金 融 硏 究 Journal of Money & Finance | Vol. 33 | No. 4 | 2019. 12
http://dx.doi.org/10.21023/JMF.33.4.2

고빈도 자료를 이용한 머신러닝 모형의 예측력 비교·분석: KOSPI200 선물시장을 중심으로

박석진* · 정재식**

— 국문초록 —

본 연구에서는 KOSPI200 선물의 틱(tick) 데이터를 활용하여 머신러닝 모형의 예측력을 분석한다. 첫째, 미시구조론(microstructure)의 함의를 이용해 바(bar)를 구성했을 경우와 둘째, support vector machine, random forest와 같은 머신러닝(machine learning) 모형을 이용했을 경우 선물가격의 상승과 하락 방향에 대한 예측력이 향상되는지를 분석했다. 분석 결과 시장에 새로운 정보가 유입되는 시점을 기준으로 봉을 구성했을 때, 그리고 머신러닝 모형을 이용했을 때 예측력이 더욱 향상되는 것으로 나타났다. 머신러닝 모형의 예측력은 모형의 훈련에 사용되는 데이터의 양이 많아짐에 따라 더욱 향상되는 것으로 나타났다. 특히 거래량의 표본추출 기간을 정보의 유입여부에 따라 조정함으로써 예측력이 향상되었다는 결과는 통상적인 시간에 따라 가격의 상승 하락을 기록하고 분석하는 것은 유의미한 정보의 손실이 있음을 알 수 있다.

핵심단어 : 고빈도자료, 머신러닝, 미시구조론, 바(bar), 사적정보
JEL 분류기호 : G14, G17

투고일 2019년 10월 15일; 수정일 2019년 12월 16일; 게재확정일 2019년 12월 19일

* 제1저자, 서강대학교 경제학과 대학원 박사과정(전화: 02-705-8179, E-mail: sukjinp1@gmail.com)
** 교신저자, 서강대학교 경제학과 교수(전화: 02-705-8704, E-mail: cschung7@gmail.com)

Ⅰ. 서론

현대 금융 시장을 과거와 구분 짓는 특징은, 크게 '빅 데이터(big data)'와 '고빈도(high-frequency)'로 꼽을 수 있다. 과거에는 저장조차 할 수 없었을 정도의 많은 데이터, 그리고 존재하지 않았던 새로운 형태의 데이터가 금융 시장 분석에 사용되고 있다. 또한, 모바일 기술 및 컴퓨터 하드웨어의 발전으로 인해 투자자들은 장소에 구애받지 않고 금융 상품을 거래할 수 있게 되었으며 컴퓨터 알고리즘을 이용한 알고리즘 트레이딩 또한 계속해서 늘어나는 추세이다. 이러한 시장의 변화로 말미암아 현대 금융 시장은 거래가 밀리초(millisecond) 단위에서 일어나는 등 예전과는 비교할 수 없을 정도로 빠르게 움직이는 시장이 됐다. 특히 미시구조론 분야에서는 알고리즘 트레이딩이 늘어나면서 시장에서 정보가 발생하고 그것이 가격에 반영되는 양상이 완전히 달라졌다고 분석된다.[1)]

이러한 금융 시장의 변화와 함께 금융 시장을 분석하는 방법론에도 많은 발전이 있었는데, 머신러닝 모형들을 금융 시장 분석에 응용하기 시작한 것이 대표적인 예시다. 머신러닝이란 컴퓨터가 주어진 데이터에 어떤 속성이 있는지, 어떤 패턴이 존재하는지를 스스로 학습할 수 있도록 하는 알고리즘을 의미한다. Artifical neural network(ANN), support vector machine(SVM) 그리고 random forest(RF) 등 많은 머신러닝 모형들이 금융 시장 분석에 이용됐으며 이들은 기존의 계량경제학 방법론과 비교해 금융 데이터의 복잡하고 비선형적인 특성을 상대적으로 잘 잡아내는 것으로 나타났다.[2)]

현재까지 머신러닝 모형들을 활용한 금융 연구의 방향성은 크게 두 가지로 요약된다. 첫째는 기존의 연구들에 사용되지 않았던 머신러닝 모형들을 연구에 접목하여 모형의 예측력을 최대한 끌어올리는 경우(윤종문, 2019; Laborda and Laborda, 2017)이다. 둘째는 자연어 처리(natural language processing) 알고리즘이나 감성 분석(sentiment analysis)을 통해 텍스트 데이터를 변수로 활용하는 등, 머신러닝 기법을 통해 기존에는 금융 연구에 사용되지 않았던 변수들을 분석에 활용(김용석, 조성욱, 2019; Gentzkow, Kelly, and Taddy, 2017)하고 있다.

머신러닝 모형과 이를 이용한 새로운 변수들이 금융 분야에서 높은 관심을 얻고 있는 반면에 모형에 사용되는 데이터를 어떻게 구성할 것인지에 대한 논의는 거의 이루어지지 않고 있다. 특히 유가증권시장 연구에서 많은 비중을 차지하는 거래 데이터의

1) 고빈도 금융시장에 대한 미시구조론 분석은 O'Hara(2015)에서 확인할 수 있다.
2) Lopez de Prado(2018b)는 금융 분야에 머신러닝을 성공적으로 접목한 사례들을 정리했다.

경우 원자료인 틱 데이터로부터 시가, 고가, 저가, 종가 및 거래량 등 정보를 요약적으로 담고 있는 바(bar)를 어떤 기준에 의해 구성할 것인지에 대한 확립된 기준은 아직 존재하지 않는다.

현재 실무 및 연구에 가장 많이 쓰이는 것은 일정한 시간, 그리고 거래횟수마다 바를 구성한 시간 바(time bar)와 틱 바(tick bar)다. Lopez de Prado(2018a)는 이들이 직관적으로 간단하다는 장점이 있지만 금융 시장 분석에 있어 유의미한 정보를 추출하는 데는 비효율적이라고 지적한다. 이에 대한 대안으로 시장에서 정보가 발생할 때마다 바를 구성하는 정보 기반 바(information-driven bar)들을 제안하는데, 이 중 VIB(volume imbalance bar)는 주문 흐름(order flow)이 기대 수준 이상으로 매수 혹은 매도 한쪽으로 몰릴 때마다 바를 구성하는 방법이다. 이는 주문 흐름이 사적정보를 가진 정보우월자(informed trader)의 거래행위와 밀접한 관계가 있다는 미시구조론의 함의를 바탕으로 하고 있다.

Lopez de Prado(2018a)는 VIB 등 정보 기반 바들이 정보를 추출하는데 효율적이라고 주장하지만, 아직까지 그 실효성이 확인된 바가 없다. 따라서 본 연구는 첫째, 고빈도 가격 움직임 예측에 VIB를 적용해 그 실효성을 확인하고자 한다. 이를 위해 기존에 많이 쓰이고 있는 시간 바(bar), 틱 바(bar) 그리고 거래량 바(volume bar)를 비교 지표로 삼았다. 만약 VIB가 현대의 고빈도 시장에서 불규칙적으로 발생하는 정보를 효과적으로 반영한다면 기존 바들과 비교해 높은 예측력을 보일 것이다.

둘째, 고빈도 가격 움직임을 예측하는데 머신러닝 모형의 예측 성능과 기존 계량경제학 모형의 예측 성능을 비교해 머신러닝 모형의 실효성을 확인하고자 한다. 머신러닝 모형을 이용했을 때 예측 정확도가 유의미하게 향상되는지를 보기 위해 로지스틱 회귀(logistic regression)모형을 벤치마크로 활용했으며 머신러닝 모형 중 Support vector machine과 Random forest를 이용했다. 표본추출 방법에 따른 예측 정확도의 변화를 관찰하기 위해 각 바에 담긴 가장 기본적인 변수들인 수익률, 거래량, 그리고 주문 흐름을 예측변수(predictor)로 이용했다. 실증 분석은 KOSPI200 지수 선물의 틱 데이터를 이용해 진행했으며 표본 기간은 2018년 11월 21일부터 2019년 7월 2일까지다.

연구 결과를 요약하면 다음과 같다. 첫째, 틱 바와 거래량 바는 표본 외 예측에서 모형과 관계없이 50% 전후의 정확도를 기록해, 예측력이 없는 것으로 나타났으며 시간 바는 다소의 예측 정확도 상승이 있었지만 유의미한 수준은 아니었다. 둘째, VIB는 최소 65% 이상의 예측 정확도를 기록해 고빈도 가격 움직임에 대한 예측력을 가진 것으로

나타났다. 이는 VIB가 시장에서 불규칙적으로 발생하는 정보를 효과적으로 반영하는 반면에 다른 바들은 그렇지 못했기 때문으로 판단된다. 마지막으로, 예측모형의 훈련에 사용되는 훈련 데이터의 양이 커질수록 SVM과 RF의 예측 정확도가 로지스틱 회귀모형과 비교해 더욱 향상되는 것으로 나타나, 고빈도 가격 움직임을 예측하는데 있어 머신러닝 모형의 실효성을 확인했다.

서론 이후 논문의 구성은 다음과 같다. 제Ⅱ장에서는 기존의 연구들을 살펴보면서 본 연구와의 차이점을 기술한다. 제Ⅲ장에서는 VIB 및 기존의 데이터 구성방법 및 본 연구에 사용되는 머신러닝 모형들을 설명한다. 제Ⅳ장에서는 본 연구에 사용된 데이터 및 기초통계량을 제시한다. 제Ⅴ장에서는 예측방법론 및 실증 분석 결과를 제시하며, 마지막 Ⅵ장에서는 연구의 결과를 정리한다.

