[High Frequency Finance]

**Final Project**

**VKOSPI 급등예측모형**

**20203884 임채빈**

1. 프로젝트 배경

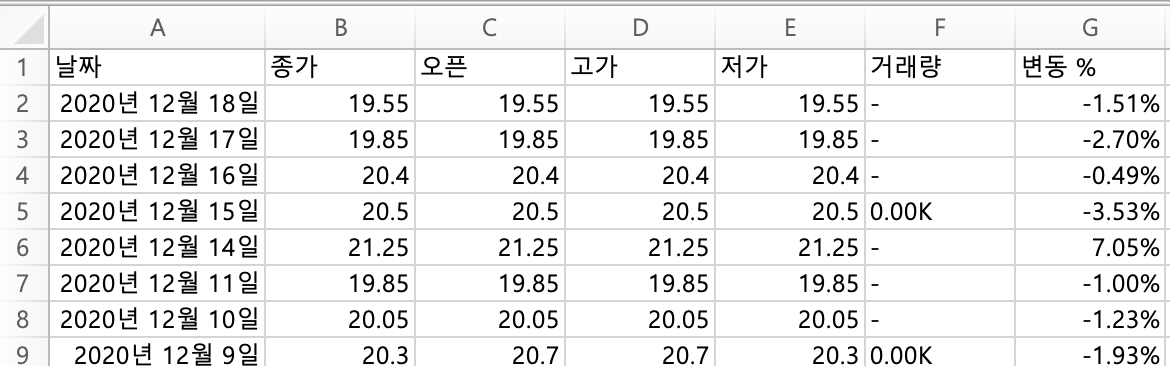
KAIST MFE(금융공학석사) 과정을 이수하면서, 계속해서 듣던 말이 있다. **“수익률 예측은 어렵고, 그에 비해 변동성 예측은 충분히 가능**하다.”라는 것이다. 이 말을 듣고 항상 들었던 생각이 ‘변동성이 충분히 예측 가능하다면, 이걸로 돈을 벌 수 있지 않을까?’라는 생각이 들었다. 변동성을 정확히 예측 가능하다면, 옵션의 hedging을 완벽하게 할 수 있을 것이다. 뿐만 아니라, 급등하는 변동성을 예측할 수 있다면, 옵션의 양매수 포지션 등으로 이익을 취할 것이라 기대할 수 있다.

이 프로젝트에서는 옵션을 다루고 싶었으나 옵션의 데이터를 하나 하나 정제하는 데에는 유동성 문제가 크게 발생하였다. 한국 옵션은 유동성이 ATM에 가까운 옵션, 그리고 만기가 얼마 남지 않은(1달 이내) 옵션만 유동성이 있고 나머지 대부분은 유동성이 떨어진다. 심지어는, 위의 경우에서도 이틀동안 거래가 되지 않는 등의 유동성 이슈가 존재하기도 한다.

따라서, 다른 관점으로 접근해 보았다. 결과적으로, VKOSPI 선물을 이용하였다. VKOSPI란, KOSPI200 옵션가격을 이용하여 옵션 투자자들이 예상하는 KOSPI200 지수의 미래변동성을 측정한 지수이다. 이는 미국 시카고옵션거래소(CBOE)가 S&P500 지수 옵션을 기초로 하여 산출 및 발표하는 변동성지수(VIX)와 유사한 한국형 변동성 지수이다. 이를 기초자산으로 한 선물을 이용하여 투자를 하면, 변동성에 대해 투자하는 것이 될 것이고, 변동성을 예측할 수 있다면 수익을 창출할 수 있을 것이라 기대할 수 있다. 즉, 일반적인 주식이나 주가지수를 예측하는 것은 시계열적 특성, Brownian motion과 같은 random walk process 때문에 힘들겠지만, 그에 비해 변동성이 예측 가능하다면 VKOSPI 선물과 같은 상품들의 흐름은 예측 가능할 것이라 생각했다.

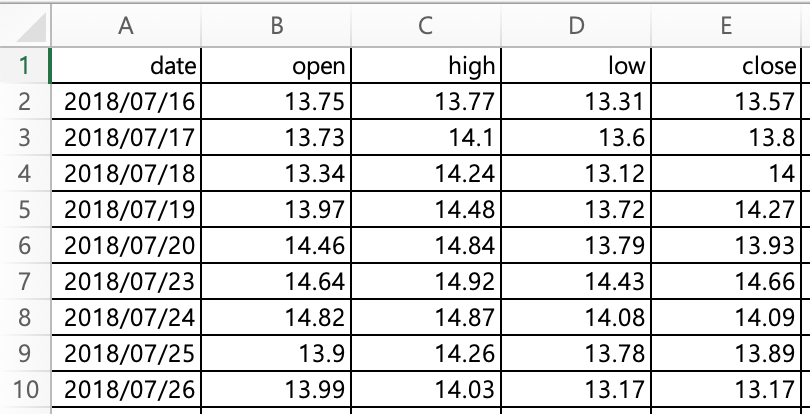
하지만, 변동성은 높은 시기와 낮은 시기가 극명하게 분리되어 나타나는 특성이 있다. 따라서 단순히 변동성을 측정하는 것 보다는 변동성, 즉 VKOSPI의 급등을 예측하는 것이 더 안전하다는 생각이 들었다. 조금 더 보수적으로 투자하되, 확실한 수익을 가져가는 전략을 구사하고 싶었다. 결국, **VKOSPI 선물 가격의 급등을 예측하고, 모델이 VKOSPI 선물 가격이 급등한다고 판단을 내리면, 그에 투자하는 전략**을 구사하였다. 여기서 선물 가격의 급등(변동성의 급등)이란, 5%이상 상승하는 경우로 정의하였다.

