77%	보유
테마	신재생에너지	피엔티	12.48%	0.00%	4.59%	21.74%	익절
테마	신재생에너지	HD현대일렉트릭	12.29%	-2.15%	17.23%	19.84%	보유
테마	의료기기	비올	10.68%	-3.49%	-0.94%	4.17%	보유
테마	의료기기	원텍	8.53%	0.00%	0.00%	-16.83%	손절

■ 포트폴리오 누적수익률 차트



- 종목 기여도



- 인덱스 기여도