Ⅱ. 선행 연구

Samuel(1959)이 머신러닝의 개념을 정립한 이후 다양한 분야에서 머신러닝이 응용됐으며 최근에는 금융 분야에서도 머신러닝을 응용한 연구가 활발하게 진행되고 있다. 머신러닝이 금융 연구에 접목되는 양상은 크게 두 가지다. 첫째는 기존 금융 모형들에 머신러닝의 방법론들을 접목해 모형의 예측 성능을 올리는 것이다. Gu, Kelly, and Xiu (2018)은 자산 가격 결정 모형에 다양한 머신러닝 모형을 적용해 전통적인 계량경제학 모형과 비교했을 때 결정계수가 개선됨을 보였다. Rossi(2018)는 decision tree를 이용한 boosting 알고리즘으로 주식 수익률과 변동성에 대해 표본 외 예측을 시행한 결과 기존의 방법론들에 비해 예측 정확도가 향상됨을 보였다. 국내 연구로는 support vector machine (SVM)을 이용해 변동성 지수인 VKOSPI의 변화를 예측해 이를 옵션 매매에 적용한 라윤선, 최홍식, 김선웅(2016) 및 K-Nearest Neighbors(K-NN) 알고리즘을 통해 Kospi200 지수 선물의 가격을 예측한 김명현, 이세호, 신동훈(2015)가 있다.

둘째는 머신러닝 및 빅 데이터 기법을 활용해 기존에 사용되지 않은 변수를 금융 분석에 응용하는 것이다. Vlastakis and Markellos(2012)은 구글 검색어 데이터를 활용해 투자자들의 정보에 대한 수요와 공급을 동태적으로 분석했다. Jegadeesh and Wu(2013)는 감성 분석을 통해 기업들의 연간 보고서, 애널리스트 보고서 등의 어조를 수량화해 그것이 시장 반응의 방향성 및 지속기간과 유의미한 관계가 있음을 보였다. 국내에서는 김동영,

박제원, 최재현(2014)가 SNS와 뉴스 기사의 감성 분석 및 머신러닝 모형을 이용하여 주가에 대한 예측 정확도가 향상됨을 보여 국내 금융 시장 분석에 있어 텍스트 분석의 중요성을 시사했다.

이와 같이 최근 머신러닝 모형을 이용한 금융 연구가 활발해지고 있는 한편, 고빈도 금융 데이터와 관련된 연구 또한 많은 연구가 진행되고 있다. Easley, Kiefer, and O'Hara (2012)은 컴퓨터 알고리즘을 이용한 초단타 매매가 많아짐에 따라 정보적 열위자/시장 조성자는 거래량에서 추출된 정보를 활용할 것을 제안했다. Easley, Lopez de Prado, and O'Hara(2016)은 매매수도(bid-ask)분류를 통한 거래량 분류 알고리즘이 사적정보를 추출하는데 적절함을 보인 바 있다. Chinco, Clark-Joseph, and Ye(2019)은 LASSO 회귀분석을 통해 1분 단위에서 주가 예측을 할 때 고빈도 수익률에 대해 예측력이 있는 정보는 불규칙적으로 발생한다는 것을 보였다. 또한, Easley, Lopez de Prado, and O'Hara(2019)은 고빈도 금융 데이터와 머신러닝 모형을 이용해 미시구조론 분야에서 사적정보와 관련된 경제변수들이 시장의 변동성과 호가 스프레드 등의 시장 변수들에 대해 예측력을 가진다는 것을 보였다. Lopez de Prado(2018a)는 데이터 구성, 모형 학습, 성능 측정 등 금융 머신러닝 방법론에 대해 심도 깊은 분석을 제공했으며, 특히 미시구조론의 함의를 바탕으로 데이터를 구성하는 VIB(Volume Imbalance Bar)를 제안했다.

본 연구는 미시구조론의 함의를 이용한 VIB가 고빈도에서 예측력을 가지는지를 머신러닝 방법론을 통해 살펴본다는 측면에서 기존의 금융 머신러닝 및 고빈도 금융 데이터 분석의 연구들과 관련되어 있다. 특히 VIB는 Lopez de Prado(2018a)가 제안했다는 점, 그리고 VIB가 미시구조론에 바탕을 두고 있다는 점에서 Lopez de Prado(2018a)와 Easley et al.(2019)과 가장 밀접한 관련이 있다. 또한, 고빈도 가격 움직임을 머신러닝 모형을 통해 예측한다는 측면에서 Chinco et al.(2019)의 연구와도 관련이 있다. 하지만 본 연구는 다음의 세 가지 측면에서 기존의 연구들과 차이점이 존재한다. 첫째, 예측 정확도를 높이기 위해 머신러닝 모형 또는 새로운 t설명변수에 집중했던 기존 연구와 달리, 본 연구는 거래량 바(bar)를 이용한 정보가 예측력에 영향을 미침을 보였다. 바(bar)를 이용한 예측력을 분석한, 즉 VIB의 실효성을 확인한 점 역시 본 연구의 기여로 판단된다. 또한, 본 연구에서는 미시적구조 관련 변수(사적정보를 담고 있는 변수)를 측정함에 일별 또는 통상적인 clock를 사용하지 않고, 뉴스 또는 이벤트가 발생하는 경우에만 미시적 구조 관련 변수를 활용하였으며, 저자들이 아는 범위에서는 처음 시도된 것으로 사료된다.[3)]

3) 박석진(2019. 2)에서는 통상적인 time clock를 활용하여 분석한 바 있다.

Ⅲ. 정보추출(Information Extraction) 모형과 예측모형

본 절에서는 틱단위로 기록된 KOSPI200 선물 자료를 어떤 형태로 기록할 것이며, 이들의 차이점과 경제학적 함의가 무엇인지 살펴본다. 또한 본고에서 사용되는 머신러닝 알고리즘의 특징을 개괄한다.

가격의 변화를 측정하는 가장 일반적인 방법은 영업시간을 기준으로 측정하는 것이다. 가격의 변화율을 기록할 때 월간, 일간, 시간별 등으로 기록하고 있다. 이는 인간의 경제활동 시간(human clock)을 기준으로 측정하는 것으로, 경제사건이 발생하는 시간(event clock)과는 다를 가능성이 크다. 한국 증시가 15시 30분에 폐장[4]되지만, 한국 경제 및 기업 관련 뉴스는 지속적으로 생산되고 있다. 24시간 거래되는 국제통화 외환시장 및 중국 관련 뉴스는 계속 발생하고 있다. 일별 변동성 및 수익률 계산은 회계적 목적 및 계약에는 매우 편리할 수 있으나, 정보의 측정 측면에서 손실(loss)이 발생한다고 볼 수 있다.

영업시간 동안에 측정되는 주가의 수익률 또는 관련 통계량 역시 동일한 맥락에서 정보의 손실이 발생할 가능성이 크다. 9시 개장 후 5분간의 변동성이나 거래량이 12시 전후의 5분 단위의 그것들과 동일한 선상에서 해석되는 것은 바람직하지 않을 수 있다. 사적정보 관련 정보의 유입이나, 투자자들의 뇌동매매와 같은 비경제적 요인은 시장참여자들의 투자행위에서 찾는 것이 필요하며, 이러한 관점에서 다양한 거래 바(bar)를 이용해서 정보가 존재하는지, 선물가격 예측에 도움이 되는지 살펴볼 필요가 있다.

다양한 바(bar)의 구성을 통해 시장에 유입된 정보를 측정해보고 이를 예측모형에 활용한다. 여기서 바(bar)는 시장에 유입된 정보를 측정하는 하나의 단위로 볼 수 있다. 즉 1시간이 아니라 비정기적 시간(irregular time span)에 발생한 뉴스 이벤트로 볼 수 있다. 바(bar)는 시작시점, 종결시점, 듀레이션, 시가, 종가, 고가, 저가, 거래량 등의 정보를 담고 있다. 통상 한국 증권사에서 제공되는 HTS의 바(bar)와 같은 개념이지만, 본고에서 활용되는 다양한 바(bar)는 경제적 사건을 기준으로 만들어졌다고 볼 수 있다. 예를 들어 혼합분포모형(mixture model)의 경우 거래건수가 시장에 유입된 정보의 양으로 측정하고 있으며 이는 틱 바(tick bar)를 통해 측정할 수 있다. 또한, 거래량이 공적정보를 담고 있다면 이는 특정 규모 이상의 거래량을 담고 있는 거래량 바(volume bar)를 통해 측정할 수 있다.

4) 정규시간(즉 장후 시간 외 종가, 시간 외 단일가 시간 제외).

머신러닝 시대를 맞아 다양한 알고리즘이 금융시장에 활용된 바 있다. 시대에 따라 support vector machine(SVM), Random forest(RF)의 방법[5]이 가장 많은 관심을 받은 바 있어, 이들을 다양한 바(bar)의 예측력을 측정한다.

1. 정보추출 모형: 바(Bar)[6]

(1) 시간 바(Bar), 틱바(Tick Bar) 및 거래량 바(Trading Volume Bar)

바(bar)를 구성하는 가장 표준적인 방법은 틱 데이터를 일정한 길이로 나누는 방법이다. 그 중에서 일정 시간마다 바를 구성한 것이 시간 바(bar), 일정 거래횟수가 기록될 때마다 바를 구성한 것이 틱 바(bar), 그리고 일정 거래량이 누적될 때마다 바를 구성한 것이 거래량 바(bar)다.

이중 일반적으로 사용되는 것은 시간 바와 틱 바로, 국내 증권사에서 제공하는 HTS(home trading system)에서도 조회할 수 있다. 거래량 바는 시간 바와 틱 바만큼 대중적으로 사용되지는 않지만, Easley, Lopez de Prado, and O'Hara(2016) 등의 학술 연구들에서 이용했다. 이들은 직관적으로 이해하기 쉽다는 장점이 있지만, 분석결과가 데이터의 시작점이 어디인지에 따라 가변적일 수 있다. 또한, 가격의 움직임에 영향을 미치는 경제적 사건들은 불규칙적으로 발생하기 때문에 일정한 간격으로 데이터를 구성하는 이들 바의 특성상 이러한 정보를 효과적으로 반영하기 힘들다고 볼 수 있다.