1. 프로젝트 사용 데이터
   1. VKOSPI 선물 일별 데이터



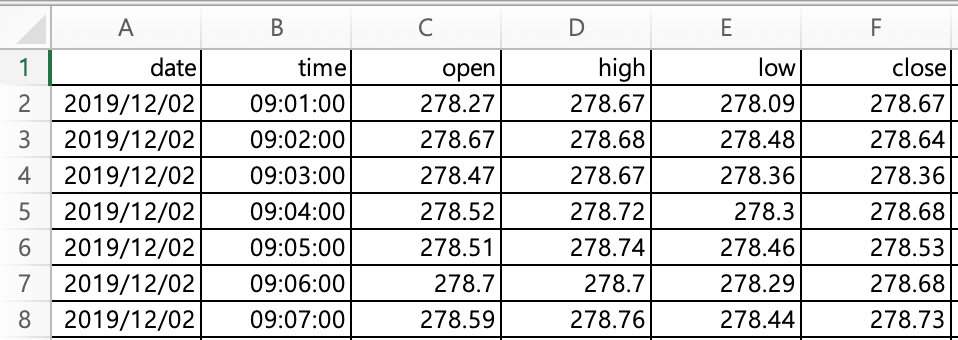
이는 우리의 종속변수(Y)를 포함하고 있는 데이터이다. 이 데이터의 ‘종가’ 부분의 수익률이 5%이상 올랐는지 여부를 종속변수(Y)로 사용하였다.

* 1. VKOSPI 지수 일별 데이터



이는 우리가 항상 포함하는 feature인 VKOSPI 주가지수 데이터이다. 이 중에서 ‘close’의 수익률을 factor로 사용하였다.

* 1. KOSPI200 분봉 데이터



분봉데이터를 이용해 하루 동안의 수익률 표준편차를 구하거나, 의 식을 사용하여 Intraday volatility를 구하였다. (여기서 r은 수익률)

1. 모델 설명
   1. 데이터 전처리

* 데이터 기간 : 2019-12-02 ~ 2020-08-07
* KOSPI200 분봉데이터의 경우 일별 RV를 구하여 일별 데이터로 변환하였음.
  1. Decision Tree

Classification model로는 Decision tree를 사용하였음. 그 이유로는, Decision tree의 분류가 잘 일어나는지 눈으로 관찰하고, 해석할 수 있기 때문.

* 1. Features(독립변수)

1. VKOSPI INDEX 수익률 : 모든 모델에 공통으로 포함되는 1번째 feature

VKOSPI INDEX ‘close’의 변화율(수익률)을 1번째 feature로 사용하였다.

1. 하루 동안의 수익률 표준편차 or

여기서 RV는 1분, 2분, 3분, 5분, 7분, 10분, 15분, 20분 간격의 frequency를 이용하여 RV를 구하였음. 그리고, 각 frequency를 사용하여 모델의 2번째 feature로 활용하였음.

즉, 1번째 feature로는 VKOSPI INDEX, 2번째 feature로는 N분(N = 1, 2, 3, 5, 7, 10, 15, 20) frequency의 RV를 쓰거나, 하루 동안의 수익률 표준편차를 사용하였음.

이를 이렇게 사용한 이유는, VKOSPI INDEX는 선물 가격의 전체적인 흐름과 함께 움직인다고 생각하였고 세부적인 흐름은 Intraday동안 발생한 변동성이 결정할 수 있다고 생각하고 이렇게 독립변수를 정의하였음.

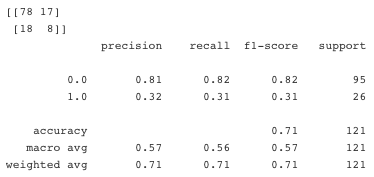
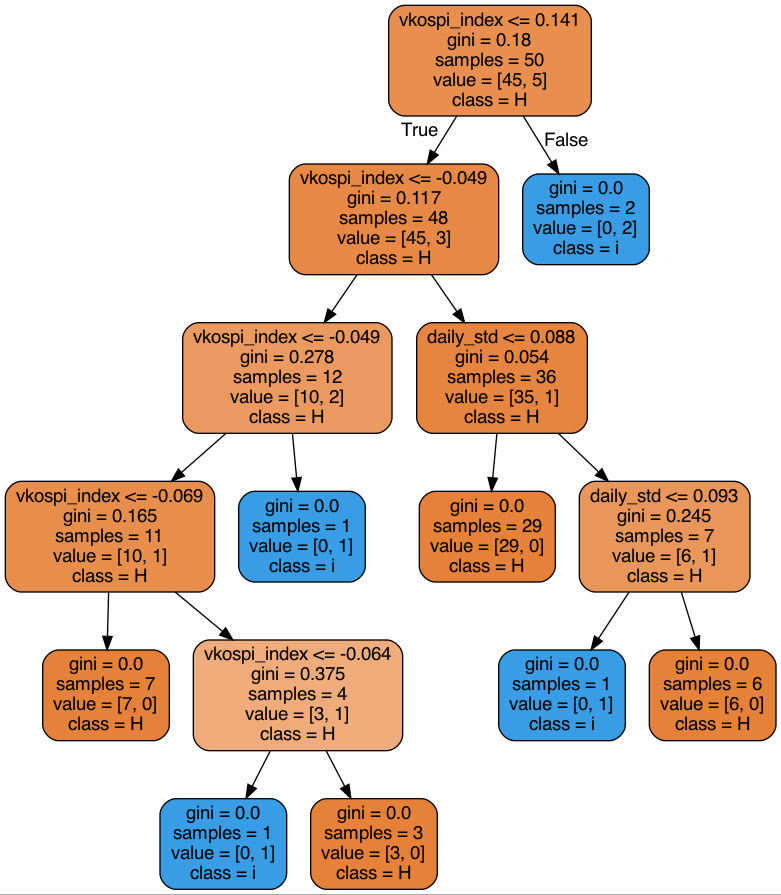
* 1. Label(종속변수)

VKOSPI 선물 수익률이 5%이상 올랐으면 True, 올라가지 않았으면 False.

