https://doi.org/10.15813/kmr.2022.23.1.005

자본시장 IT시스템 효율적 용량계획 모델: 심리지수 활용을 중심으로

Effective Capacity Planning of Capital Market IT System: Reflecting Sentiment Index

이국형 (Kukhyung Lee) 연세대학교 정보대학원[1)]
김미예 (Miyea Kim) 창원대학교 경영대학[2)]
박재영 (Jaeyoung Park) 연세대학교 정보대학원[3)]
김범수 (Beomsoo Kim) 연세대학교 정보대학원[4)]

〈 국문초록 〉

최근 COVID-19, 동학개미운동 등 투자환경의 변화로 시스템 처리 허용 수준을 상회하는 트랜잭션이 발생하고 이로 인해 전산장애가 자본시장에서 빈번하게 나타나고 있다. 자본시장 IT시스템들은 장애 영향도가 매우 큰 시스템들로서, 2020년에 예측하지 못한 큰 규모의 트랜잭션이 상당한 기간 유입되어 전산장애가 급증하였다. 다수의 기업들이 높은 수준의 IT시스템 용량계획 정책을 유지하고 있던 상황임에도 불구하고, 이를 상회하는 트랜잭션이 유입된 것은 용량계획에 대한 새로운 접근 방법이 필요함을 시사하고 있다. 이에 본 연구는 다양한 머신러닝 기법을 활용하여 자본시장 IT시스템 용량계획 모델들을 개발하고 성능을 비교 분석한다. 또한, 동학개미운동과 같이 예측하기 힘든 투자자의 행동을 반영할 수 있는 심리지수를 예측에 활용함으로써 용량계획 모델의 성능을 높인다. COVID-19 기간을 포함한 실증데이터를 이용하여 본 연구에서 개발한 용량계획 모델은 실무에서 활용 가능한 수준의 높은 성능과 안정성을 가질 수 있다. 본 연구는 기업의 비용 효율성과 IT시스템 용량 변경에 수반되는 운영상의 제약을 모두 고려한 최적의 파라미터를 제시하였는데, 이것은 자본시장 도메인에서 유용하게 사용될 수 있다. 또한, 본 연구는 투자자의 심리를 반영하는 심리지수가 IT 시스템 용량계획에 중요한 예측요인이 될 수 있는 것을 입증함으로써, 심리지수가 다양한 수요예측에 적극적으로 활용될 수 있음을 보여준다.

주제어: 효율적 용량계획, 심리지수, VKOSPI, 자본시장 IT 시스템, 지식경영

1) 제1저자, leekukhyung1@gmail.com
2) 제2저자, miyea.kim615@gmail.com
3) 제3저자, inyourface33@gmail.com
4) 교신저자, beomsookim@gmail.com

논문접수일 2022년 01월 26일 | 1차 수정 2022년 02월 25일 | 게재확정일 2022년 03월 02일

1. 서론

자본시장은 IT시스템의 장애 영향도가 매우 커서 용량계획이 매우 중요한 영역이다. IT시스템 용량계획 실패는 곧 시스템 중단으로 이어지며 이것은 자본시장에 큰 혼란을 줄 수 있다. 실제로, 최근 COVID-19, 동학개미운동 등 투자환경의 변화로 자본시장 IT시스템의 처리 허용 수준을 상회하는 트랜잭션 발생 및 시스템 중단이 빈번하게 보고되고 있다. 2020년에 예측하지 못한 큰 규모의 트랜잭션이 발생하여 1분기 기준으로 전 분기 대비 약 2배 이상(94건→191건)의 전산장애가 발생하였다(한국금융신문, 2020). 이는 동학개미 운동이라 불리는 역대 최대 규모의 개인투자자 트랜잭션 유입이 원인이었으며, 국내 증권사들의 IT 시스템은 이와 같은 막대한 트랜잭션 유입을 감당하기에 역부족이었던 것으로 분석된다(한국금융신문, 2020). 국내 자본시장 중에서도 트랜잭션의 유입건수 및 변동성이 증권시장에 비해 상대적으로 더 큰 파생상품 시장의 경우, COVID-19 확산 전인 2020년 1월 대비 2~3월 평균 호가(주문) 건 수가 59% 증가(2,223만 건→3,536만 건)하였고, 일별로는 2020년 3월 27일에 최대 호가 건 수를 경신하며 4,904만 건을 기록한 후, 다시 6월 18일 5,818만 건을 기록하며 역대 최대 수치를 기록하였다.

자본시장 IT시스템은 증권사의 원장시스템, 거래소의 매매체결시스템 등을 의미하는 것으로, 개인투자자의 호가를 입력받아 이를 처리하는 시스템이며, 이들 시스템의 트랜잭션 규모를 결정하는 핵심이 바로 호가이므로, 효율적인 IT시스템 용량계획에 있어서 일별 최대 호가 건수를 예측하는 것이 무엇보다 중요하다. 하지만, 현재 자본시장 대부분의 기업들은 상당한 수준의 과용량 IT시스템 용량계획 정책을 수립하고, 초기 구축 후에 새로운 시스템으로 교체할 때까지 장기간 사용하는 방식을 취하고 있다. 이와 같은 IT시스템 용량계획으로는, 동학개미운동 사례에서 알 수 있듯이, 갑작스런 트랜잭션 증가에 제대로 대처할 수 없다. 따라서 IT시스템 용량계획에 대한 새로운 접근방법이 필요한 상황이다.

이에 본 연구는 머신러닝 기법을 바탕으로 과거 자본시장 데이터를 활용하여 미래 특정기간 동안 IT시스템이 감당해야 할 최대 호가건수(즉, 최대 IT시스템 용량)를 예측하고자 한다. 또한, 본 연구는 IT시스템 용량 예측에 있어서 과거 자본시장 데이터 뿐 아니라 개인투자자의 행동 변화를 고려한다. 자본시장은 수많은 요인들에 의해 영향을 받으며 전통적인 금융경제학으로는 설명할 수 없는 다양한 현상들이 발생하는데, 이를 행동 재무학 관점에서 접근하는 연구가 많이 진행되고 있으며, 최근에 주목받고 있는 것이 바로 투자자 심리이다. 투자자 심리를 정량화한 것이 투자자 심리지수인데, 자본시장 IT시스템에 유입되는 트랜잭션은 이를 이용하는 투자자의 심리에 따라 변화하므로, IT시스템 용량계획에 투자자 심리지수를 활용할 경우 더욱 정확한 용량계획 모델 수립이 가능할 것이다.

요약하면, 본 연구의 목적은 자본시장의 효율적인 IT시스템 활용을 위해 자본시장 시계열 데이터와 투자자 심리지수를 바탕으로 일별 최대 호가 건 수를 예측하는 IT시스템 용량 계획 모델을 제안하고 다양한 머신러닝 기법을 활용하여 모델 성능을 비교 분석한 후에 최적의 IT시스템 용량 계획 모델을 제안하는 것이다.

2. 이론적 배경 및 선행연구

2.1. 용량계획

용량계획(capacity planning)은 개략적인 IT시스템 아키텍처와 응용 업무를 기반으로 IT시스템에 요구되는

성능 요구사항과 성능을 결정하는 것이다(Menasce & Almeida, 1998). 이러한 용량 계획은 응용프로그램에 접근하는 사용자의 수, 서버시스템에 접속하는 동시 접속자 수, 서버시스템에 의해서 수행되어야 하는 피크율, 서버시스템에 필요한 여유율 등을 다룬다(나종회, 최광돈, 2004). 이는 성능 요구사항을 CPU의 형태나 수, 메모리의 형태나 크기, 디스크의 크기 등 IT시스템의 요구사항으로 변환하는 용량산정(capacity sizing)과는 다른 것으로, IT시스템을 사용하는 사용자의 규모나 사용자로 인해 발생하는 트랜잭션의 규모를 의미한다.