(2) VIB(Volume Imbalance Bar)

기존 바들에 대한 대안으로 de Prado(2018a)는 시장에 새로운 정보가 발생하는 시점들을 기준으로 바를 구성할 것을 제안하며, 이를 VIB(volume imbalance bar)로 명명한 바 있다. 미시구조론의 함의를 근간으로, 주문흐름(order flow)이 기대 수준에서 벗어날 때를 시장에 새로운 정보가 발생한 시점으로 간주한다. 사적정보를 가진 정보우월자들이 언제 거래에 나서는지, 어떻게 거래하는지, 그리고 그것이 시장에 미치는 영향이 무엇인지 등은 미시구조론에서 중요하게 다루는 주제다. Glosten and Milgrom(1985)과 Easley et al.(1997) 등 정보우월자들의 거래행위를 다룬 연구들에 의하면 이들이 거래에 나설 때 가지고 있는 정보의 성격에 따라 주문 흐름이 매도나 매수 한쪽으로 몰리게 된다.

---

5) Boosting(adaboost, xgboost) 역시 최근 많이 활용되고 있으나, 이는 underfitting에 적합하며, 금융시장 모형은 overfitting과 분산이 더 큰 문제로 판단되기 때문에 본고에서는 분석하지 않았다.

6) 본 절은 Lopez de Prado(2018a)를 참고해 작성했다.

주문 흐름은 매수주도 주문 거래량에서 매도주도 주문 거래량을 뺀 것으로[7], 사적정보를 가지지 않은 정보열위자들은 거래에 나서는 경우 매수나 매도 주문을 진행할 확률이 동일하며 이들의 주문 흐름은 정보의 유무와 상관없이 평균적으로 0이 된다. 반면에 정보우월자들은 오로지 사적정보가 있을 때만 거래에 나서는데, 그들이 가지고 있는 정보가 좋은 정보인지, 아니면 나쁜 정보인지에 따라 매수 혹은 매도 거래만 진행하게 된다. 즉 시장에 새로운 정보가 발생하면 정보우월자들의 거래로 인해 주문 흐름이 균형에서 벗어나게 된다.

이를 바탕으로 VIB는 주문 흐름이 기대 수준에서 벗어날 때 시장에 새로운 사적정보가 발생했다고 보고 그러한 시점들을 기준으로 바를 구성한다. 먼저, T시점의 주문 흐름 θ_{T}는 다음과 같이 정의된다.

$$\theta_{\mathrm{T}}=\sum_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{b_t}\mathrm{V_t}$$

b_t는 체결된 거래가 매수주도 거래인지, 또는 매도주도 거래인지에 따라 1 또는 -1의 값을 가지며 V_t는 체결 거래량을 의미한다.

둘째, 하나의 바이 시작될 때 θ_{T}의 기댓값은 다음의 수식으로 나타낼 수 있다.

$$\begin{aligned}E_0[\theta_T] &= E_0[\sum_{t|b_t=1}^{T} V_t] - E_0[\sum_{t|b_t=-1}^{T} V_t] \\ &= E_0[T](P[b_t=1]E_0[V_t|b_t=1] - P[b_t=-1]E_0[V_t|b_t=-1])\end{aligned}$$

위 식에서 $P[b_t=1]E_0[V_t|b_t=1]$을 V^+, $P[b_t=-1]E_0[V_t|b_t=-1]$을 V^-으로 표기하면,

$$V^+ + V^- = E_0[T]^{-1}E_0[\sum_t V_t] = E_0[V_t]$$

으로 나타낼 수 있다. 즉, 체결 거래량에 대한 초기 기댓값은 매수주도 거래량에 의한 부분과 매도주도 거래량에 의한 부분으로 분해할 수 있다. 이를 이용해 $E_0[\theta_T]$를 다음과 같이 나타낼 수 있다.

---

7) 매수주도 거래는 시장가 주문을 통한 매수 주문을 의미하며, 매도주도 거래는 시장가 주문을 통한 매도 주문을 의미한다.

$$E_0[\theta_T] = E_0[T](V^+ - V^-) = E_0[T]E_0(2V^+ - E_0[V_t])$$

위 식을 바탕으로, VIB는

$$T^* = \arg\min_T \{|\theta_T| \geq E_0[T]|2V^+ - E_0[V_t]|\}$$

을 만족하는 T^*들을 기준으로 바를 구성하는 방법이다. T^*를 추정하기 위한 임계치는 $E_0[T]$와 $|2V^+ - E_0[V_t]|$의 곱으로 이루어져 있는데, 이들은 각각 이전에 구성된 바들의 T값 및 θ_T들의 지수 이동평균으로 추정한다.

$|2V^+ - E_0[V_t]|$은 주문 흐름의 불균형 정도에 대한 기댓값을 나타내는데, θ_T가 기대 수준 이상으로 균형에서 벗어나면 T^*는 작아지며, 반대로 θ_T가 기대 수준에서 크게 벗어나지 않는다면 T^*는 커진다. 미시구조론에 의하면 θ_T가 기대 수준에서 벗어나는 경우는 정보우월자들이 거래에 나설 때이기 때문에, VIB는 이에 맞춰 바들을 구성함으로써 정보우월자들이 가진 사적정보를 반영한다고 이해할 수 있다. 따라서 각각의 바에는 동일한 양의 정보가 담겨 있으며, 정보우월자들이 많이 거래할수록 더 많은 바가 형성된다.

2. 머신러닝 모형[8)]

머신러닝은 데이터가 주어졌을 때 그것의 속성이나 패턴 등을 컴퓨터가 스스로 학습할 수 있도록 하는 일련의 알고리즘들을 의미한다. 그중에서 지도학습(supervised learning) 모형들은 종속변수(dependent/feature variables)가 존재해 그것과 예측 변수(predictor variable) 사이의 관계를 도출하는 것을 목표로 한다. 지도학습 모형은 종속변수가 연속적인 값을 가지는지, 또는 특정 카테고리(category)에 속하는 범주형 데이터인지에 따라 회귀(regression)와 분류(classification)로 나눌 수 있다. 자산의 가격 수준을 예측하는 모형들은 회귀에 속하며 가격의 움직임을 예측하는 모형들은 분류에 속하는 것으로 이해할 수 있다.

본 연구에서는 바의 종류에 따른 예측력의 변화를 확인하기 위해 분류 모형 중

8) 본절의 내용은 Murphy(2002), Geron(2017), James, G., D. Witten, Hastie, and Tibshirani(2017)을 요약 정리했다. 확률적 머신러닝은 Murpy, python을 이용한 실증 분석은 Geron, R을 이용한 분석은 Jame et al.(2017)을 참조했다.

Support vector machine(SVM)과 Random forest(RF)를 사용한다. SVM은 분류 문제에 있어 종속변수와 예측변수 사이에 비선형 관계가 존재할 때 이를 더 고차원의 공간으로 투사(mapping)함으로써 선형관계로 변환하는 한편, RF는 여러 개의 decision tree들을 훈련한 후 과반수 선출(majority voting)로 종속변수를 예측한다. 이번 절에서는 SVM과 RF 모형에 대해 더 구체적으로 소개를 한다.

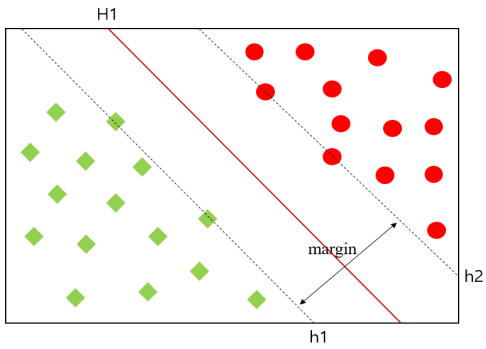
(1) Support Vector Machine(SVM)

SVM은 서로 다른 집단에 속한 범주형 데이터를 분류하는 최적의 초평면을 찾는 모형이다. 먼저 SVM을 훈련하는데 이용하는 데이터 집합 D를 다음과 같이 정의한다.

$$D=\left\{(x_i,\ y_i)|x_i\in\mathbb{R}^m,\ y_i\in\{-1,\ 1\}\right\}_{i=1}^{n}$$

각 x_i는 m차원의 실수 벡터를 나타내며 y_i는 x_i가 어떤 범주에 속해있는지에 따라 1 또는 -1의 값을 가진다. 다음으로 y_i의 값에 따라 데이터를 기하학적으로 분리할 수 있을 때, 그 경계면을 초평면(hyperplane)이라 부르며 초평면과 가장 가까운 데이터들을 서포트 벡터(support vector)라고 정의한다. SVM은 이러한 서포트 벡터들을 바탕으로 두 분류 집단 사이의 공백(margin)을 최대화하는 초평면을 찾으며, 그 구조를 <Figure 1>과 같이 나타낼 수 있다.