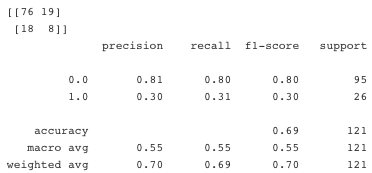
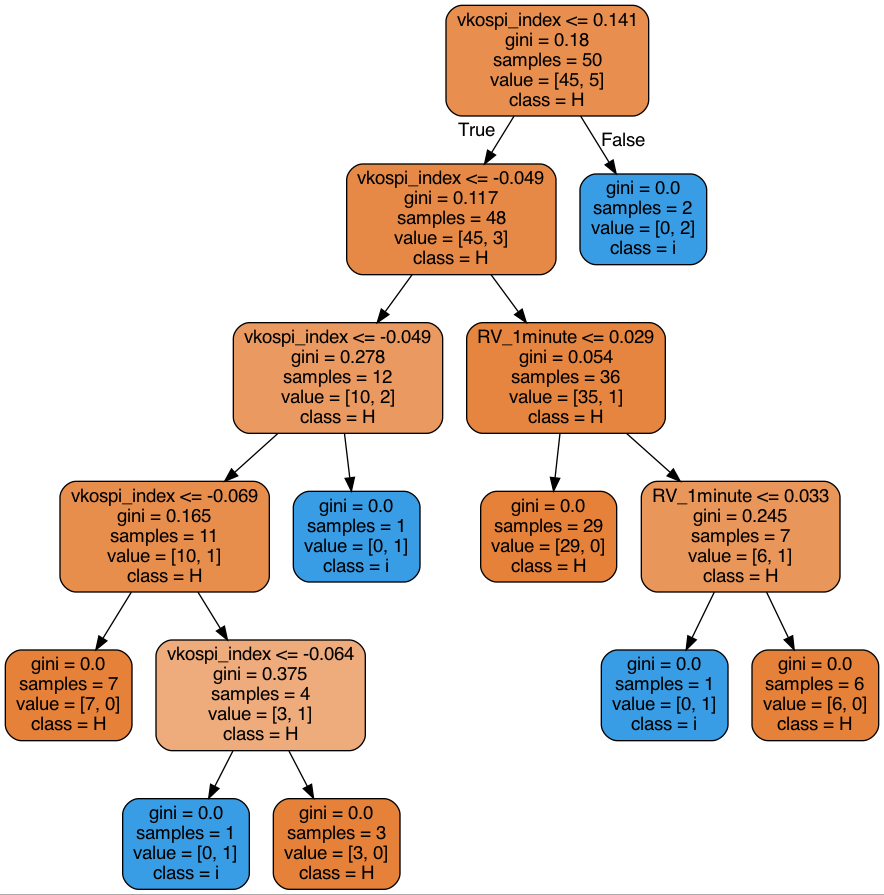
1. Classification 결과
   1. 모델 분류 성능

모델을 정의할 때에는 독립변수 기준으로 정의하고, 결과를 나열하겠다. 결과에는 confusion matrix와 classification report가 포함된다. 좌측 표의 좌측 상단에는 confusion matrix가 배치되고, 우측 하단에는 classification report가 배치된다. 또한, 우측 그림에는 Decision tree가 분류되는 모습을 배치하였다.

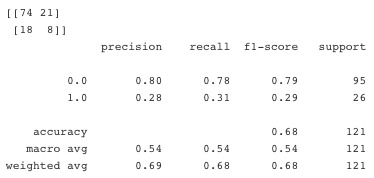
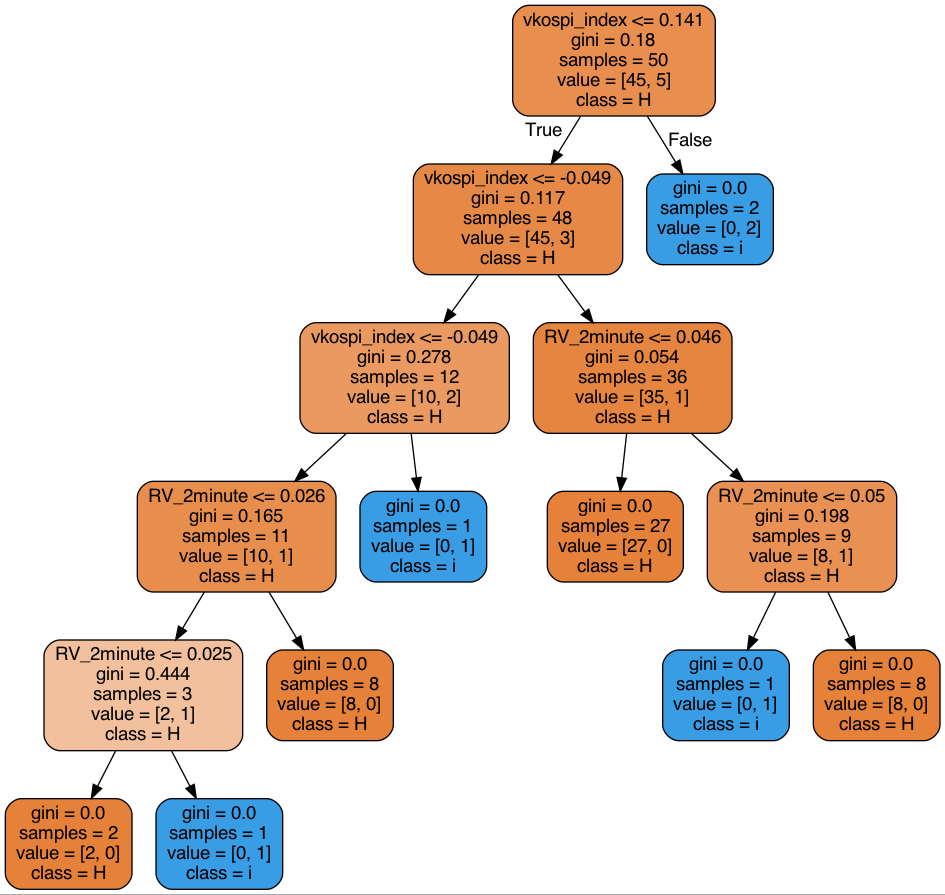
* + 1. VKOSPI INDEX + Intraday 표준편차