최근 COVID-19, 동학개미운동 등 투자환경의 변화로 자본시장 IT시스템의 처리 허용 수준을 상회하는 트랜잭션 발생 및 전산장애가 빈번하게 보고되고 있다. 증권업계를 포함한 자본시장은 IT시스템의 장애 영향도가 매우 커서 용량계획이 매우 중요한 영역으로, 2020년 예측하지 못한 큰 규모의 트랜잭션이 발생하여 1분기 기준으로 전분기 대비 약 2배 이상(94건 → 191건)의 전산장애가 발생하였다(한국금융신문, 2020). 국내 자본시장 중에서도 트랜잭션의 유입건수 및 변동성이 증권시장에 비해 상대적으로 더 큰 파생상품 시장의 경우, COVID-19 확산 전인 2020년 1월 대비 2~3월 평균 호가(주문)건수가 59% 증가(2,223만건 → 3,536만건)하였고, 일별로는 2020년 3월 27일에 최대 호가건수를 경신하며 4,904만건을 기록한 후, 다시 6월 18일 5,818만건을 기록하며 역대 최대 수치를 기록하였다. 이는 동학개미 운동이라 불리는 역대 최대 규모의 개인투자자 트랜잭션 유입이 원인이었으나, 국내 증권사들의 IT시스템은 이와 같은 막대한 트랜잭션 유입을 감당하기에 역부족이었던 것으로 분석된다(한국금융신문, 2020). COVID-19의 확산 및 개인투자자의 유례없는 자본시장 참여는 예측하기 힘든 불확실성이며, 이는 기존 IT시스템의 용량계획으로는 대처할 수 없다는 것을 보여주고 있는 것으로, IT시스템 용량계획의 새로운 접근방법이 필요함을 시사하고 있다.

2.2. 수요 예측 선행연구

IT시스템 관점에서의 용량계획은 넓은 관점에서 사용자에 대한 수요 예측으로도 해석될 수 있다. 수요를 정확히 예측하는 것은 기업이 전략을 수립하는데 있어서 매우 중요한 요소이며, 더 나아가 사람들의 행동이 예측하기 힘들거나 사회적 여파가 매우 큰 서비스에 대한 수요 예측은 더욱 중요해진다. 이러한 수요를 더욱 더 정확하게 예측하기 위해 업계 및 학계 모두 머신러닝을 활용하여 활발히 연구를 진행하고 있다. 예를 들어, 대도시의 교통난 해결 및 스마트시티에 대한 교통체계 마련을 위해 자전거, 택시 등 교통 수요 예측(Du et al., 2020; Xiao et al., 2020), 기후 변화 대응을 위한 빌딩·도시·도(州)의 장·단기 전력 및 급수 수요 예측(Muralitharan & Vishnuvarthanc, 2018; Guo et al., 2018), 제품 판매량 및 기업의 공급 관리망(supply chain management) 수요 예측(Seyedan & Mafakheri, 2020; Noh et al, 2020), 전자 상거래 수요 예측(Tugay & Oduducu, 2020) 등 머신러닝을 활용한 수요 예측은 교통·전력·제조·소매 등 특정 산업에 국한하지 않고 전방위적으로 매우 중요하게 여겨지고 있다. 금융 분야에서도 머신 러닝을 활용한 예측 연구가 일부 진행되고 있는데, 주식의 수익을 예측(Chen et al., 2015)하거나 이를 위한 주가의 방향성 예측(Nelson et al., 2017; Weng et al., 2017) 등 시장의 움직임 및 패턴을 찾아내는 것에 초점을 맞추고 있다. 또한 외환시장 환율예측 연구(임현욱 등, 2021), 최근에는 암호화폐 수익률 변동예측(김은미, 2021), 암호화폐 가격 예측(원종관, 홍태호, 2021)에 대한 연구가 진행되었다. 이처럼 기존 연구는 시장에 참여하는 수요를 예측하는 연구에는 관심을 거의 가지지 않았는데, 최근 상당한 규모의 IT

시스템에도 불구하고 시스템 장애가 발생한 사례에서 알 수 있듯이, 이제는 금융시장과 자본시장에서의 수요 예측이 중요한 이슈로 부각되었다.

IT시스템 관점에서의 수요예측 연구도 일부 이루어졌다. 네트워크 트래픽 예측(Mozo & Gómez-Canaval, 2018), 클라우드 워크로드 예측(Kumar et al., 2020; Yu et al., 2018) 등과 같은 분야에서 머신러닝이 활용되고 있다. 하지만, 이러한 연구는 대다수 클라우드나 클라우드를 활용한 데이터 센터의 이용률 예측 및 최적화가 대부분이며, 자본시장 IT시스템 용량예측 연구는 거의 찾아볼 수 없다. 자본시장 IT시스템 용량예측 실패는 사회에 큰 혼란을 야기하는 만큼 효율적으로 정확하게 예측하는 것이 요구된다. 따라서 본 연구는 기존 연구에서 거의 관심을 가지지 않았던 하지만 중요한 문제인 자본시장 IT시스템 용량계획 모델을 개발하고자 하며, 특히, 용량계획에 있어서(즉 호가 건수를 예측하는데 있어서) 투자자 심리지수를 고려하고자 한다.

2.3. 투자자 심리지수

심리적 영향과 편견이 투자자의 재무활동에 영향을 미치는 것을 연구하는 학문 분야인 행동 재무학에서 연구되는 다양한 구성 개념들 중에서 최근에 크게 주목받고 있는 것 중에 하나가 투자자 심리(investor sentiment)이다. 심리는 시장참여자의 태도를 반영하는 매우 독특하고 효율적인 변수(Xing et al., 2018)이며, 투자자가 결정을 내릴 때는 그들의 감정 또는 마음의 상태가 그 결정에 영향을 미치는데(Lucey & Dowling, 2005), 많은 연구에서 이를 폭넓게 투자자 심리로 언급하고 있다. 투자자 심리의 개념에 대해서는 그 대상과 범위에 따라 제각각 다양하게 정의하고 있지만 미래 주식시장 활동에 대한 투자자의 낙관주의/비관주의(Baker & Wurgler, 2006)라는 정의가 가장 폭넓게 받아들여지고 있다.

투자자의 심리를 정량화하기 위한 노력은 오래전부터 지속적으로 연구되어 왔는데, 최근에 정량화된 투자자 심리 지수와 실제 자본시장 현상들의 상관관계가 많이 밝혀지면서 급속도로 성장하였다. 투자자 심리를 정량화한 투자자 심리지수(investor sentiment index)는 데이터를 수집 및 측정하는 방식에 따라 크게 3가지 유형으로 분류된다(López- Cabarcos et al., 2019). 첫 번째 유형은 설문 수집 방법에 따라 생성되는 투자자 심리 지수로 미시간 대학의 Survey Research Center에서 월별로 생성하는 Michigan Consumer Sentiment Index (MCSI)가 이에 해당한다(Qiu & Welch, 2004). 두 번째는 자본시장에서 생성되는 변수 또는 변수들의 조합으로 신규 생성한 프록시(proxy) 지수이다. Baker Wurgler index (Baker & Wurgler, 2006), EURSent Index(Reis & Pinho, 2020), VKOSPI Sample Entropy(Cho, 2016) 등이 이에 해당한다. 마지막 유형은 인터넷 내 SNS 등을 통한 사용자들의 상호작용 정보를 수집하여 생성한 마이크로블로깅(micro-blogging) 지수이다. New York Times와 같은 온라인 신문(Buckman et al., 2020), 페이스북(Siganos et al., 2017) 등 웹 트랜잭션 및 SNS 정보를 이용하여 생성한 심리지수가 이에 해당한다. 이러한 심리지수는 포트폴리오 자산 평가 및 배분(Xing et al., 2018; Cho, 2016)에 활용되는가 하면, 시장의 변동성(Liu, 2015; Liang et al., 2020)이나 주가의 방향성을 예측(Oh & Sheng, 2011; Makrehchi et al., 2013)하거나, 직접적으로 주식 수익률과 인과관계를 입증(Reis & Pinho, 2020; Han et al., 2011; Bagchi et al., 2013; Lee & Ryu, 2014)하는 연구에 활용되고 있다.