<Figure 1> Structure of SVM



두 분류 집단을 구분하는 선형 경계면을 $f(x)=w^Tx+b$으로 정의하면 최적의 경계면을 찾는 문제는 두 분류 집단의 서포트 벡터들 사이의 거리인 $\frac{2}{\|w\|^2}$을 최대화하는

문제로 나타낼 수 있는데, 이는 $\frac{1}{2} \| w \|^2$을 최소화하는 것과 동치다. 마지막으로 각 집단에 속하는 데이터들이 경계면을 기준으로 같은 방향에 위치하도록 $y_i(w^T x + b) \geq 1$의 제약조건을 적용하여 비용함수를 최소화하는 w와 b를 계산하면 최적의 경계면을 찾을 수 있으며, 이를 수식으로 표현하면 다음과 같다.

$$\min_{w,b} \frac{1}{2} \| w \|^2$$

$$s.t.\ y_i(w^T x_i + b) \geq 1 \quad \text{for } i = 1, \cdots, N$$

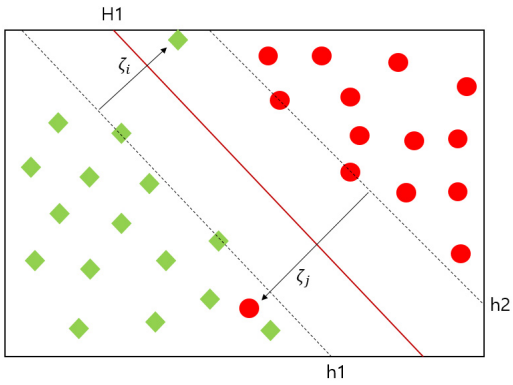
위의 SVM 모형은 선형 경계면에 의해 데이터가 완전하게 두 집단으로 분류되는 경우를 산정하고 있다. 하지만 이 모형은 데이터가 완전하게 선형 분리가 되지 않는 데이터들을 다루는 경우에는 적용할 수 없는데, 이 경우에는 유화변수(slack variable) ζ_i과 패널티 변수(penalty variable) C를 도입하여 해결한다.

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \zeta_i$$

$$s.t.\ y_i(w^T x_i + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 1 \quad \text{for } i = 1, \cdots, N$$

ζ_i는 데이터가 SVM의 마진 안에 위치할 수 있도록 해주는 역할을 하는 반면, C는 마진 안에 위치한 데이터들에 대해 부과하는 패널티로 이해할 수 있다. 이를 그림으로 나타내면 <Figure 2>와 같다.

<Figure 2> Structure of SVM with ζ and C



마지막으로 두 분류 집단을 나누는 초평면 자체가 비선형인 경우가 있다. 이 경우에는 커널 함수(Kernel function) $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ 을 이용해 데이터를 원 공간(input space)으로부터 선형으로 분류할 수 있는 고차원의 특징 공간(feature space)으로 투사한 뒤 두 집단을 나누는 초평면을 찾는다. 이러한 커널-SVM(Kernel-SVM)을 그림으로 나타내면 <Figure 3>과 같다. 커널-SVM에 적용하는 커널 함수는 다양한 형태를 가질 수 있는데, 기존 연구들에서 가장 많이 사용된 커널 함수들은 다음과 같다.

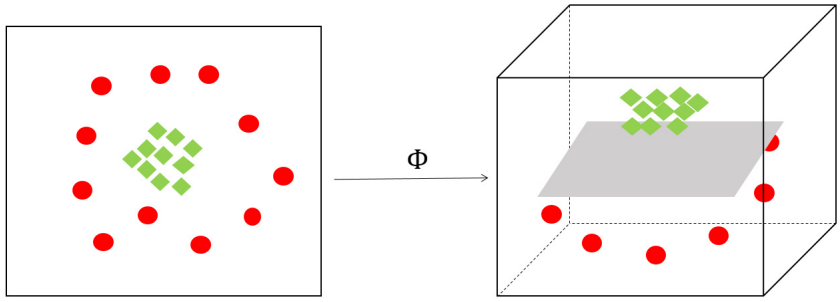
$$Linear: \ K(x_i, x_j) = x_i^T x_j,$$

$$Polynomial: \ K(x_i, x_j) = (x_i^T x_j + c)^d, \ c > 0,$$

$$Sigmoid: \ K(x_i, x_j) = \tanh\{a(x_i^T x_j) + b\}, \ a, b \geq 0,$$

$$Gaussian: \ K(x_i, x_j) = \exp - \frac{\| x_i - x_j \|^2}{2\sigma^2}, \sigma \neq 0.$$

<Figure 3> Structure of Kernel-SVM.



(2) Random forest(RF)

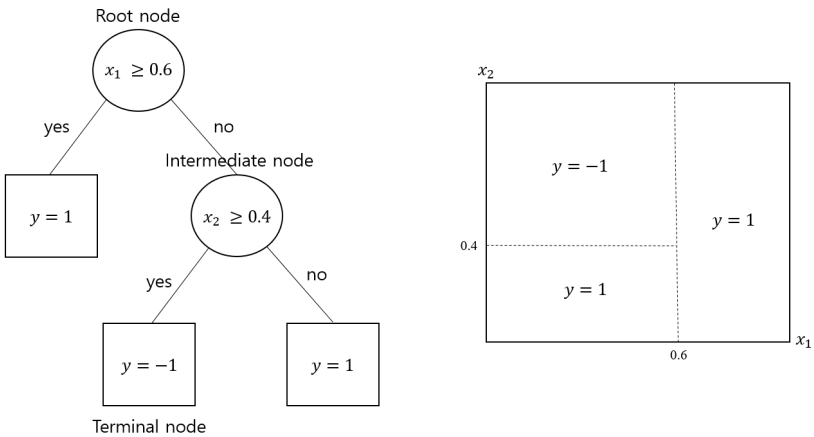
RF는 다수의 독립적인 decision tree를 훈련해 이들의 다수결을 바탕으로 예측치를 도출한다. 먼저, decision tree는 여러 개의 분류 규칙을 통해 데이터를 여러 개의 부분집합으로 나눠 예측치를 도출한다. <Figure 4>는 일반적인 decision tree의 구조를 나타낸다.

먼저 데이터는 초기 분리지점인 root node를 시작으로 나눠지기 시작하며 분기가 거듭될수록, 즉 중간 마디(intermediate node)들을 계속해서 지날수록 계속해서 나눠지며

각 집합에 속한 데이터의 개수는 줄어든다. 분류 문제에 적용되는 경우 끝 마디(terminal node)에 모인 데이터들의 범주에 따라 데이터를 각각의 집단으로 분류한다.

decision tree는 각각의 노드에서 데이터를 분류할 때 분류 뒤 각 영역의 불순도(impurity)를 최소화하는 방향으로 학습을 하며, 불순도는 엔트로피(entropy), 지니 계수(Gini index), 또는 오 분류 오차(misclassification error)를 이용해 측정한다.

<Figure 4> Structure of decision tree.



Decision tree는 모형의 학습이 간단하고 모형의 구조를 시각화할 수 있다는 장점이 있지만, 학습에 사용된 데이터가 조금만 달라지더라도 모형의 구조가 달라질 수 있다는 불안정성 및 표본 외 예측에서 높은 예측 오차를 보이는 과적합(overfitting)[9] 문제가 발생할 확률이 높다는 단점 또한 지닌다.

RF는 개별적인 decision tree가 지닌 불안정성 및 과적합 문제를 보완하기 위해 다수의 훈련된 decision tree를 사용하는 앙상블(ensemble)[10] 학습 방법이다. RF는 배깅(bagging)[11] 알고리즘을 통해 각 decision tree를 훈련할 때 무작위성을 도입한다. 즉, 각각의

9) 과적합은 머신러닝 모형이 훈련 데이터를 과하게 학습해, 잡음(noise)에 해당하는 데이터까지 모형화하는 현상이다. 과적합이 발생하면 훈련 데이터에 대해서는 높은 예측 정확도를 보이지만, 모형화에 이용하지 않은 새로운 데이터가 주어졌을 때 예측 오차가 크게 나타나게 된다.

10) 앙상블 학습은 다수의 머신러닝 모형을 학습해 그 모델들의 예측 결과를 기반으로 하나의 예측치를 도출하는 방법론이다.

11) 배깅 알고리즘은 부트스트래핑(bootstrapping)을 통해 훈련데이터에서 샘플을 여러 번 추출해 각 모형을 학습시켜 결과를 집계하는 알고리즘이다.

decision tree들은 같은 데이터를 바탕으로 훈련되는 것이 아니라 전체 데이터에서 무작위로 추출된 표본들을 바탕으로 훈련된다. 이렇게 서로 다른 데이터들을 통해 훈련된 decision tree들을 사용해 예측치를 도출함으로써 데이터에 따라 모형에 달라지는 불안정성을 해소한다.

더 나아가, 각각의 decision tree를 훈련할 때 사용되는 데이터와 마찬가지로 예측변수들 또한 무작위로 선택된 일부분만 사용하게 된다. 예를 들어 자산의 가격 변동성을 예측하는데 거래량이 가장 높은 예측력을 가지는 예측변수라고 하자. 이런 경우에 모든 예측변수를 이용해 decision tree들을 훈련하는 경우 대부분 거래량을 중심으로 모형이 구성되어 decision tree들이 유사한 구조를 가지게 된다. RF는 각 마디에서 데이터를 나눌 때 무작위적으로 선택된 예측변수들을 이용함으로써 나무들 사이의 유사도를 낮춘다.

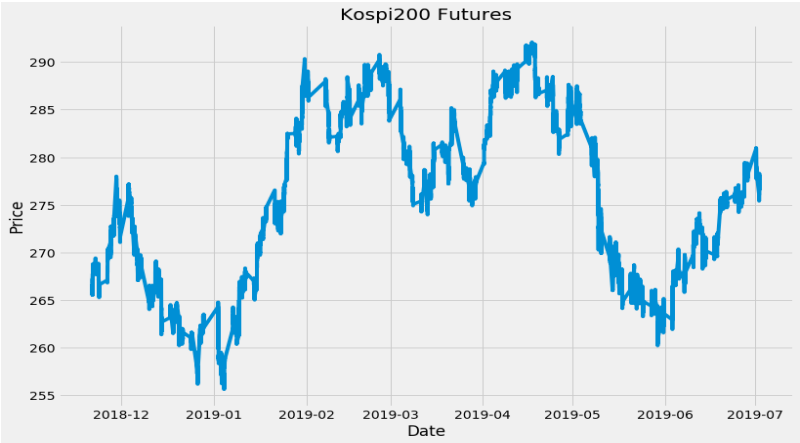
요약하면, decision tree는 일반적으로 훈련 데이터(training data)에서의 성능은 탁월하다. 그러나 실험 데이터(test data)에서의 예측 성능은 떨어지는 과적합 문제와 예측분산이 높아지는 경향을 보인다. RF는 배깅 알고리즘을 통해 decision tree를 다수 생성함으로써 표본 외 예측에서의 예측 오차를 줄이는 것으로 이해될 수 있다.