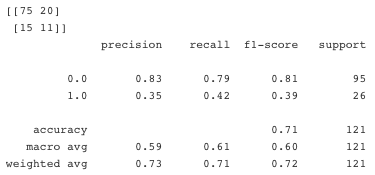
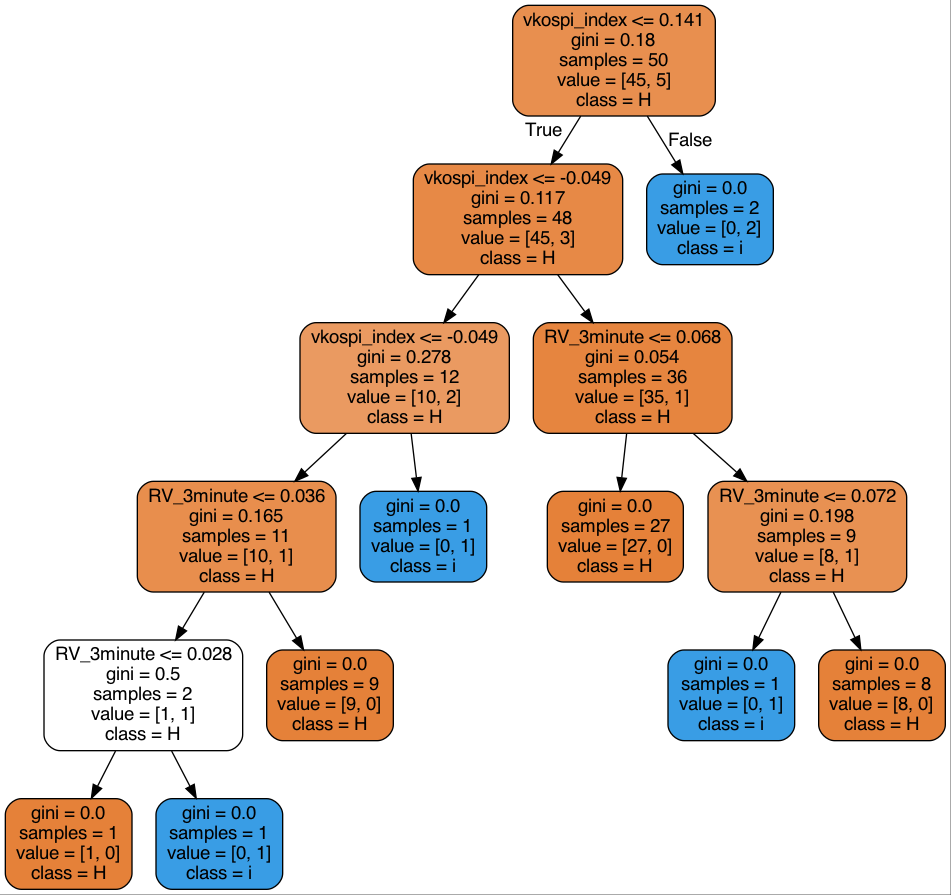
* + 1. VKOSPI INDEX + RV(frequency : 1분)

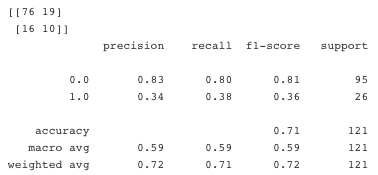
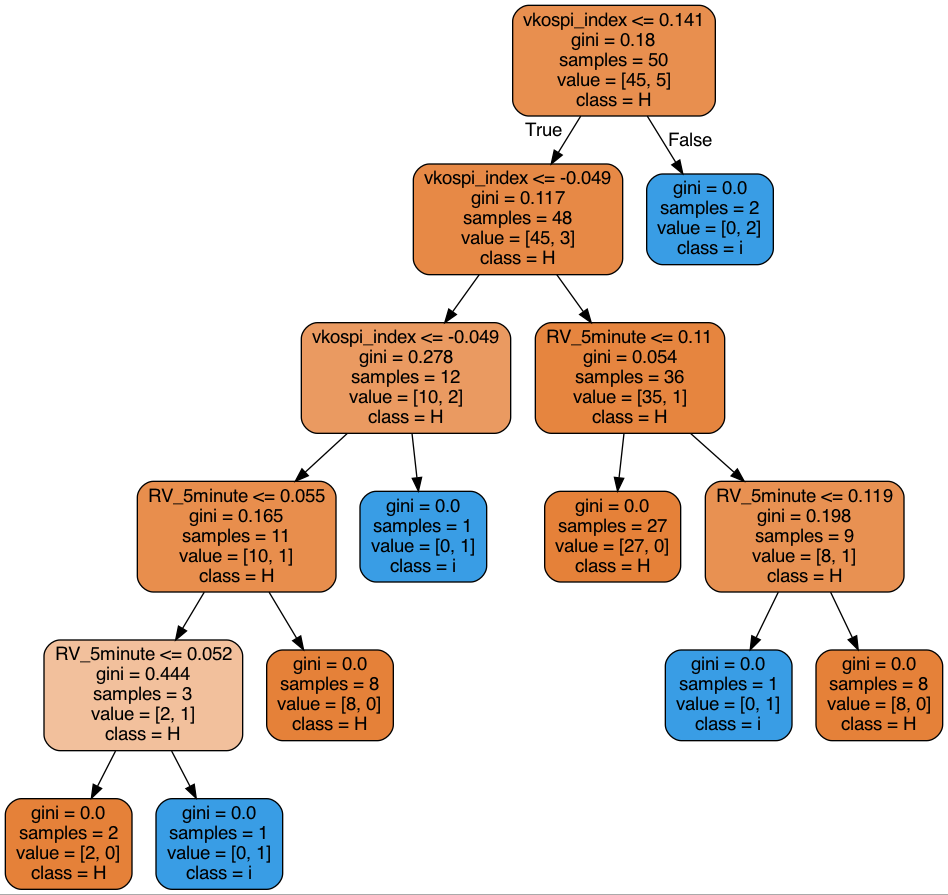
* + 1. VKOSPI INDEX + RV(frequency : 2분)