한편, Pedro Manuel Nogueira Reis와 Carlos Pinho는 유럽의 공포지수라 불리는 Euro Stoxx 50 변동성 지수(VSTOXX)와 금 가격, 독일국채 수익률 등을 조합하여 EURSent Index를 신규로 생성하였는데, 이 때 사

용한 VSTOXX와 유사한 것이 바로 본 연구에서 사용하고자 하는 KOSPI 200 변동성 지수(VKOSPI)이다. VKOSPI는 한국의 유가증권시장(KOSPI) 200개 종목을 기초자산으로 하는 KOSPI200 옵션 가격을 이용하여 미래 변동성에 대한 시장의 기대치를 나타내는 실시간 지수(KRX, 2009)로, 한국거래소(KRX)가 발표하는 지수이다. VKOSPI는 현재 시장 참여자들이 향후 시장의 변동성을 어느 정도로 보고 있는지를 나타내는 수치로, 주식시장의 변동성이 클 것이라 예상하는 투자자가 많아질수록 지수가 증가하고 반대로 변동성이 작을 것이라 예상할수록 지수가 감소한다. VKOPSI가 상승할 때 자본시장에 폭락이 자주 나타나 공포지수라고도 불린다.

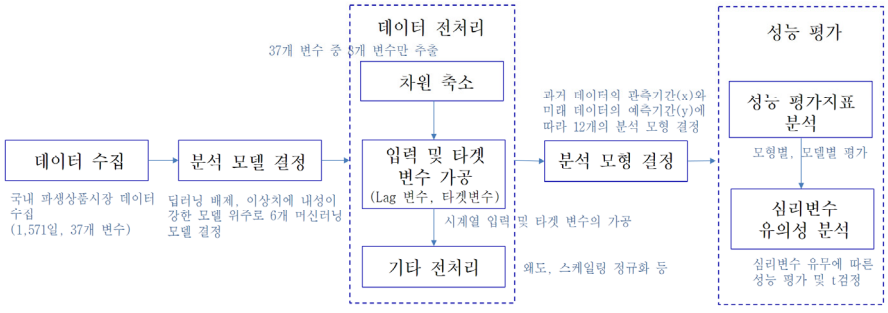
심리지수는 행동 재무학 분야에서 여러 가지 목적에 따라 적극적으로 활용되고 있다. 특히 자본시장에서 기존에 설명될 수 없던 많은 현상들이 심리지수와 인과관계가 있음이 실증적으로 입증됨에 따라 심리지수에 대한 관심이 증대되고 있다. 심리는 시장참여자의 태도를 반영하는 매우 독특하고 효율적인 변수(Xing et al., 2018)이며, 투자자가 결정을 내릴 때는 그들의 감정 또는 마음의 상태가 그 결정에 영향을 미치는데(Lucey & Dowling, 2005), 많은 연구에서 이를 폭넓게 투자자 심리로 언급하고 있다. 이전 연구를 보면, 투자자 심리지수가 주가의 급등 또는 급락이나 시장의 극단적인 움직임과도 연관이 있음이 입증되었다(Piccoli & Chaudhury, 2018). 이것은 COVID-19와 같은 외부 충격에 따라 발생한 투자자들의 적극적인 주식투자 활동이 심리지수와 연관이 있음을 시사하고 있다. 따라서, 투자자의 실제 행동인 호가 건 수를 예측하는데 있어서 투자자의 심리상태를 보여주는 공포지수(VKOPSI)가 중요한 역할을 할 수 있다.

3. 연구 설계

3.1. 연구절차

본 연구는 과거 특정기간의 데이터를 이용하여 미래 특정기간 동안 IT시스템이 감당해야 할 최대 호가건수를 예측한다. 다시 말해, 특정 기준일 시점 기준으로 과거 x일(관측기간)의 데이터를 관측하고, 이를 이용하여 기준일 이후 y일(예측기간)동안 필요할 호가건수의 최대치를 예측하는 모델을 개발한다. 이 과정에서 투자자의 심리를 반영하는 심리지수도 포함하여 모델링함으로써 의미 있는 성능 강화가 나타나는지도 확인한다.

본 연구에서는 Python 3.7 및 Scikit-learn 0.24.1 라이브러리를 이용하여 전체적인 분석을 진행하였으며, 연구 절차는 크게 데이터 수집, 분석 모델 결정, 데이터 전처리, 분석 모형 결정, 성능 평가의 순으로 구성되며 아래 <그림 1>에 도식화하였다.



〈그림 1〉 연구절차

3.2. 데이터 수집

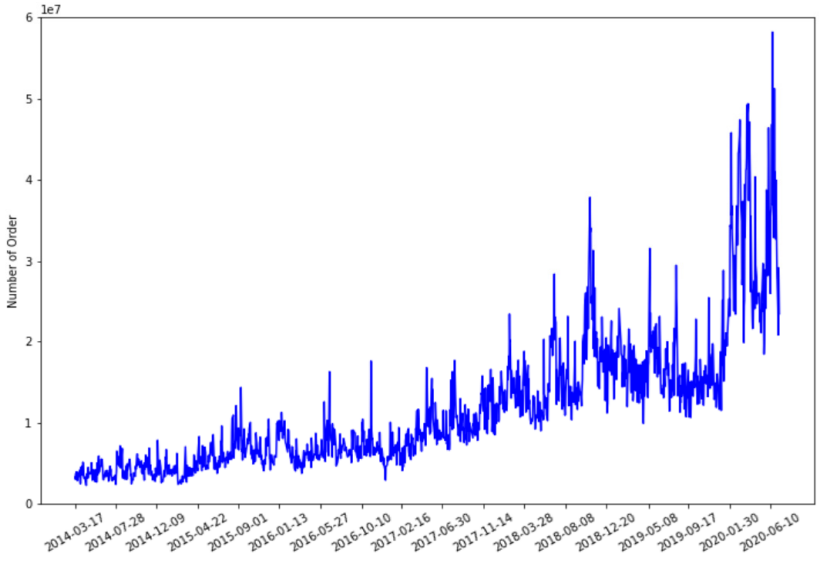
본 연구는 COVID-19 기간을 포함한 실증 분석을 위해 국내 자본시장 파생상품 시장을 대상으로 데이터를 수집하였다. 자본시장은 가격이라는 단일 측정지표가 있고, 다양한 이론에 근거하여 정형화되어 운영되고 있으며, 특히 파생상품 시장은 사람들의 심리에 매우 민감한 시장이므로 본 연구에 적절하다고 할 수 있다. 데이터는 자본시장의 모든 데이터가 집약되는 한국거래소의 정보데이터시스템(http://data.krx.co.kr)을 통해 이차자료 형태로 제공받았으며, 일별 종가 기준의 데이터이다. 데이터는 2014년 3월 3일부터 2020년 7월 24일까지 파생상품 시장의 주요한 지표들로 구성되어 있으며, 영업일 기준 1571개의 데이터가 존재한다. 이들 지표 중 본 연구에서 예측하고자 하는 일 별 호가 건 수의 추이는 <그림 2>와 같다.

3.3. 분석모델

어떤 머신러닝 분석 모델을 활용하여 모델링할 것인지 고려하기 위해서는 데이터의 양고 특성을 고려하여야 한다. 본 연구는 약 7년 치 데이터를 마련하였음에도 불구하고 영업일 기준 1571개 수준으로 데이터가 많지 않고, 미래의 특정기간 동안의 최대 호가건수를 예측하는 것으로 연구의 목적상 이상치도 매우 많다. 특히, 이상치의 경우, 학습 데이터의 과적합 발생 가능성을 높여, 연구 모델의 안정성을 저해할 수 있다고 알려져 있다(Aggarwal, C., 2017). 따라서 본 연구는 이러한 상황을 고려하여, 본 연구에서는 딥러닝이 아닌 전통적인 머신러닝 모델들을 활용하였으며, 이들 중에서도 이상치에 대해 높은 안정성을 보이는 Lasso, Ridge 선형회귀모델과 회귀트리모델의 앙상블 모델인 RandomForest와 XGBoost를 활용하였다.

먼저 선형회귀모델인 Ridge, Lasso는 학습 데이터를 이용하여 학습을 진행할 때 경사하강법을 이용하여 Mean Squared Error(평균제곱오차, MSE)가 최소화되도록 목적함수를 설정하는데, Lasso와 Ridge는 MSE에 L1, L2 페널티를 추가하여 목적함수를 변경함으로서 과적합을 어느 정도 탈피할 수 있는 모델이다.

다음으로 RandomForest와 XGBoost 모델은 앙상블 계열의 모델이다. 앙상블이라는 것은 하나의 모델을



〈그림 2〉 국내 파생상품시장 호가건수 추이(2014.3.3.~2020.7.24.)