Ⅳ. 데이터 및 기초통계량

본 연구에서는 Kospi200 지수 선물의 실시간 체결 데이터를 실증 분석에 이용한다. 표본 기간은 2018년 11월 21일부터 2019년 7월 2일까지 총 150 거래일이며[12], 총 데이터 관측 수는 525만 6743개다. 데이터는 국내증권사의 HTS(home trading system)를 이용해 수집했다. <Figure 5>는 Kospi200 지수 선물의 시계열을 도식화한 것이다. 표본 기간이 150 거래일이라는 점에서 각 바(bar) 구성법들이 일별 이상의 저빈도에서 가격에 대해 예측력을 가지는지 분석하기에는 부족하다. 하지만 총 관측 수가 500만 개 이상이라는 점에서 바(bar) 단위의 고빈도에서 분석을 진행하기에는 충분하다고 판단된다. 특히 분석 기간 중 Kospi 선물이 등락을 거듭하고(지속적인 상승이나 하락이 아닌) 있어 짧은 거래일수가 문제가 되지 않는다고 판단된다.

12) 자료기간은 이용 가능성(availability)으로 임의적 기간이며, 특정한 경제적 논거에 의해 결정된 것은 아니다.

<Figure 5> Time series of Kospi200 Futures



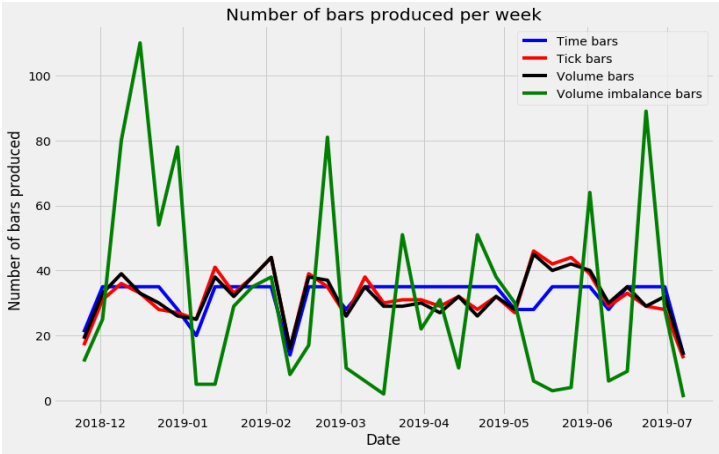
다음으로는 각 바(bar) 구성법들에 따라 데이터들을 구성했다. 각 바(bar)를 통해 추출한 정보가 고빈도 가격 움직임에 대해 예측력을 가지는지 비교하기 위해서 각 바의 구성 빈도를 비슷하게 맞춰야 한다. 앞에서 다뤘듯이 VIB(Volume Imbalance Bar)는 주문 흐름이 기대 수준에서 벗어날 때마다 바(bar)를 구성하기 때문에 사전적으로는 총 몇 개의 바(bar)가 구성되는지 알 수가 없다. 따라서 먼저 VIB를 이용해 바를 구성한 후 시간 바(bar), 틱 바(bar) 및 거래량 바(bar)는 VIB로 구성된 바의 개수에 맞게 기준을 설정했다.

원자료를 바탕으로 VIB를 적용해 데이터를 구성한 결과 총 1,038개의 바(bar) 가 만들어졌다. 시간 바(bar)는 60분마다, 틱 바는 5006번의 거래가 있을 때마다, 그리고 거래량 바는 29206계약이 거래될 때마다 바(bar)를 구성해 각각 1049개, 1050개, 그리고 1049개의 바를 만들었다. <Figure 6>은 각 바(bar) 구성법들을 이용해 구성한 바들의 주별 개수를 나타낸다.

먼저 시간 바의 경우 시간에 따른 바(bar) 개수의 변동성이 거의 없으며, 휴일 등으로 장이 열리지 않았던 때만 바(bar) 개수가 감소한 것을 볼 수 있다. 다음으로 틱 바(bar), 거래량 바(bar)는 시장에서 일정량의 활동이 있을 때마다 바(bar)가 구성된다[13]. 이들은 시간 바(bar)와 비교해 상대적으로 변동성이 큰 편인데, 이는 거래횟수 및 거래량의 일별 변동성이 반영된 것으로 이해할 수 있다.

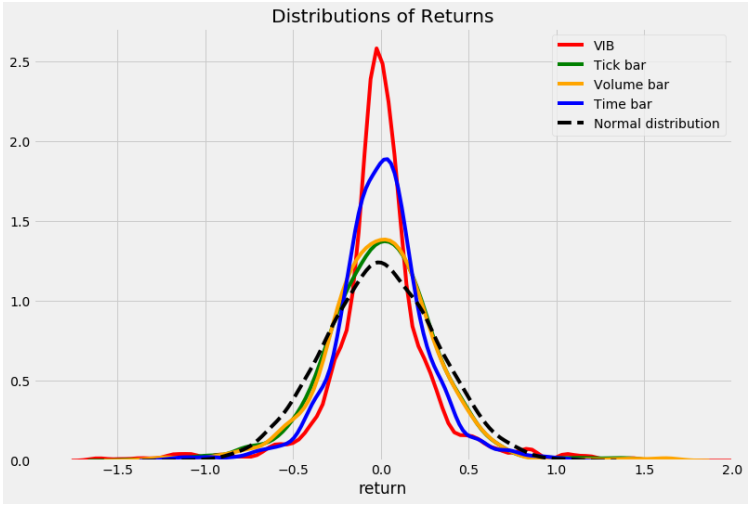
13) 앞서 언급한 거래량 및 틱 바(bar)의 숫자를 조정해도 생성되는 바(bar)의 숫자는 유사하였다.

<Figure 6> Number of Bars Produced Per Week



VIB는 모든 바(bar) 중에서 주별로 생성되는 바(bar) 개수의 변동성이 가장 크다. 앞서 언급했듯이 VIB는 정보를 가진 정보우월자들이 거래에 참여할 때마다 바(bar)를 구성하는 것으로 간주한다. 따라서 바(bar)들이 많이 구성된 때는 정보우월자들이 거래에 많이 나섰을 때이며, 반대로 바들이 많이 구성되지 않았을 때는 정보우월자들이 거래에 많이 나서지 않았던 때로 이해할 수 있다.

<Figure 7> Distributions of Returns



각 바(bar)로 구성된 Kospi200 지수 선물의 수익률 분포는 <Figure 7>에 나타냈다. 틱 바(bar)와 거래량 바(bar)의 수익률 분포가 정규분포에 가장 가깝게 나타났으며 시간 바(bar)의 수익률은 이들과 비교해 고첨도(leptokurtic)의 분포를 보인다. 이러한 차이는 시간 바(bar)를 사용할 경우 거래량의 일중 계절성에 노출되기 때문이다[14]. 마지막으로 VIB는 다른 바들에 비해 정규분포와 가장 다른 분포를 보인다. 이는 각 바의 길이가 일정하지 않은 VIB의 특성으로 인한 것으로 이해할 수 있다.[15]

Ⅴ. 실증 분석

1. 실증 방법론 개요

본 절에서는 종속변수와 예측변수는 어떻게 설정하는지, 각 머신러닝 모형의 하이퍼파라미터(hyperparameter)는 어떻게 선택하는지 및 예측치를 어떻게 도출하는지 등 머신러닝 모형을 이용한 실증 분석 과정을 전반적으로 설명한다.

먼저, 머신러닝 모형이 예측할 종속변수의 경우는 가격이 증가하는지 또는 감소하는지에 따라 0과 1의 값을 부여했다. 기존의 연구에서는 가격의 방향성을 측정하기 위해 각 바(bar)의 종가를 기준으로 사용한 것에 반해, 본 연구에서는 현재 바(bar)의 종가와 다음 바(bar)의 고가를 비교해 가격의 방향성을 판단했다.

종가 대신 고가를 사용한 이유는 첫째, 종가를 이용하는 경우 가격이 따르는 경로에 대한 정보가 누락 됨에 따라 가격의 방향성에 대한 측정 오차가 발생할 가능성이 존재하기 때문이다. 둘째, VIB(Volume imbalance bar)는 시장에서 사적정보를 가진 정보우월자들의 거래 활동이 있을 때마다 바(bar)를 구성하기 때문에 각 바(bar)의 종가는 해당 바(bar)가 구성되는 동안 발생한 정보에 영향을 받는다. 따라서 종가보다는 고가를 기준으로 가격의 방향성을 측정하는 것이 옳다고 판단된다. <Table 1>은 각각의 바(bar)마다 종속변수를 어떻게 설정했는지 요약한다.

---

14) 9시 개장 후 5분 동안의 거래량이 12시 후 5분 동안의 거래량보다 월등하게 높은 것처럼, 거래량에는 일중 패턴이 존재한다. Lopez de Prado(2018a)는 타임 바(bar)의 경우 거래량의 일중 계절성에 영향을 많이 받기 때문에 이에 대한 대안으로 틱 바(bar) 및 거래량 바(bar)를 사용할 것을 주장했다.

15) VIB의 경우 각 바의 길이가 일정하지 않은 것은 시장에 사적정보가 불규칙적으로 발생하는 것으로 이해할 수 있다.