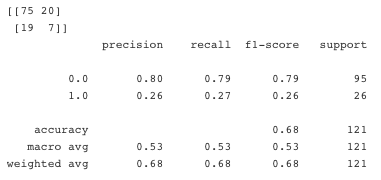
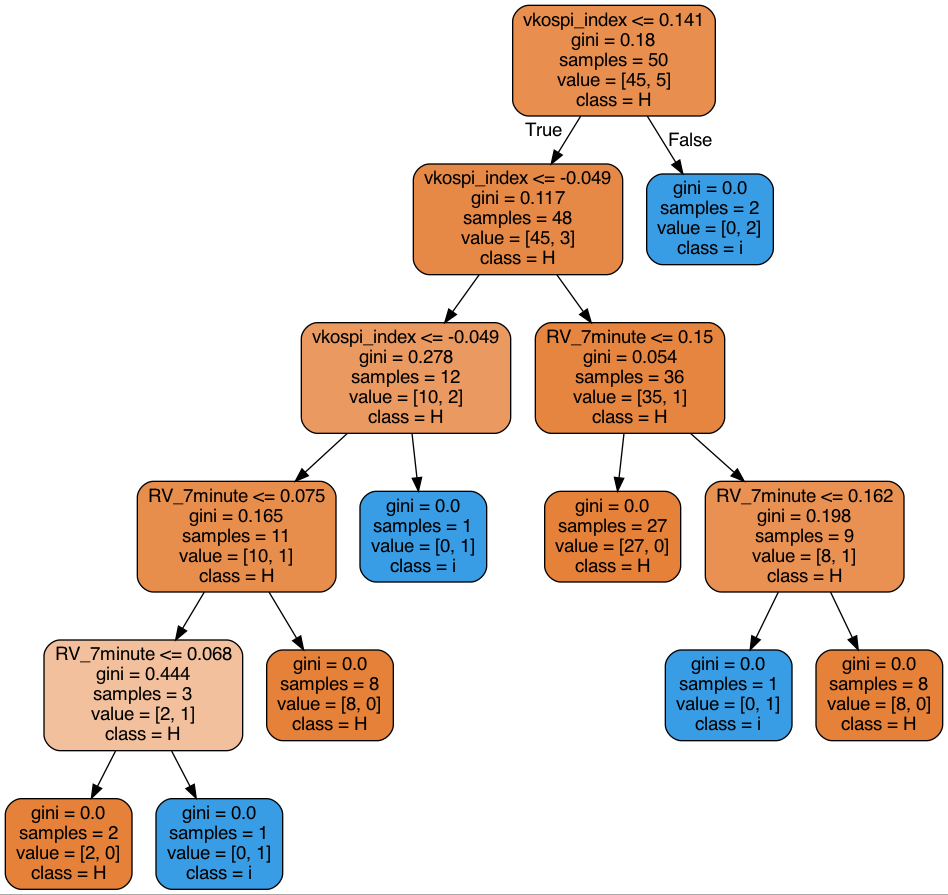
* + 1. VKOSPI INDEX + RV(frequency : 3분)

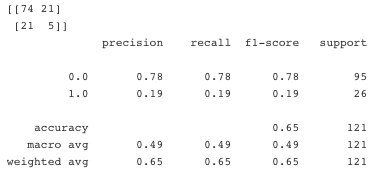
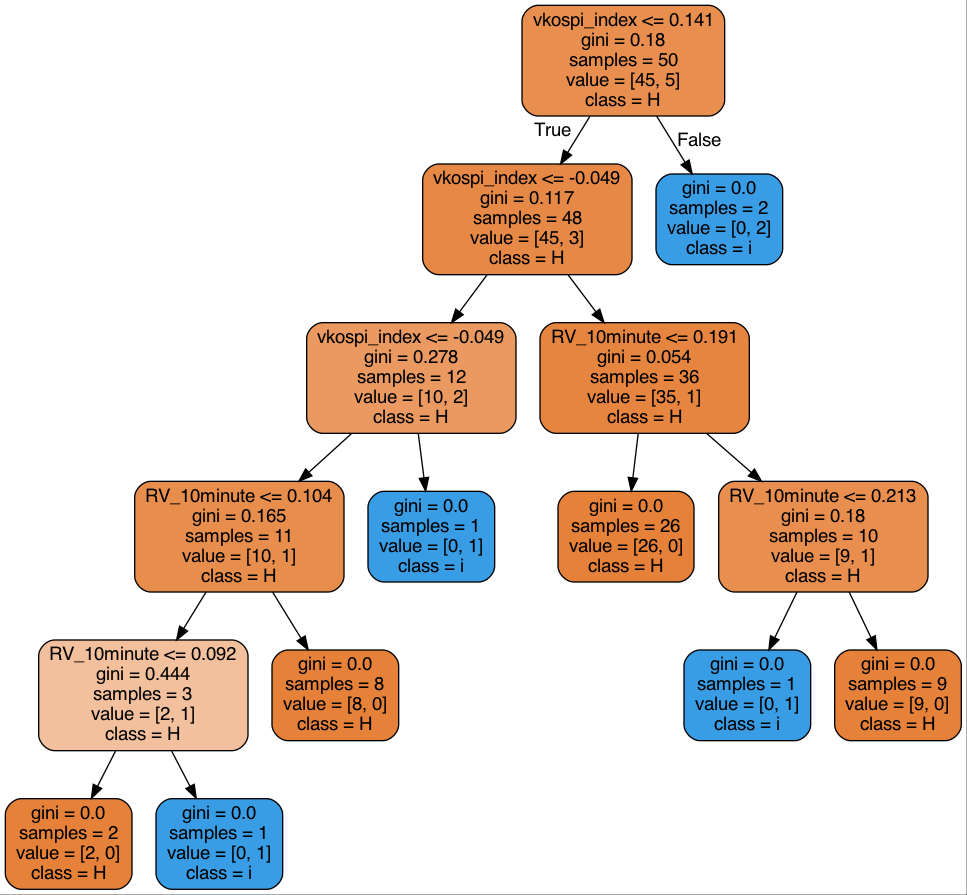
* + 1. VKOSPI INDEX + RV(frequency : 5분)