이용할 경우, 학습 데이터를 너무 잘 학습하는 과적합이 발생할 수 있기 때문에, 여러 개의 모델들을 사용하여 종합적으로 판단하는 모델을 일컫는다. 앙상블 계열은 크게 배깅과 부스팅으로 나뉘는데, 먼저 배깅(bagging)이란, 원본 데이터에서 임의로 데이터를 추출하여 계속 추가하는 복원 추출을 통해 데이터와 변수를 무작위로 정해진 개수만큼 뽑은 뒤, 이를 독립적인 하나의 모델에 학습하고, 또 다른 임의의 데이터 셋을 구성하는 과정을 반복하여 학습한 여러 개의 독립적인 모델을 종합적으로 판단하여 예측하는 방법이다. RandomForest는 이런 배깅을 이용한 분석 모델로 기본 독립적인 모델로는 Decision Tree를 이용하며, 복원 추출된 서로 다른 여러 개의 데이터 셋으로 학습된 여러 개의 독립적인 Decision Tree를 이용하여 Test set을 예측한 후, 그 결과값을 평균하는 분석모델이다. 반면에 부스팅(boosting)은, 배깅과 다르게 여러 개의 모델이 독립적이지 않다. 먼저 복원 추출을 통해 마련된 임의의 학습 데이터를 학습한 하나의 모델을 이용하여 실제 테스트 데이터를 이용하여 예측한다. 이후 다시 복원 추출을 통해 마련된 다른 임의의 학습 데이터를 이용하여 다른 모델을 학습할 때, 직전 모델에서 잘못 예측된 데이터에 가중치를 부여함으로써, 잘못된 예측에 대해서는 더욱 더 큰 패널티를 제공해주는 방식이다. 즉, 여러 개의 모델을 순차적으로 학습 및 예측하여, 오류를 개선해나가는 특징이 있는 모델이다. XGBoost는 부스팅 모델의 하나인 GBM(Gradient Boosting Model)에서 발전된 모델로, GBM은 잘못된 예측에 대한 가중치를 업데이트할 때 경사하강법을 이용하는 방식이다. XGBoost는 이러한 GBM의 성능을 높임과 동시에 Ridge, Lasso와 같이 과적합에 강한 내성을 주기 위해 목적함수에 정규화항을 추가한 모델이다.

마지막으로 이상치에 대한 영향도를 비교하기 위하여 이상치에 대한 처리 메커니즘이 없는 Decision Tree와 최근 많은 연구에서 높은 성능과 안정성을 인정받는 Support Vector Machine(SVM)을 포함하였다. SVM은 주로 판별식에 사용되는 분석 모델이며 표본 공간 내 클래스간 분류를 위해 선형회귀모델과 유사하게 벡터를 설정하되, 벡터의 두께를 나타내는 마진(margin)을 주어 분류하는 모델로, 마진(margin)은 두 클래스 집단 간의 거리를 의미한다. SVM을 회귀모델로 사용하기 위해 소개된 모델이 바로 Support Vector Regression (SVR)이며 최대한 많은 데이터를 마진(margin)안에 포함하고, 마진(margin)밖에 존재하는 데이터에 대해 에러를 줄여 그 에러를 최소화하는 방향으로 회귀를 진행하는 모델이다. 즉, 일반적인 회귀모델은 해당 데이터를 설명할 수 있는 선에 대해 에러를 계산하는 방식이라면, SVR은 마진(margin) 밖에 존재하는 데이터들에 대해서만 에러를 계산하는 방식이다.

머신러닝 모델들은 다양한 파라미터의 값에 따라 결과가 달리 나타나거나 과적합이 발생할 수 있다. 따라서, 본 연구에서는 각 머신러닝 적용 시 과적합을 방지하기 위하여 교차검증을 적용하고, 하이퍼파라미터를 설정하고 GridSearchCV를 사용하여 최적의 파라미터를 찾아 분석하였다.

3.4. 데이터 전처리

다변량 시계열 데이터를 분석하기 위해서는 변수별 시계열 트랜드를 반영할 수 있도록 각 변수별로 lag변수(d-1, d-2 등)를 생성하여 이를 새로운 변수로 추가하는 것이 필요하다. 그러나, lag 변수는 시계열 트랜드를 관측하는 x의 값이 커질수록 급속도로 증가하여, 해당 모델이 학습 데이터에 과적합될 확률이 매우 높다. 본 연구에 사용되는 일별 데이터도 약 7년 치의 데이터임에도 불구하고, 1571건으로 적은 편이므로,

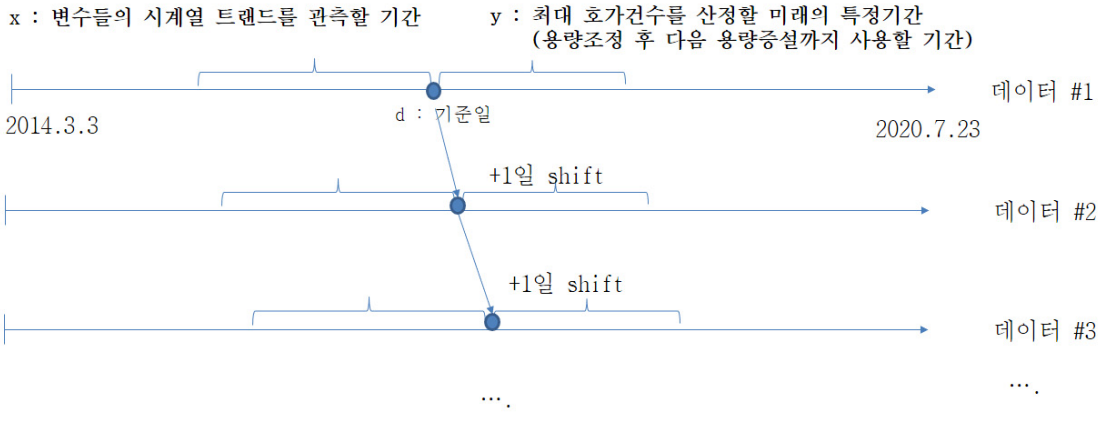
변수가 많아질수록 그 위험성이 높아진다. 따라서 본 연구에서는 타겟변수에 가장 영향력이 큰 최소한의 변수들만을 선별하여 lag 변수를 생성하는 것이 중요하다. 이 때 타겟변수라 함은 y기간 동안의 최대 호가건수를 의미하며 이는 호가건수 변수를 가공함으로써 결정된다.

최소한의 핵심변수들을 선별하기 위하여 먼저 수집된 37개 변수 간 상관계수를 구한 후, 0.9 이상의 상관계수를 보이는 변수들 쌍에서 타겟변수와 상관계수가 가장 높은 변수를 유지하고 나머지 변수 11개를 삭제하였다. 이후, 옵션거래량, 선물거래량 등과 같이 동일한 유형으로 분류될 수 있는 변수들을 그룹별로 묶은 후, 각 그룹 내에서 타겟변수와 가장 상관계수가 높은 변수 11개를 추출하였다. 마지막으로, 타겟변수와의 상관계수가 0.01 이하로 매우 낮은 변수들을 삭제하여 최종적으로 8개 변수를 선택하였다. 최초 수집된 37개 변수와 최종 선정된 8개 입력변수는 각각 <부록>, <표 1>에 기술되어 있으며, lag 변수의 생성 및 타겟변수의 생성을 도식화하면 <그림 3>과 같다.

<표 1>에서 보듯이, 본 연구에서 사용되는 입력변수들이 각각 가질 수 있는 값들의 범주는 모두 상이하다. 가령 호가건수는 최대 5,820만 건에 육박하나, VKOSPI지수는 최대 69.24이다. 이러한 변수별 스케일 차이가 크게 나타날 경우, 머신러닝 알고리즘은 범주가 큰 변수들에 더욱 민감하게 반응하게 된다. 따라서 본 연구에서는 이를 방지하기 위해 모든 변수들의

〈표 1〉 입력변수 기술 통계량

입력변수	Mean	*SD*	Min	Max
호가건수	11,647,600	8,342,356	2,256,015	58,188,050
체결건수	1,644,038	6,991,982	3,656,880	5,890,272
거래량(선물)	2,271,546	1,632,755	4,683,670	1,398,524
거래종목수(옵션)	515	190	74	914
투자자수	47,417	5,613	31,670	73,942
거래증거금필요액	4,982,515	1,291,644	2,846,330	1,108,360
K200현물가격(종가)	272.05	26.85	199.28	338.83
VKOSPI	15.27	6.19	9.72	69.24



〈그림 3〉 입력 및 타겟변수 생성

범주를 0에서 1사이로 변환하는 Min-Max 정규화를 수행하였다.