<Table 1> Dependent Variable

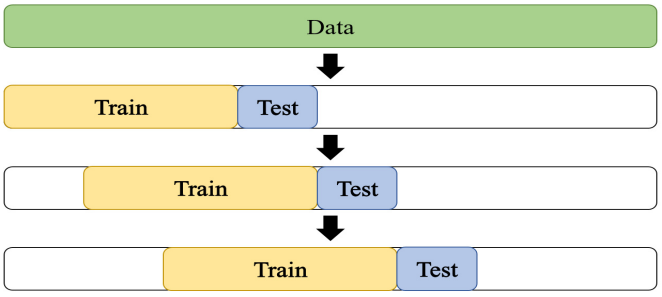
<Table 1> presents how the dependent variable is labeled. $\triangle P_t \equiv P_t^h - P_{t-1}^c$ denotes the price change between the close, high price of two adjacent bars. Then, y_t is labeled 1(0) if $\triangle P_t$ is bigger(smaller) than the threshold.

	Time bar	Tick bar	Volume bar	VIB
Threshold	$\triangle P_t \geq 0.35$	$\triangle P_t \geq 0.5$	$\triangle P_t \geq 0.5$	$\triangle P_t \geq 0.2$
$y=1$	52.5%	52%	53%	51.5%
$y=0$	47.5%	48%	47%	48.5%

종속변수는 다음 바(bar)의 고가가 해당 바(bar)의 종가보다 임계치 이상으로 상승하는 경우에는 1로 설정했으며, 그렇지 못한 경우에는 0으로 설정했다. 임계치는 바마다 각 범주에 속하는 데이터의 비율을 최대한 비슷하게 맞추도록 설정했다[16].

종속변수를 예측하는데 필요한 설명변수로는 각 바의 수익률, 거래량 및 주문 흐름이 사용되었다. 미시구조론은 거래량과 주문 흐름이 정보와 밀접한 관련이 있다고 주장한다. 앞서 언급했듯이 주문 흐름은 사적정보의 거래 행위와 밀접한 관련이 있으며, 거래량은 시장에 유입되는 정보의 양과 비례해 시장정보의 대리변수로 볼 수 있다[17]. 따라서 이러한 예측변수 설정은 미시구조론의 함의를 바탕으로 사적정보 및 시장정보를 이용해 고빈도 가격 움직임을 예측하는 것으로 이해할 수 있다.

<Figure 8> Structure of Rolling Window



16) 일반적으로 머신러닝의 분류 모형들은 각각의 범주(0 또는 1)에 속한 종속변수의 비율이 균형적이지 않으면 학습이 제대로 이루어지지 않는다. 따라서 임계치를 종속변수의 비율을 기준으로 설정하는 것은 이러한 문제를 방지하기 위한 것으로 이해할 수 있다.

17) 시장정보는 시장에서 발생하는 모든 정보를 지칭한다. 이는 뉴스가 발생했을 때 그에 반응하여 투자자들이 거래에 나서, 거래량이 증가하는 현상으로 이해할 수 있다.

다음으로, 각각의 머신러닝 모형들을 훈련하고 표본 외 예측 성능을 비교·분석하기 위해 본 연구에서는 롤링 윈도우(rolling window) 방식을 따른다. 롤링 윈도우는 금융 분야에서 예측모형의 성능을 측정할 때 빈번하게 사용되는 방법론으로, 그 구조를 도식화하면 <Figure 8>과 같다. 그림에서 볼 수 있듯이, 롤링 윈도우는 윈도우의 크기를 고정한 채 데이터를 따라 이동하면서 예측치를 도출하는 방법이다. 본 연구에서는 윈도우의 크기를 400바(bar)[18]로 구성했으며, 각각의 윈도우에 속한 훈련 데이터로 머신러닝 모형을 훈련한 후 그 다음 바(bar)의 가격 움직임을 예측한다.

각 예측모형의 표본 외 예측 성능을 평가할 지표로 재현율(recall), 정밀도(precision), 그리고 정확도(accuracy)를 사용한다. 이들은 예측모형의 예측 결과를 요약한 혼동행렬(confusion matrix)을 바탕으로 정의된다. 혼동행렬을 시각화하면 <Table 2>와 같다.

<Table 2> Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	tp	fp
	Negative	fn	tn

이를 바탕으로 재현율, 정밀도 및 정확도는 다음과 같이 정의된다.

$$Precision = \frac{tp}{tp+fp}$$

$$Recall = \frac{tp}{tp+fn}$$

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

혼동행렬에서 fn은 실제값이 참인데 거짓으로 예측을 잘못한 경우를 나타내는 제1종 오류(type Ⅰ error)를 나타내며, fp는 반대로 실제값이 거짓인데 참으로 잘못 예측하는 제2종 오류(type Ⅱ error)를 나타낸다. 따라서 정확도가 예측모형의 전반적인 예측 성능을 측정한다면, 재현율과 정밀도는 예측모형이 제1종 오류와 제2종 오류에 얼마나 강건한지 측정한다.

18) 표본 외 예측을 진행하는데 다양한 윈도우 크기를 적용했으나 분석결과 유의미한 차이는 발생하지 않아, 본 논문에서는 윈도우 크기를 400바로 했을 때의 결과를 보고한다.

마지막으로, SVM과 RF는 학습을 진행하기 전에 사용자가 하이퍼파라미터(Hyper-parameter)를 설정해야 한다. 본 연구에서는 그리드 탐색(grid search)을 통해 최적의 하이퍼파라미터 조합을 도출한다. 그리드 탐색은 사전적으로 하이퍼파라미터들이 취할 수 있는 값들을 미리 정해두고, 모든 가능한 조합 중 가장 좋은 성능을 보이는 하이퍼파라미터 조합을 도출한다. 롤링 윈도우의 각 윈도우마다 하이퍼파라미터를 설정하는 것이 가장 이상적이지만, 계산량이 너무 많아지는 문제가 있어 각각의 바마다 전 표본을 이용하여 하이퍼파라미터들을 설정했다. <Table 3>은 모형별로 설정해야 하는 하이퍼파라미터들의 종류 및 그들이 취할 수 있는 범위를 나타낸다.

<Table 3> Hyperparameters and Candidate Values

<Table 3> presents hyperparameters, and the candidate values of each hyperparameter. Each hyperparameter is chosen via the grid search method.

SVM		Random Forest	
C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	Number of Trees	[1, 10, 20, 30, 40, 50]
$Gamma$	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	Maximum depth	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
$Kernel$	Linear, RBF, Sigmoid	Predictors per node	[1, 2, 3]

2. 실증 분석 결과

(1) 바(bar)별 표본 외 예측 결과

각 바(bar)별 표본 외 예측 결과는 <Table 4>에 요약되어 있다. 틱 바(bar), 그리고 거래량 바(bar)는 0.5 내외의 예측 정확도를 기록해, 랜덤워크와 차이가 없어 예측력이 없는 것으로 분석되었다. 시간 바(bar)는 이들에 비해 예측 정확도가 약간 상승했지만 크게 유의미한 수준은 아닌 것으로 판단된다. 반면, VIB(Volume Imbalance bar)로 구성한 데이터는 최소 0.65의 예측 정확도를 기록해 가격의 움직임에 대한 예측력을 가진다는 것을 보여주고 있다.[19]

19) 본 논문에서는 서로 다른 모형들의 예측력을 비교하는데 통계적 검정을 사용하지 않지만, 다양한 설정에서 VIB의 예측 성능 지표들이 다른 바(bar)들에 비해 일관되게 높다는 점에서 결과가 유의미하다고 판단된다.

<Table 4> Out of Sample Forecast Results

<Table 4> presents the out-of-sample forecast results of Logistic regression, SVM and Random forest. Forecast ability of each model is evaluated by Precision, Recall and Accuracy.

Bar type	Model	Precision	Recall	Accuracy
Time bar	Logistic regression	0.54	0.69	0.54
	SVM	0.52	0.59	0.52
	Random forest	0.54	0.58	0.54
Tick bar	Logistic regression	0.51	0.76	0.49
	SVM	0.51	0.79	0.50
	Random forest	0.49	0.54	0.48
Volume bar	Logistic regression	0.53	0.70	0.52
	SVM	0.52	0.76	0.51
	Random forest	0.50	0.52	0.48
VIB	Logistic regression	0.67	0.73	0.65
	SVM	0.70	0.66	0.65
	Random forest	0.69	0.76	0.68

구체적으로 시간 바, 틱 바, 그리고 거래량 바의 예측 결과를 살펴보면 로지스틱 회귀모형과 SVM의 경우 재현율이 정밀도에 비해 높다. 즉, 제2종 오류가 상대적으로 크게 나타났다는 것을 의미한다. 이러한 결과는 주로 예측모형의 성능이 좋지 않고 종속변수가 특정 범주에 몰려있는 경우 나타나게 된다. 즉, 목적변수가 더 많이 몰려있는 방향으로 예측치를 도출하는 것으로 이해할 수 있다. <Table 5>를 살펴보면 바의 종류를 막론하고 가격이 오르는 경우가 떨어지는 경우보다 근소하게 많음을 확인할 수 있는데, 이러한 경향성이 결과에 반영된 것으로 판단된다.

<Table 5> Confusion Matrix of Time Bar-Logistic Regression Model

		Actual	
		Positive	Negative
Predicted	Positive	228	193
	Negative	102	126

Chinco et al.(2019)은 주가 움직임에 영향을 미치는 정보는 불규칙하게 발생하며 주가에 빠르게 반영된다고 분석한다. 이는 가격이 오를지, 아니면 떨어질지 예측할 때 임의의 시점에서 예측하기보다 정보가 발생했을 때, 그리고 그 정보가 가격에 반영되기 전에 예측해야 함을 시사한다. 즉, 위의 결과는 기존의 바(bar)들은 정보 발생 시점을

제대로 포착하지 못해 예측력이 낮지만 VIB는 시장에서 정보가 발생하는 시점들을 효과적으로 반영해 예측력이 높게 나타나는 것으로 해석된다.