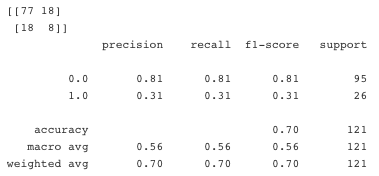
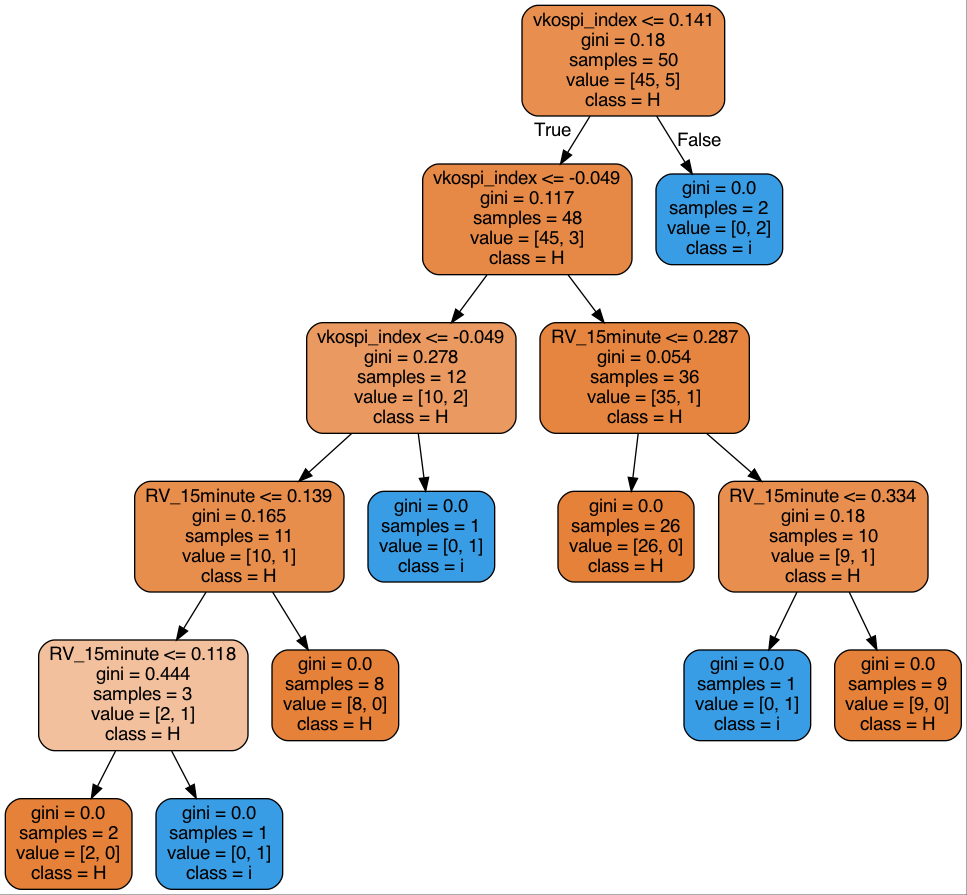
* + 1. VKOSPI INDEX + RV(frequency : 7분)

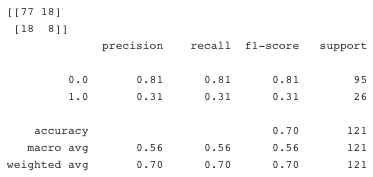
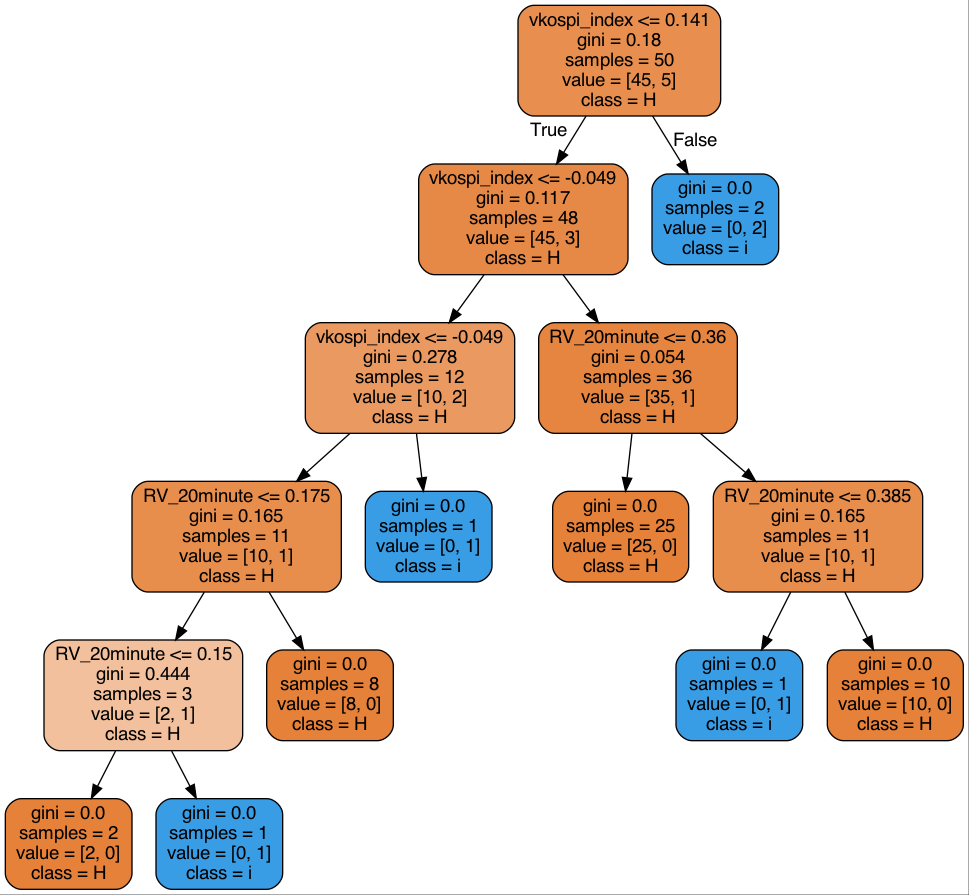
* + 1. VKOSPI INDEX + RV(frequency : 10분)

* + 1. VKOSPI INDEX + RV(frequency : 15분)

* + 1. VKOSPI INDEX + RV(frequency : 20분)

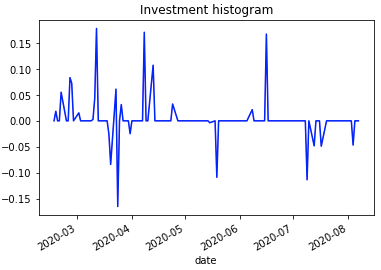
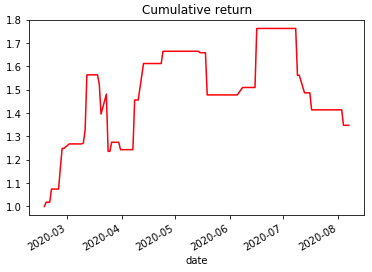
 