이 외에도, 결측치 확인과 데이터 타입 변환 작업을 수행하였으며, 최종 전처리 과정을 거친 후, 학습·검증·테스트 데이터를 7:1:2의 비율로 나누어 분석 모델에 사용할 데이터를 준비하였다.

3.5. 분석모형

<그림 3>에서 나타나듯, x는 예측을 위해 사용할 과거 관측기간이며, y는 d일 기준으로 타겟변수인 최대 호가건수를 결정할 기간이다. 이러한 x와 y의 값은 매우 중요한데, 이는 이들 x, y값이 성능 뿐 아니라, 기업들의 실무적 제약 등 비즈니스적 의미가 있기 때문이다. 기업은 특정 시점 또는 주기적으로 미래에 IT시스템이 필요할 용량을 산정한 후, 이를 기반으로 용량 변경을 수행하는데, y를 작게 설정하는 것은 짧은 주기로 용량 변경을 수행한다는 의미로, 잉여용량이 줄어들어 비용효율성은 개선될 것이나, 반대로 시스템 변경을 위한 테스트 수행 및 잦은 시스템 변경으로 인한 장애 위험이 증가한다. 반대로 y를 길게 설정하는 것은 시스템 변경 주기를 길게 가져간다는 의미로, 장애위험은 감소하나 비용효율성이 저하될 것이다. x의 경우, 얼마 동안의 과거기간을 관측해야만 가장 예측력이 좋은지를 확인하는 것으로, 이 값이 크면 클수록 많은 변수를 예측에 활용한다는 뜻이므로 예측 성능이 개선될 수 있으나, 일정 수준 이상의 너무 많은 변수를 활용할 경우, 핵심이 되는 변수의 영향력이 희석되어 오히려 성능이 감소하는 구간이 발생할 수 있다. 이에 더하여, lag 변수가 기하급수적으로 늘어 과적합이 발생하여 모델의 안정성 또한 저해될 수 있다. 반대로 x가 작을 경우, 짧은 기간만을 관측하여 미래를 예측하므로 예측 성능이 낮아질 수 있다. 따라서, x는 값을 증가해가면서 성능 및 과적합 여부를 관측하여 적정한 구간을 파악하고, y는 기업별 운영 상황에 부합하도록 비용효율성 및 변경·장애 위험성 간의 균형있는 최적의 값을 선택하는 것이 매우 중요하다.

이러한 모든 점을 고려하여 본 연구는 <표 2>의 모형들을 수립하였다. x는 120일(6개월) 이상을 설정할 경우, 성능이 급격히 감소하여 120일 이내에서 20일

〈표 2〉 분석모형

모형	x	y	lag 변수 개수	VKOSPI 포함
모형 1	20일	20일	160	O
모형 2	40일	20일	320	O
모형 3	40일	40일	320	O
모형 4	60일	20일	480	O
모형 5	60일	40일	480	O
모형 6	60일	60일	480	O
모형 7	80일	20일	640	O
모형 8	80일	40일	640	O
모형 9–1	80일	60일	640	O
모형 9–2	80일	60일	560	X
모형 10	120일	20일	960	O
모형 11	120일	40일	960	O
모형 12	120일	60일	960	O

간격으로 설정하였다. y는 60일(3개월)을 초과할 경우 비용효율성이 낮은데다가, 증권사시스템을 포함하여 전체 자본시장IT시스템의 실제 변경 주기가 3개월 인 것을 감안하여 60일 이내로 설정하였다. <표 2>의 모형에서 모형 1, 2, 3, 4, 5, 6, 7, 8, 9-1, 10, 11, 12는 최적의 x와 y값과 가장 우수한 성능을 보여주는 머신러닝 모델을 찾기 위한 것이며, 모형 9-1, 9-2는 가장 성능이 좋은 모형에서의 심리지수인 VKOSPI의 유의성을 확인하기 위한 것이다.

4. 연구결과

4.1. 평가지표

본 연구에서는 예측값과 실제값의 모든 차이의 절댓값 합을 의미하는 MAE(Mean Absolute Error)와 이를 백분율로 변환한 MAPE(Mean Absolute Percentage Error)를 평가지표로 활용한다. 또한 본 연구에서는 추가적으로 MPE(Mean Percentage Error)를 활용한다. MPE는 모델이 과소 예측 또는 과대 예측 여부를 확인하기 위해 사용하는 것으로, 음수일 경우 과대 예측, 양수일 경우 과소 예측을 의미한다. 본 연구는 시스템의 용량계획을 예측하는 것으로 과소 예측을 회피해야 한다. 즉 최소 수준의 과대 예측을 하여야만 의미있는 모델이 될 수 있기 때문이다. 각 평가지표의 계산방식은 <표 3>과 같다.

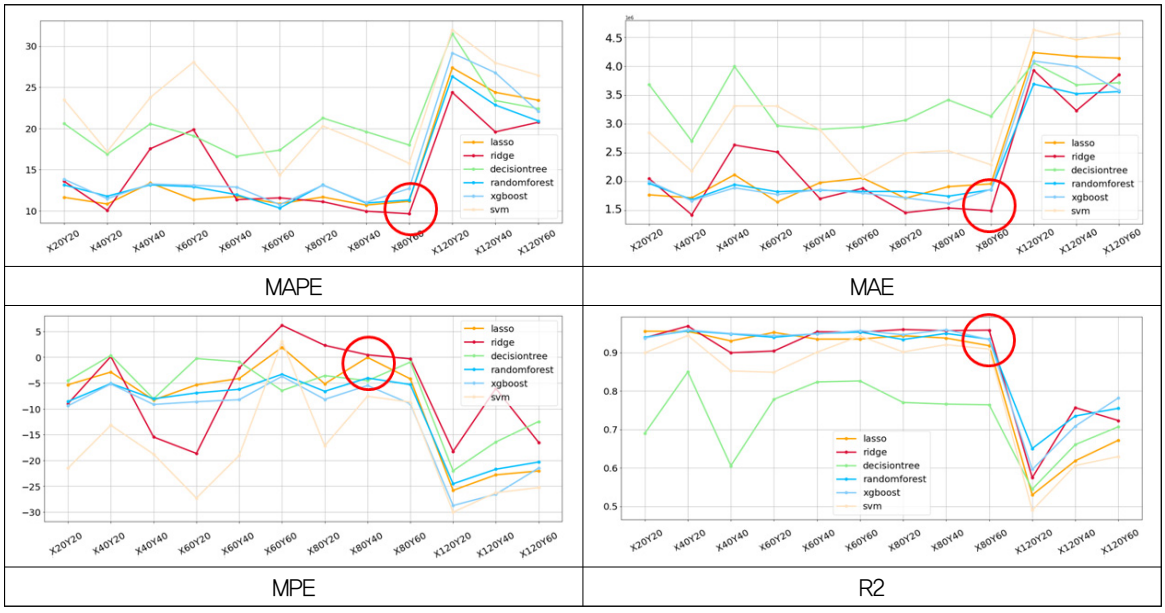
4.2. 성능평가

앞서 언급한대로, 비용효율성 및 변경/장애 위험성 간의 균형 있는 최적의 x, y값을 선택하는 것은 매우 중요하므로, 우선 <표 2>의 전체 모형들에 대해서 성능 비교가 필요하다. <표 2>의 각 분석 모형은 주어진 x, y 조합에 대해 생성된 입력데이터는 각각의 머신러닝 모델에 동일하게 적용될 수 있도록 설계되었다. 각 머신러닝 적용 시 과적합을 방지하기 위하여 교차검증을 적용하고, 하이퍼파라미터를 최적화하기 위해 GridSearchCV를 사용하여 분석하였다. 각 분석 모형에 대해 평가지표별 성능 비교는 <그림 4>와 같다.