다른 한편으로, VIB의 높은 예측 정확도는 바(bar)가 형성되는 시간(duration)이 불규칙하기 때문일 가능성도 존재한다. 가격의 변동성은 시간에 비례하기 때문에 고가를 기준으로 종속변수를 설정하는 본 연구의 설계상 바가 형성되는데 걸리는 시간이 길어질수록 가격이 오른 것으로 분류될 가능성이 높아진다. VIB의 경우도 바가 형성되는데 걸린 시간과 고가, 시가 차이의 상관계수가 0.57로 나타나 이러한 경향이 반영된 것을 확인할 수 있다. 만약 바가 형성되는 시간에 패턴이 존재한다면 <Table 5>에서 확인한 VIB의 높은 예측력은 정보를 효과적으로 반영했기 때문이라고 해석하는 것보다 예측모형이 이러한 패턴을 파악했기 때문이라고 해석해야 한다.

이러한 가능성을 확인하기 위해 목적변수를 다르게 설정하여 표본 외 예측을 다시 진행했다. 다음 바의 고가를 기준으로 가격이 올랐는지, 떨어졌는지 판단하는 대신 이번에는 각 바의 종가를 기준으로 가격의 방향성을 판단했다. 바가 형성되는 시간과 시가, 종가 차이의 상관계수가 0.02로 나타나 변동성 패턴의 영향을 분석하는데 종가를 기준으로 목적변수를 설정하는 것이 적절하다고 판단된다.

앞서 언급했듯이 VIB는 시장에서 정보가 발생할 때마다 바를 구성하기 때문에, 각 바의 종가는 서로 다른 정보를 담고 있어 이를 기준으로 가격의 움직임을 판단하는 것은 부적절할 수 있다. 하지만 새로운 정보가 발생해 바가 형성되기 전까지 가격은 직전 바에서 발생한 정보의 영향을 받기 때문에, 이것이 해당 바의 종가에도 다소 영향을 미칠 것으로 판단된다. 따라서 만약 VIB의 높은 예측력이 가격 움직임에 영향을 미치는 정보를 효과적으로 포착하기 때문이라면 종가를 기준으로 가격 움직임을 측정했을 때도 예측력의 차이가 나타나야 한다. 목적변수를 재설정해 진행한 표본 외 예측 결과는 <Table 6>에 요약했다.

먼저, 시간 바와 틱 바, 거래량 바의 경우 예측 정확도가 <Table 5>의 결과와 크게 달라지지 않은 것을 확인했다. 따라서 이들의 경우에는 가격의 방향성을 어떤 기준에 따라 판단하는지와 관계없이 가격의 움직임에 대한 예측력이 없다. VIB의 경우에는 앞선 결과에 비해 예측 정확도가 다소 감소했다. 하지만 로지스틱 회귀모형의 경우를 제외하면 SVM과 RF의 경우 0.6에 가까운 예측 정확도를 기록해 종가를 기준으로 가격의 움직임을 측정했을 때도 예측력을 지닌 것을 확인했다.

<Table 6> Out of Sample Forecast Results based on Close Prices

<Table 6> presents the out-of-sample forecast results of <Table 4> in an alternative setting where the dependent variable is labeled based on close prices.

Bar type	Model	Precision	Recall	Accuracy
	Logistic regression	0.50	0.83	0.49
Time bar	SVM	0.51	0.85	0.51
	Random forest	0.54	0.58	0.54
	Logistic regression	0.53	0.80	0.53
Tick bar	SVM	0.52	0.89	0.51
	Random forest	0.49	0.54	0.48
	Logistic regression	0.49	0.63	0.50
Volume bar	SVM	0.51	0.55	0.52
	Random forest	0.48	0.47	0.48
	Logistic regression	0.57	0.52	0.55
VIB	SVM	0.60	0.53	0.58
	Random forest	0.60	0.56	0.59

이러한 결과는 <Table 5>에서 확인한 VIB의 높은 예측력이 바가 형성되는 시간 때문이 아님을 확인해준다. 따라서 VIB가 보인 높은 예측 정확도는 그것이 불규칙적으로 발생하는 정보를 효과적으로 포착하기 때문이다. 마지막으로, <Table 6>에서 VIB의 예측 정확도가 앞선 결과에 비해 다소 감소한 것은 바가 형성될 때 새롭게 발생한 정보가 종가에 영향을 미치기 때문으로 판단된다.

(2) 예측모형에 따른 예측력 비교 · 분석

바의 종류에 따라 예측력이 다르게 나타난다는 결과는 예측모형을 선택하기에 앞서 데이터를 효과적으로 구성하는 것이 중요함을 의미하지만, 예측 정확도를 최대한 높이기 위해서는 적절한 예측모형을 선택하는 것 또한 중요하다. 따라서 본 절에서는 머신러닝 모형을 통해 앞서 확인한 VIB의 예측력을 더욱 높일 수 있는지 분석한다.

먼저, <Table 5>의 표본 외 예측 결과를 보면 VIB는 로지스틱 회귀모형과 SVM으로는 0.65, RF는 0.68의 예측 정확도를 기록했다. 즉, 머신러닝 모형의 예측 정확도가 로지스틱 회귀모형과 비교해 높지만 크게 유의미한 수준은 아니다.

그러나 머신러닝 모형의 예측 성능은 모형을 훈련하는데 사용되는 데이터의 양에 크게 영향을 받으며, 특히 변수들 사이에 비선형적 관계가 존재하는 경우에 이런 경향이 더욱 두드러지게 나타난다. 따라서 모형을 훈련할 때 사용하는 데이터의 크기를 다양하게 설정해 훈련 데이터의 양에 따라 모형의 예측 성능이 어떻게 변하는지를 살펴봐야 정확한

비교가 가능하다. 이를 위해 각 모형별로 윈도우의 크기를 점진적으로 늘려가며 표본 외 예측 성능이 어떻게 변하는지를 분석했다. 윈도우의 크기별 표본 외 예측 결과는 <Table 7>에 요약했다.

<Table 7> Out of sample forecast results based on rolling window size.

<Table 7> presents out of sample forecast results of different rolling window size. All results are based on VIB(Volume imbalance bar).

Window size	Model	Precision	Recall	Accuracy
	Logistic regression	0.62	0.62	0.61
50	SVM	0.60	0.56	0.59
	Random forest	0.63	0.63	0.62
	Logistic regression	0.62	0.65	0.62
100	SVM	0.63	0.56	0.60
	Random forest	0.64	0.63	0.62
	Logistic regression	0.65	0.67	0.63
200	SVM	0.66	0.64	0.64
	Random forest	0.66	0.67	0.64
	Logistic regression	0.67	0.69	0.65
300	SVM	0.69	0.64*	0.65
	Random forest	0.68	0.69	0.66
	Logistic regression	0.67	0.73	0.65
400	SVM	0.70	0.66	0.65
	Random forest	0.69	0.76	0.68
	Logistic regression	0.68	0.71	0.65
500	SVM	0.72	0.67	0.67
	Random forest	0.72	0.74	0.70
	Logistic regression	0.69	0.75	0.68
600	SVM	0.74	0.70	0.70
	Random forest	0.72	0.73	0.70
	Logistic regression	0.68	0.76	0.67
700	SVM	0.73	0.71	0.69
	Random forest	0.72	0.71	0.69
	Logistic regression	0.72	0.79	0.69
800	SVM	0.78	0.76	0.73
	Random forest	0.77	0.74	0.71
	Logistic regression	0.66	0.81	0.70
900	SVM	0.74	0.78	0.75
	Random forest	0.74	0.72	0.74

먼저, 세 모형 모두 윈도우 크기를 늘려갈수록 예측 정확도가 향상되는 것으로 나타났다. 이는 모형의 종류와 관계없이 훈련 데이터의 양이 많아질수록 예측력이 향상됨을 의미한다. 그러나 모형별로 훈련 데이터양의 증가에 따른 예측 정확도의 향상 정도는

다소 차이가 존재한다. 로지스틱 회귀모형의 경우 가장 작은 윈도우에서는 0.62, 그리고 가장 큰 윈도우에서는 0.7을 기록해 0.08의 예측 정확도 향상이 있었던 반면 RF와 SVM은 각각 0.12와 0.15의 향상을 기록해 훈련 데이터양의 증가에 따른 예측 성능 향상은 머신러닝 모형이 우월한 것으로 나타났다.

윈도우 크기별로 예측 결과를 살펴보면 300 이하에서는 세 모형이 거의 동일한 예측 정확도를 기록했다. 특히, 100 이하에서는 로지스틱 회귀모형의 예측 성능이 SVM보다 앞서는 것을 확인할 수 있다. 그러나 윈도우의 크기가 점점 커질수록 SVM과 RF 예측 성능이 로지스틱 회귀모형을 앞지르게 되며 윈도우 크기가 900에 이르러서는 정확도에서 4~5%의 차이가 나는 것을 확인할 수 있다.

이러한 결과는 훈련 데이터양이 늘어나면서 RF와 SVM은 목적변수의 비선형적 결정경계를 정확하게 학습하기 때문이지만, 로지스틱 회귀모형의 예측 성능 또한 크게 뒤떨어지지 않는 것으로 보아 비선형성의 정도가 크지는 않은 것으로 판단된다. 이는 이용 가능한 데이터의 양이 많지 않을 때 머신러닝 모형으로 예측하는 것은 효율적이지 않으며, 머신러닝 모형을 이용할 때 사전적으로 많은 데이터를 확보해야함을 의미한다.

Ⅵ. 결론

본 연구에서는 미시구조론의 함의를 이용해 바를 구성했을 때, 그리고 머신러닝 모형을 이용해 고빈도 가격 움직임을 더욱 정확하게 예측할 수 있는지를 분석했다. 분석 결과 미시구조론의 함의를 이용한 VIB(Volume Imbalance Bar)는 예측력을 지닌 것에 반해 기존에 많이 쓰이는 시간 바, 틱 바(bar) 및 거래량 바(bar)는 예측력이 없는 것으로 나타났다. 또한, 훈련 데이터양을 늘릴수록 머신러닝 모형의 예측 성능이 로지스틱 회귀모형보다 향상되는 것으로 나타났다.