* 1. 실제 투자 시뮬레이션 결과

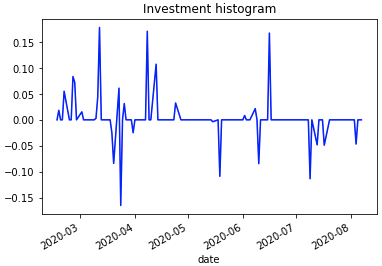
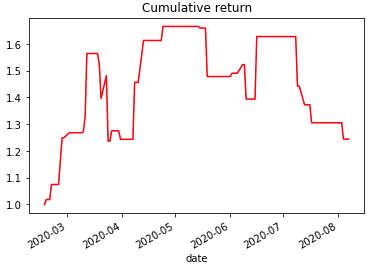
모델 별로 두 가지 그래프를 뽑았다. 좌측 그래프는 각 투자기간 별 수익률이다. 투자를 하지 않은 구간에는 0의 수익률이 발생하였고, 투자를 한 구간에서는 그 시점에 실현한 수익이 발생한다. 선물 계약의 경우 daily settlement를 한다는 점에서 당일 투자한 결과를 실현하는 이러한 특성을 잘 반영된다고 볼 수 있다.

우측 그래프는 전략의 cumulative return을 의미한다. 각 수익률이 누적으로 계산되어 시점 별로 누적 수익률 분포를 보여준다. 이를 통해, 전략의 중간 수익률과 최종 수익률을 관찰할 수 있다.

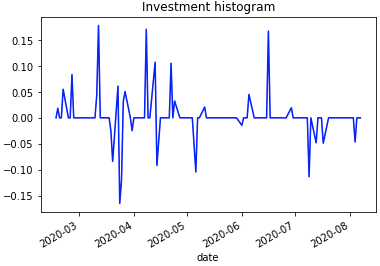
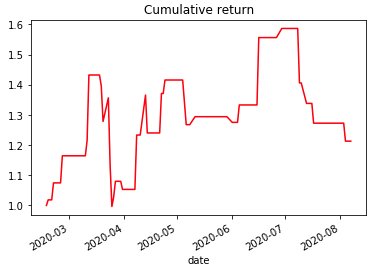
* + 1. VKOSPI INDEX + Intraday 표준편차

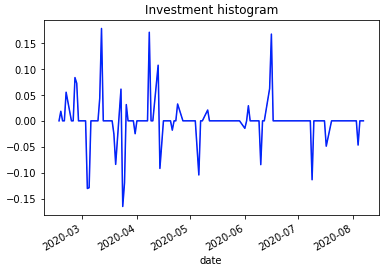
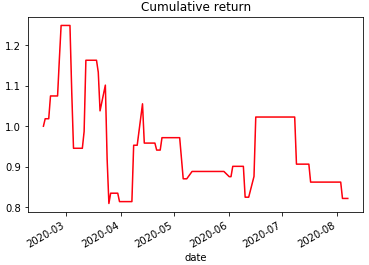
* + 1. VKOSPI INDEX + RV(frequency : 1분)

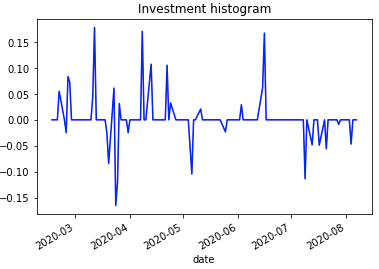
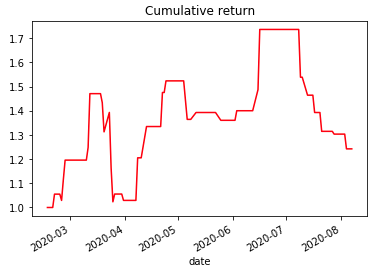
* + 1. VKOSPI INDEX + RV(frequency : 2분)

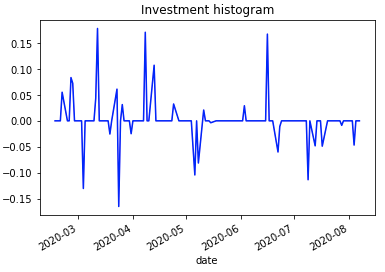
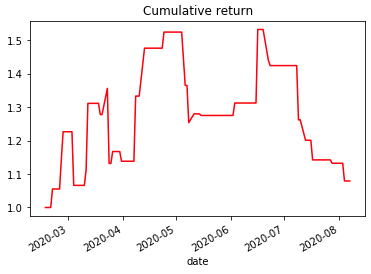
* + 1. VKOSPI INDEX + RV(frequency : 3분)

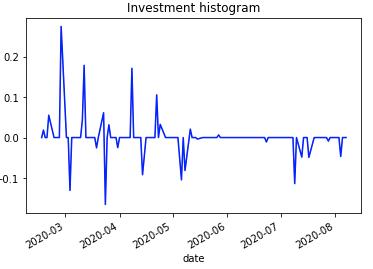
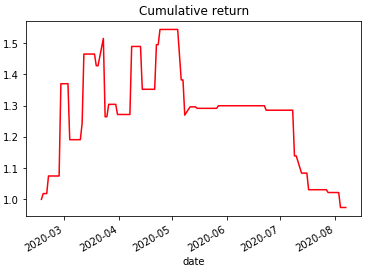
* + 1. VKOSPI INDEX + RV(frequency : 5분)

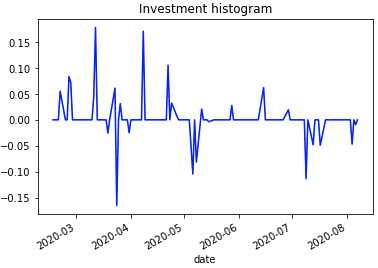
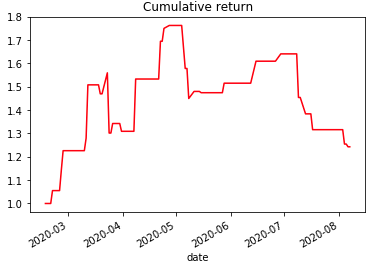
* + 1. VKOSPI INDEX + RV(frequency : 7분)