먼저 MAPE 성능을 보면, x가 60~80일 사이, 즉, 약 3~4개월의 관측기간을 이용하여 예측할 때 10% 수준의 준수한 성능을 보여주었다. MAPE 9~10% 수준의 최적 성능을 나타낸 것은 x가 80일, y가 60일인 모형 9-1로 이는 4개월의 관측기간 데이터를 이용하여 향후 3개월 동안의 최대 호가건수를 예측하는 경우이며

〈표 3〉 평가지표 계산식 및 설명

평가지표	계산식
MAE	$\frac{1}{n}\sum_{i=1}^{n}(\lvert 실제값 - 예측값 \rvert)$
	과소 또는 과대 예측 오차 각각의 크기가 상계되지 않고 반영
MAPE	$\frac{1}{n}\sum_{i=1}^{n}(\frac{\lvert 실제값 - 예측값 \rvert}{실제값})\times 100$
	MAE를 백분율로 변환
MPE	$\frac{1}{n}\sum_{i=1}^{n}(\frac{실제값 - 예측값}{실제값})\times 100$
	모델이 과소예측(+) 또는 과대예측(−)인지를 판단 MAE가 낮다는 전제하에, MPE값이 작을수록 실제값 대비 상대적으로 예측오차가 작음을 의미



〈그림 4〉 모형별 성능 결과

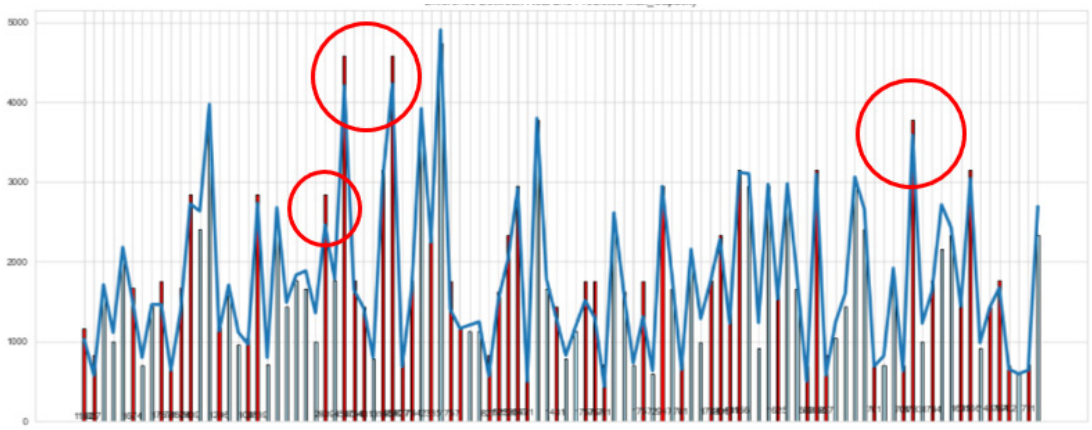
MAE도 동일 결과를 보인다. 이는 자본시장의 IT시스템 변경주기를 고려해볼 때 충분히 현실적인 기간이다. 아무리 MAPE와 MAE가 낮다고 하더라도, MPE가 양수라면 성능이 우수하다고 할 수 없다. 용량계획이라는 것은 과소 예측 시 시스템 장애를 유발할 수 있기 때문이다. 따라서 본 연구에서는 과소 예측이 필요하므로 MPE가 음수여야 한다. MAPE와 MAE에서도 가장 우수한 성능을 나타내었던 모형 9-1이 MPE도 음수이므로 해당 모형이 예측 오차도 최소화하면서 과소 예측을 회피하는 적절한 모형임을 확인할 수 있다. 마지막으로 각 모형의 R2를 분석한 결과, 모형 10, 11, 12를 제외하고 대다수의 모형들의 R2가 90% 이상으로 설명력이 충분한 것으로 나타났다. 특히, MAPE와 MAE 성능이 가장 우수한 모형 9-1이 R2도 95% 이상으로 나타나 설명력 또한 가장 우수한 것을 확인할 수 있었다.

상기의 분석을 통해 찾아낸 최적의 모형 9-1(x 80일, y 60일)에 대해 각 머신러닝별 비교를 위해 더욱 심층적인 분석을 진행하였다. 동일한 모형에 대해서 비교하는 것으로, 동일 x, y에 따라 전처리 과정을 통해 생성되는 입력데이터가 각각의 머신러닝 모델에 동일하게 적용될 수 있도록 설계되었다. 과적합 분석을 위해 학습 및 테스트 데이터에 대한 각각의 테스트 결과를 모두 생성하였다.

모델별 분석 결과, <표 4>와 같이 Ridge, Lasso, RandomForest 및 XGBoost와 같이 과적합에 강한 모델들이 전반적으로 우수한 성능을 나타내었다. 특히, 본 연구가 제안하는 Ridge 모델의 경우, 0.9582의 높은 설명력과 9.65%의 낮은 MAPE, 그리고 음수의 MPE를 보여주고 있어 매우 우수한 성능을 보인다. 더욱이, 과적합을 판단할 수 있는 학습 및 테스트 데이터 예측 결과 간 차이 또한 크지 않다. 이는 연구의 목적상 이상치가 다수 존재하는 데이터의 특성에 따른 것으로 풀이된다. 선형회귀 계열인 Ridge는 Lasso와 매우 유사한데, 학습 데이터를 이용하여 학습을 진행할 때 경사하강법을 이용하여 Mean Squared Error(평균제곱오차, MSE)가 최소화되도록 목적함수를 설정할 때, MSE에 L2 페널티를 추가하여 목적함수를 변경함으

〈표 4〉 모델별 성능 결과

모델명	MAPE(%)	MAE(만건)	MPE	R2	학습 · 테스트간 MAPE 차이
Ridge	**9.65**	**1,493,037**	**−0.32**	**0.9582**	**1.61**
Lasso	11.16	1,958,946	−4.21	0.9177	0.07
Decision Tree	23.59	3,569,350	−14.76	0.6681	5.39
SVM	15.78	2,290,528	−8.72	0.9089	0.09
RandomForest	11.29	1,832,811	−5.99	0.9377	1.88
XGBoost	12.71	1,863,303	−9.04	0.9336	4.49



〈그림 5〉 Ridge 예측 시각화

로서 이상치에 따른 과적합을 어느 정도 탈피할 수 있는 모델이다. Lasso는 MSE에 L1 페널티를 추가하는 모델이며, L1과 L2 페널티를 추가하는 것을 각각 L1, L2 정규화라 하며 이를 적용한 Lasso와 Ridge의 목적함수는 다음과 같다.

$$Obj = \frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i)^2 + \alpha\sum_{j=1}^{m}|w_j| \quad \text{(L1, Lasso)}$$

$$Obj = \frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i)^2 + \alpha\sum_{j=1}^{m}w_j^2 \quad \text{(L2, Ridge)}$$

where y_i *실제값,* Y_i *예측값,* w_j *가중치,* α *페널티*

L1의 경우 MSE 외에 가중치 페널티를 추가하였고, L2는 가중치의 제곱에 대한 페널티를 추가한 것이다. L1에서 α 값이 높아질수록 몇몇 가중치들은 0으로 수렴하고 이에 따라 변수의 수도 감소하는 효과가 있어 변수 선택의 효과가 있다. 즉 변수를 줄여주어 구불구불한 선을 펴주며 정규화 하는 것이다. 이와는 다르게 L2의 경우 α 이 커질수록, w 의 절대값을 0의 방향으로 수렴토록 하며, 선의 기울기를 감소시켜 복잡도를 줄여주는 효과가 있다. 본 연구에서 GridSearchCV를 이용하여 목적함수를 최소화하는 α 값을 모두 탐색하여 결정하였으며, 가장 우수한 성능을 보인 Ridge의 하이퍼파라미터 α 값은 2.8로 나타났다.