특히 같은 원자료를 사용하더라도 어떤 기준에 따라 바를 구성하냐에 따라 머신러닝 모형의 예측력이 유의미하게 차이가 난다는 발견은 머신러닝 모형의 접목 및 새로운 예측변수의 도입에 치중한 기존 국내외 연구들에 새로운 경험적 사실을 추가했다. 미시구조론의 함의를 이용해 바를 구성했을 때 머신러닝 모형의 예측력이 가장 높아진다는 발견은 앞으로의 금융 머신러닝 연구에 있어서 기존의 미시구조론 연구들을 활용하는 것이 중요함을 시사한다.

하지만 데이터의 표본 기간이 상대적으로 짧아 바(bar) 구성법에 따른 예측력이 일별 이상의 저빈도에서도 유의미한 차이가 나는지를 분석하지 못한 점, 가장 기본적인 변수들인 수익률, 거래량 및 주문 흐름만을 예측변수로 했다는 점, 그리고 서로 다른 모형들의 예측력 차이를 비교하는데 통계적 검정을 사용하지 않은 점은 본 연구의 한계점으로 뽑힌다. 따라서 향후 연구들에서는 미시구조론 변수들을 예측변수로 사용했을 때 머신러닝 모형의 예측 성능에 어떤 영향을 끼치는지, 저빈도 등 다양한 빈도에서도 바(bar) 구성법에 따라 예측 성능이 유의미하게 차이가 나는지 등을 분석하는 것과 머신러닝 모형을 이용한 분석 결과에 적용 가능한 통계적 검정을 제시하는 것은 금융 연구에 머신러닝 방법론을 효율적으로 응용하는데 중요할 것으로 판단된다.

<참 고 문 헌>

1. 김동영, 박제원, 최재현, “SNS와 뉴스기사의 감성분석과 머신러닝을 이용한 주가예측 모형 비교 연구,”『한국IT서비스학회지』, 제13권 제3호, 2014, 221-233.
   (Translated in English) Kim, D. Y., J. W. Park, and J. H. Choi, “A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles,” *Journal of Information Technology Services* 13(3), 2014, 221-233.
2. 김명현, 이세호, 신동훈, “K-Nearest Neighbors(K-NN) 알고리즘을 통한 KOSPI200 선물지수 예측효과 연구,”『대한경영학회지』, 제28권 제10호, 2015, 2613-2633.
   (Translated in English) Kim, M. H., S. H. Lee, and D. H. Shin, “Predictability Test of K-Nearest Neighbors(K-NN) Algorithm: Application to the KOSPI 200 Futures,” *Korean Journal of Business Administration* 28(10), 2015, 2613-2633.
3. 김용석, 조성욱, “한국어 텍스트 분석과 적용: 머신러닝을 통한 증권발행신고서의 비정형화된 텍스트 분석,”『한국증권학회지』, 제48권 제2호, 2019, 215-235.
   (Translated in English) Kim, Y. S. and S. W. Joh, “Text Analysis for IPO firms in Korea: Analysis of Korean Texts in Registration Statements via Machine Learning,” *Korea Journal of Financial Studies* 48(2), 2019, 215-235.
4. 라윤선, 최홍식, 김선웅, “서포트 벡터 머신을 이용한 VKOSPI 일 중 변화 예측과 실제 옵션 매매에의 적용,”『지능정보연구』, 제22권 제4호, 2016, 177-192.
   (Translated in English) Ra, Y. S., H. S. Choi, and S. W. Kim, “VKOSPI Forecasting and Option Trading Application Using SVM,” *Journal of Intelligence and Information Systems* 22(4), 2016, 177-192.
5. 윤종문, “딥러닝 신경망을 이용한 신용카드 부도위험 예측의 효용성 분석,”『금융연구』, 제33권 제1호, 2019, 151-183.
   (Translated in English) Yoon, J. M., “Effectiveness Analysis of Credit Card Default Risk with Deep Learning Neural Network,” *Journal of Money and Finance* 33(1), 2019, 151-183.
6. Chinco, A., A. Clark-Joseph, and M. Ye, “Sparse Signals in the Cross-section of Returns,” *Journal of Finance* 74(1), 2019, 449-492.
7. Easley, D., N. M. Kiefer, and M. O’Hara, “One day in the life of a very common stock,” *Review of Financial Studies* 10, 1997, 805-835.

8. Easley, D., M. Lopez de Prado, and M. O'Hara, "The Volume Clock: Insights into the High-Frequency Paradigm," *The Journal of Portfolio Management* 39(1), 2012, 19-29.
9. Easley, D., M. Lopez de Prado, and M. O'Hara, "Discerning Information from Trade Data," *Journal of Financial Economics* 120(2), 2016, 269-286.
10. Easley, D., M. Lopez de Prado, and M. O'Hara, "Microstructure in the Machine Age," *Working Paper*, 2019.
11. Geron, A., "Hands-On Machine Learning with Scikit-Learn and TensorFlow," *O'Reilly*, 2017.
12. Gentzkow, M., B. Kelly, and M. Taddy, "Text as data," *NBER working paper* No. 23276, 2017.
13. Glosten, L. and P. Milgrom, "Bid, Ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of Financial Economics* 14(1), 1985, 71-100.
14. Gu, S., B. Kelly, and D. Xiu, "Empirical Asset Pricing via Machine Learning," *Chicago Booth Research Paper* No. 18-04, 2018.
15. James, G., D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R," *Springer*, 2017.
16. Laborda, R. and J. Laborda, "Can tree-structured classifiers add value to the investor?," *Finance Research Letters* 22, 2017, 211-226.
17. Jegadeesh, N. and D. Wu, "Word power: A new approach for content analysis," *Journal of Financial Economics* 110(3), 2013, 712-729.
18. Lopez de Prado, M., *Advances in Financial Machine Learning*, John Wiley & Sons, 2018a.
19. Lopez de Prado, M., "Ten Applications of Financial Machine Learning," Available at SSRN 3365271, 2018b.
20. Murphy, K., "Machine Learning: A Probabilistic Perspective," MIT Press, 2012.
21. O'Hara, M., "HIgh frequency market microstructure," *Journal of FInancial Economics* 116(2), 2015, 257-270.
22. Rossi, A., "Predicting Stock Market Returns with Machine Learning," Working paper, University of Maryland, 2018.
23. Samuel, A. L., "Some Studies in Machine Learning Using the Game of Checkers," IBM *Journal of Research and Development* 3, 1959, 210-229.
24. Vlastakis, N. and R. Markellos, "Information demand and stock market volatility," *Journal of Banking & Finance* 36(6), 2012, 1808-1821.

< Abstract >

Forecasting Ability of Machine Learning Algorithms using High-frequency Data: KOSPI200 Futures

Suk Jin Park* · Chae Shick Chung**

This paper investigates the effectiveness of machine learning algorithms and microstructure theory in predicting high frequency price movement. While accurately predicting future prices of financial assets has always been a major concern for the financial sector, recent developments in analytics tools and accessibility to new data have stimulated academics to pursue research.

There are two main ways in which machine learning algorithms are incorporated into financial research. The first is to increase the predictive power of models by adopting machine learning techniques that have not been used in previous studies (Yoon, 2019; Laborda and Laborda, 2017). Secondly, machine learning algorithms are also used to identify new predictive variables (Kim and Joh, 2019; Gentzkow et al., 2017).

On the other hand, there is very little discussion regarding the criteria to construct structured dataset from raw financial data. While microstructure theory argues that active informed traders leave characteristic footprints in market data, incorrectly structured data may fail to extract this information effectively. In this aspect, de Prado(2018a) suggested VIB(volume imbalance bar) based on implications of microstructure theory, which sample bars when informed traders are active.

Therefore, this study examines whether or not VIBs contain predictive information regarding future price movement. Using tick data of KOSPI 200 futures, we constructed VIBs and three standard bar types widely used by practitioners and academics: time bar, tick bar and volume bar. Then,

* First Author, Ph.D Student, School of Economics, Sogang University (+82-2-705-8179, E-mail: sukjinp1@gmail.com)

** Corresponding Author, Professor, School of Economics, Sogang University, (+82-2-705-8704, E-mail: cschung7@gmail.com)

we produced out of sample predictions of one-bar ahead price movement and compared prediction performances of different bar types. In order to test the effectiveness of machine learning algorithms, we used logistic regression as the benchmark and compared the prediction accuracy with SVM(support vector machine) and random forest, two machine learning algorithms widely applied in financial research.

The results of the analysis can be summarized as follows. First, the prediction accuracies of time bar, tick bar and volume bar were no better than a random walk. On the other hand, the prediction accuracy of data constructed with VIB was 65% at least, implying that it contains predictive information regarding future price movement. Chinko et al.(2019) argues that predictive information of returns are sparse and short-lived. Therefore it is better to predict price movements when an information event takes place, and before that information is reflected in the price than predicting them at a random time. This result shows that while VIB incorporates predictive information by effectively identifying the presence of informed traders, standard bar types fail to capture this information.

Second, as the size of training data increases, prediction accuracies of SVM and random forest outperform the prediction accuracy of logistic regression. While there is no significant difference when the training data is small, the gap widens with more training data and eventually resulting in a 5% difference in the biggest training data size. This result implies that machine learning algorithms may enhance prediction accuracy given large data.

This study shows that even though the same raw data is used, prediction accuracy of machine learning algorithms may differ depending on the criteria of how the structured dataset is constructed. Synchronizing bar constructions with information flows may capture predictive information. On the other hand, sampling bars based on chronological time may lead to a significant loss of information. Therefore, while the majority of financial machine learning research focus on model implementation and producing new predictor variables, this research shows that proper construction of structured data is also an important feature.

Keywords : High Frequency Data, Machine Learning, Microstructure, Bar, Private Information

JEL Classification : G14, G17