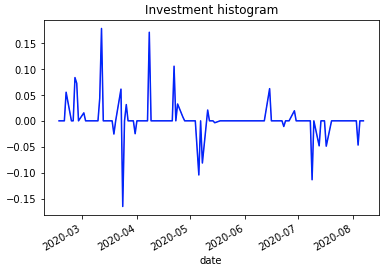
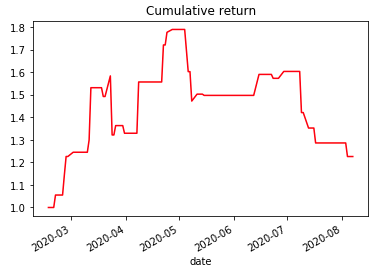
* + 1. VKOSPI INDEX + RV(frequency : 10분)

* + 1. VKOSPI INDEX + RV(frequency : 15분)

* + 1. VKOSPI INDEX + RV(frequency : 20분)

1. discussion
   1. 분류 성능 관련 discussion

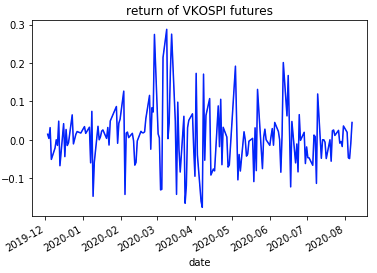
실제 급등여부를 분류하는 Decision tree의 정확도는 여러 모델이 0.7정도로 공통적으로 비슷했다. 하지만, 실제로 중요한 것은 precision(급등이라 예측한 것 중 실제 급등인 것)일 것이다. 왜냐면, 실제 투자 수익률로 귀결되는 부분은 precision이기 때문이다. Precision은 대체로 0.3 내외로 관측되었다. 결과를 표로 나타내면 아래와 같다. (단, 모델(1)~(9)는 위 4번 항목의 모델 순서 그대로 반영하도록 하겠다)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 모델 | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Accuracy | 0.71 | 0.69 | 0.68 | 0.71 | 0.71 | 0.68 | 0.65 | 0.70 | 0.70 |
| Precision | 0.32 | 0.3 | 0.28 | 0.35 | 0.34 | 0.26 | 0.19 | 0.31 | 0.31 |

위의 성능 지표를 보면, 분류 성능이 가장 좋게 관찰되는 부분은 **모델(3)(= VKOSPI INDEX + RV(frequency : 2분))**이고, 가장 좋지 않게 관찰되는 부분은 **모델(7)(= VKOSPI INDEX + RV(frequency : 10분))**이다.

추가적으로, Decision tree를 그림으로 나타낸 것의 leaf 부분을 보면, 모델이 급등이라고 판단하는 경우(파란색 네모박스)와 그렇지 않다고 판단하는 경우(주황색)에 대부분의 지니계수가 0임을 볼 수 있다. 이는 모델이 어느 정도 잘 분류한다고 볼 수 있다는 판단을 내릴 수 있었다. 실제 판단이 잘 안되는 경우에는 층이 깊어져도 지니계수가 줄어들지 않는 경우를 볼 수 있다. 물론, 이것이 overfitting이냐는 문제는 학습 데이터 셋의 양을 조절함으로써 검증을 해보아야 할 것이다.

* 1. 실제 투자 결과 관련 discussion



위 그래프는 데이터의 전체 기간 동안 VKOSPI 선물의 수익률 분포이다. VKOSPI 선물의 경우, 전체적으로 수익률의 등락이 크지만 특히 3월과 4월에서의 수익률 분포의 변화가 크게 나타난다. Decision tree 모델도 실제로 3월과 4월에 투자빈도가 높으며, 3~4월에는 전체적으로 선물가격 상승부분이 하락부분보다 많이 분포하기 때문에 3~4월의 cumulative return은 대개 상승하는 형태를 보인다. 하지만, 5월 이후부터는 전체적으로 변동성의 급등 예측을 잘 하지 못하는 것으로 보인다. 특히 2020년 6월 중순에 나타난 급등은 RV의 frequency를 높임에 따라 예측하지 못하는 모습이 보인다.

Classification 성능 중 frequency가 0.3에 그치는 데에도 불구하고 성능이 난 이유는, 모델이 “VKOSPI 선물 가격이 5% 이상 급등하는 경우에 True”라고 예측했기 때문일 것이다. 실제로 모델이 True라고 판단하였을 때 투자를 한 경우를 보면 5%이상 상승하진 않았지만 상승을 하였기 때문에 수익을 취하는 구간이 존재한다. 이렇게 보수적인 방식으로 투자를 실행하였기 때문에 모델이 생각하기에 급등한다고 판단하였지만 그것이 설령 틀렸더라도, 손해를 보기 보다도 이득을 보는 구간 또한 존재하기 때문에 precision이 좋지 않음에도 불구하고 결과적으로 (+)수익을 낸 것으로 보인다.

하지만, 5월 이후에는 모델이 큰 성적을 내지 못하는 것을 보아 모델에 대한 세부적인 분석이 필요할 것으로 보인다. 실험을 실행한 데이터의 기간은 COVID-19로 인해 시장의 변동성이 매우 큰 기간이었다. 이 때문에, 시장이 횡보하는 구간에서도 모델이 이와 같이 실행되는지에 대한 검증이 필요할 것이다. 또한, 임의적인 판단으로 모델이 50일동안 학습하고 moving window를 실행해 보았는데 이에 대한 세부적인 고려도 필요할 것으로 보인다. 학습 기간을 늘리거나 줄여보고, moving window를 하는 것이 좋은지 아니면 특정 구간만 학습을 하는 것이 결과적으로 좋은지는 더 많은 데이터 셋을 바탕으로 더 많은 실험을 거듭 반복해보아야 할 것이다.

1. 향후 과제

앞서 언급한 것과 같이, overfitting 여부를 관찰하기 위해 학습 데이터 셋의 양과 decision tree의 hyper parameter에 대한 더 세부적인 튜닝이 필요할 것으로 보인다. 뿐만 아니라, 학습 기간 정도를 고정시킨 채로 moving window를 하는 것이 옳은가 하는 여부도 관찰해 보아야 할 것이다. 또한 추가적으로 volatile한 시장보다 횡보하는 시장에서도 제시된 모델이 잘 작동하는지 추가적인 데이터로 검증할 필요가 보인다.