Ridge의 예측 결과를 구체적으로 살펴보기 위하여 약 300여개의 테스트 데이터에 대해 실제값과 예측값, 그리고 그 차이를 일부 시각화한 그래프는 <그림 5>와 같다. 그래프 내 파란색 꺾은선은 예측값을, 막대는 실제값을 의미하는데, 막대가 붉은 색일 때는 실제

값이 예측값보다 커 과소 예측된 경우이며, 하늘색인 경우는 실제값이 예측값보다 작아 과대 예측된 경우이다. 전체 그래프를 살펴보면, 예측값의 대다수가 실제값과 거의 유사하여 매우 정교한 예측을 하고 있음을 짐작할 수 있다. 다만, 차이가 크지 않더라도 전체 테스트 데이터 286개 중에 빨간색으로 과소 예측된 경우가 76건(약 26%)으로 다수 존재하므로, 실제 용량계획 모델을 실무에 적용하기 위해서는 여유용량을 일정 수준 마련할 필요가 있다. 본 연구에서는 여유용량을 별도로 4.3장에서 분석하였다.

4.3. 필요 여유용량 추가 분석

다수의 모델들이 우수한 성능을 나타내고는 있지만, IT시스템의 용량계획은 시스템 장애에 직접적인 영향을 줄 수 있는 중요한 사안이며 특히나 미션 크리티컬한 자본시장 IT시스템에서는 더욱 더 그러하다. 앞서 <그림 5>에 나타나듯이 과소 예측으로 인해 시스템 용량이 부족한 케이스가 발생할 수 있기에 예측 모델의 결과에 여유용량을 일정 수준 마련할 필요가 있다. 여유용량을 계산하기 위해서 각각의 분석 모델별로 테스트 데이터 전체에 대한 예측값과 실제값의 차이를 각각 구한 후 이중에서 가장 큰 값을 계산하였다. 이렇게 계산한 값만큼의 여유용량을 각 용량계획 모델에 적용하면 시스템 장애 가능성을 매우 낮출 수 있을 것으로 판단된다. 각 분석 모델별 필요 여유용량은 <표 5>와 같다.

본 연구에서 사용한 데이터는 2014년 3월 3일부터 2020년 7월 24일까지 약 7년간의 국내 파생상품 시장의 데이터로, 이 기간 중 호가건수는 평균적으로 1,165만 건이며, 최대치는 5,818만 건이다. 가장 우수한 성능을 보여준 Ridge 기준으로 필요한 여유용량은 지난 7년간의 평균치 대비 40%, 최대치 대비 약 8% 수준이다. 본 연구가 제안하는 Ridge 용량계획 모델에 추가적으로 약 467만 건의 여유용량을 운영한다면, 자본시장 IT시스템 용량계획 모델을 실무에 적용할 때 더욱 더 안정성을 강화할 수 있다.

4.4. 심리지수 유의성

본 연구에서는 심리지수의 유의성을 살펴보기 위해 가장 최적의 성능을 나타내는 모형 9-1(x 80일, y 60일)에 대해서 심리지수인 VKOSPI를 제거한 모형 9-2를 추가하여 분석하였다. 모든 데이터 셋은 동일하나, 모형 9-1의 데이터 내 VKOSPI와 관련된 lag 변수들만 제거된 데이터가 모형 9-2에 사용된다. <표 6>은 본 실험에서 가장 우수한 성능을 보인 Ridge에 대해 모형 9-1과 9-2를 비교한 것으로, VKOSPI 관련 변수들이 제거된 모형 9-2는 R-square 0.0183, MAPE 3.03%의 의미 있는 성능 저하가 나타났다. 또한, 모형 9-2는 MPE가 양수로 나타나 과소 예측된 것을 알 수 있다. 따라서 VKOSPI는 용량계획을 실무에 적용하기 위해 필수적인 변수이다.

모형 9-1과 9-2에 대한 Ridge 모델 간의 성능의 차이가 우연이 아닐 확률이 높은지를 확인하는 과정이 필요하다. 즉, 두 모델의 성능 결과 간의 통계적 유의미성을 검증하기 위해 각 모델별 테스트 데이터에 대한 예측값을 각각 표본 집단으로 삼아 이들 간에 통계적 유의미성을 검증하는 t-검정을 실시하였다. 본 연구에서는 각각의 표본 집단에 대해 Shapiro Wilk 검정을 실

〈표 5〉 필요 여유용량

Ridge	Lasso	Decision Tree	SVM	RandomForest	XGBoost
467만 건	587만 건	1,153만 건	516만 건	637만 건	1,196만 건

〈표 6〉 모형 9-1(심리지수 포함), 9-2(심리지수 미포함) 성능 비교 및 t-검정

Ridge	Test set	
	모형 9-1	모형 9-2
R^2	0.9582	0.9399
MAE(건)	1,493,037	1,883,795
MAPE(%)	9.65	12.68
MPE	-0.32	6.40
Wilcoxon t-검정	2797***	

주) *** $p < .001$

시한 결과 정규분포를 따르지 않음을 확인하였다. 정규성을 만족하지 않는 경우에는 중위값을 이용하는 비모수적 방법을 사용하여야 하며, 대응표본 중심 검정을 위해서는 Wilcoxon 검정을 주로 사용한다(Wilcoxon, 1992). 본 연구에서는 Wilcoxon 검정을 실시한 결과 두 개의 표본 간 중심위치가 유의미한 차이가 있음을 확인하였다. 즉, 자본시장 IT시스템의 용량계획 모델에 VKOSPI 지수를 포함함으로써 유의미한 성능 강화가 나타났음을 확인하였다.

5. 결론

5.1. 결과 논의

본 연구는 자본시장 IT시스템 용량계획 모델 개발을 위해 Ridge, Lasso, RandomForest 등 머신러닝 기반의 모델을 개발하고 성능을 비교 분석하였다. 특히, 모델 개발에 있어서 투자자의 심리상태를 반영하는 VKOSPI를 활용하였다. 본 연구가 제안하는 심리지수를 포함한 Ridge 모델은 R-square 약 0.9582, MAPE 약 9.65% 수준의 높은 성능을 보였다. 또한, 본 연구에 제안하는 모델은 성능 뿐 아니라 안정성도 높은 모델이다. 모델링 과정에서 lag 변수가 많이 생성되었고, 호가건수의 최대치를 예측하는 본 연구의 특성상 이상치를 제거할 수 없어 과적합 발생 가능성이 매우 높은 상황임에도 불구하고, 테스트 데이터 및 학습 데이터의 성능 차이가 매우 적어 안정성이 높다. 일반적으로 안정성을 강화하기 위한 하이퍼파라미터 튜닝 과정에서 일부 성능 감소는 불가피하나, 이러한 성능 감소에도 불구하고 MAPE 9.65% 수준의 성능을 보였다는 것은 주목할 만하다.

다음으로, 본 연구는 머신러닝 모델을 이용하여 설계하는 과정에서 가장 예측력이 높은 x, y를 도출하였다. x는 예측을 위해 관측할 과거 기간이며, y는 향후 최대 호가건수를 관측할 기간인데, 특히 y가 중요하다. 기업이 특정 시점 또는 주기적으로 IT시스템에 필요한 미래의 용량을 산정하여 이를 기반으로 용량변경을 수행하는데, y를 작게 설정하는 것은 짧은 주기로 용량 변경을 수행한다는 의미로, 비용효율성은 좋아질 수 있으나, 반대로 테스트 수행 등의 기간이 짧다는 것을 의미하여 장애 위험이 증가한다. 반대로 y를 길게 설정할 경우, 장애 위험은 감소하나 비용효율성은 저하될 수 있다. 본 연구에서 도출한 x(80일), y(60일)는 자본시장의 IT시스템의 용량을 변경하는 준비 기간을 고려할 때에도 충분히 현실적인 값으로, 가장 높은 성능을 보이면서 장애가능성과 비용효율성 간의 균형 있는 최적의 값이다.

한편, IT시스템 용량 계획 실패는 시스템 장애에 직접적인 영향을 줄 수 있다. 특히 자본시장 IT시스템

장애는 사회에 큰 혼란을 야기하는 만큼 철저한 용량 계획이 요구된다. 즉, IT시스템 용량 계획 모델이 과소 예측하여 시스템 용량이 부족한 케이스가 발생하는 경우를 대비하여, 일정 수준의 여유용량을 안전정치로 마련할 필요가 있다. 본 연구에서 추가적으로 분석한 결과, 약 467만 건 수준의 여유용량을 마련할 경우, 시스템 장애 발생 가능성을 매우 낮출 수 있음을 확인하였다(표 5). 이러한 분석은 기업의 용량계획 모델 오차 범위 또한 고려할 수 있게 함으로써, 본 연구에서 개발한 용량계획 모델을 적극적으로 실무에서 활용할 수 있도록 보완하는 것이다.

데이터 품질과 관련하여, 본 연구에서는 다수의 변수들 중에서 8개의 변수만을 입력 변수로 활용하였다. 적은 변수로 매우 높은 성능의 안정적인 모델을 개발하였으나, 해당 변수들이 타겟변수에 가장 영향력이 높은 변수들이라고 단언할 수는 없다. 실제로 자본시장에서는 입력 데이터 변수 관점에서 수백 또는 수천 개의 변수들이 존재하며 이들 중에 본 연구에서 사용한 변수보다 더 중요한 변수가 있을 가능성도 있다. 이들을 다양하게 조합하여 분석한다면, 본 연구에서 제안하는 모델보다 더 높은 성능과 안정성을 갖춘 모델을 개발하는 것도 충분히 가능하다.

마지막으로, 본 연구를 통해 자본시장에서 사람들의 행동을 나타내는 심리지수인 VKOSPI를 포함하여 MAPE 약 3% 내외 수준의 의미 있는 성능 강화를 확인하였다. 이는 동학개미운동에서 보듯이, COVID-19와 같은 불확실성이 높아지는 상황에서 사람들의 행동양식이 빠르게 변화하고 그에 따른 시스템의 사용량도 매우 빠르게 변화하고 있다는 것을 반증한다. 즉, 상품 및 소비자 행위 중심의 예측에서 더 나가가 이들을 서비스하는 IT시스템의 용량을 계획하는 것에도 심리지수가 큰 도움이 될 수 있다는 것을 확인한 것이다.

5.2. 시사점

본 연구는 다음과 같이 학술적 공헌을 가진다. 먼저, 본 연구는 수요 예측 문헌에 학술적으로 기여한다. 교통·전력·제조·소매 등 다양한 분야에서 수요 예측 연구가 진행되었는데, IT시스템 용량을 예측하는 연구는 거의 찾아볼 수 없다. 본 연구는 자본시장 시계열 데이터를 바탕으로 IT시스템 용량 예측모델을 개발할 수 있음을 보임으로써, 향후 연구의 토대를 마련하였다. 향후 연구에서는 일별 데이터가 아니라 시간 단위 또는 분 단위 등으로 데이터를 가공하여 충분한 데이터를 수집할 수 있다면, 딥러닝 계열의 모델 등 다양한 모델들을 폭넓게 활용하여 성능을 예측해볼 수 있을 것이다. 더 나아가, 본 연구에서 lag 변수의 생성을 위해 제한적으로 선정하였던 변수보다 더욱 더 많은 유의한 변수들을 추가하여 모델의 성능 및 안정성 향상을 기대할 수 있다.

또한, 본 연구는 투자자의 심리를 반영하는 심리지수가 자본시장 IT시스템 용량계획에 있어서 중요한 역할을 한다는 것을 보였다. 이는 자본시장에만 국한하는 것이 아니라 다른 도메인 또는 기업에 확대 적용이 가능하기에, 다양한 분야에서 심리지수를 활용한 연구가 촉진되는 계기가 될 수 있다. 이러한 심리지수는 이미 다양한 도메인에 존재하고 있기도 하는데, 경제 분야에서 경제심리지수나 부동산 분야에서 소비심리지수, 매수우위지수 등이 이에 해당한다. 이들은 대부분 통계나 정보제공의 목적으로 사용되고 있으나, IT시스템의 용량계획을 포함한 다양한 수요예측에 유용하게 활용될 수 있을 것이다.

본 연구는 다음과 같은 실무적 시사점을 가진다. 첫째, 본 연구는 기업들이 기존의 과용량 정책 대신, 머신러닝을 활용하여 유연한 용량계획 정책을 실무적으로 수립하고 사용할 수 있도록 도움을 준다. 본 연구

는 국내 자본시장을 대상으로 COVID-19 기간을 포함한 실증 데이터를 이용한 것으로서, 증권사시스템 등을 포함한 자본시장 도메인에서 공통적으로 사용할 수 있는 용량계획 모델을 개발하였다. 본 연구의 모델은 기존의 과용량 IT시스템 용량계획에서 벗어나 기업이 원하는 시점에 용량을 예측하고 조정할 수 있게 함으로써 비용효율성을 고려한 기업의 의사결정 지원이 가능하다.

둘째, 용량계획 모델을 개발함에 있어, 기업의 비용효율성 뿐 아니라 테스트 및 변경 주기 등의 실무적 제약을 고려한 최적의 관측기간(x: 80일) 및 예측기간(y: 60일) 값을 제안하였다. 즉, 본 연구가 제안하는 모델은 실증적인 데이터를 기반으로 하고 있으며, 그에 따라 모델을 구성하는 최적의 파라미터 또한 실증적으로 밝혔기에 실무적 활용가치가 더욱 높다고 할 수 있다. 본 연구의 모델을 활용하여 용량계획을 조정할 경우 비용효율성을 달성할 수 있을 것이라 기대된다.

마지막으로, 본 연구에서 제시한 모델은 자본시장 도메인뿐 아니라, 트랜잭션 기반의 용량계획 예측이 필요한 다양한 도메인으로 확대하여 적용할 수 있다. 즉, 해당 도메인에 필요한 데이터만 마련된다면, 이를 입력으로 하여 다양한 x, y 조합의 입력 데이터로 다양한 머신러닝을 비교 분석할 수 있다. 본 연구를 통해 의미 있는 수준의 용량계획 모델 개발이 가능함이 입증되었으므로, 다양한 도메인의 기업들이 IT시스템에서 머신러닝을 활용하여 용량계획 모델을 실무적으로 수립하여 사용하는 계기가 될 수 있을 것으로 보인다.

〈참고문헌〉

[국내 문헌]

1. 김은미 (2021). 감성분석을 이용한 뉴스정보와 딥러닝 기반의 암호화폐 수익률 변동 예측을 위한 통합모형. **지식경영연구, 22**(2), 19-32.
2. 나종회, 최광돈 (2004). 정보시스템 용량산정 방식에 관한 탐색적 연구: 공공부문 H/W 규모산정을 중심으로. **한국SI학회지, 3**(2), 9-23.
3. 원종관, 홍태호 (2021). 텍스트 마이닝과 딥러닝을 활용한 암호화폐 가격 예측: 한국과 미국시장 비교. **지식경영연구, 22**(2), 1-17.
4. 임현욱, 정승환, 이희수, 오경주 (2021). 국고채, 금리 스왑 그리고 통화 스왑 가격에 기반한 외환시장 환율예측 연구: 인공지능 활용의 실증적 증거. **지식경영연구, 22**(4), 71-85.
5. KRX (2009). **변동성지수(VKOSPI) 상품의 이해**. KRX, KRX-2009-14.

[국외 문헌]

6. Aggarwal, C. (2017). ***Outlier analysis***. Springer, pp. 1-34.
7. Bagchi, D., Lee, C. S., & Ryu, D. J. (2013). An investigation of return-volatility relationship using high-frequency VKOSPI data. ***Afro-Asian Journal of Finance and Accounting, 3***(3), 258-273.
8. Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. ***Journal of Finance, 61***(4), 1645-1680.
9. Buckman, S. R., Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). News sentiment in the time of COVID-19. ***FRBSF Economic Letter, 8***, 1-5.
10. Chen, K., Zhou, Y., & Dai, F. (2015, October). A LSTM-based method for stock returns prediction: A case study of China stock market. ***In 2015 IEEE International Conference on Big Data (big data)*** (pp. 2823-2824). IEEE.
11. Cho, J. K. (2016). Market timing with the VKOSPI sample entropy indicator. ***International Journal of IT-based Business Strategy Management, 2***(1), 17-24.
12. Du, B., Hu, X., Sun, L., Liu, J., Qiao, Y., & Lv, W. (2020). Traffic demand prediction based on dynamic transition convolutional neural network. ***IEEE Transactions on Intelligent Transportation Systems, 22***(2), 1237-1247.
13. Guo, Y., Wang, J., Chen, H., Li, G., Liu, J., Xu, C., ... & Huang, Y. (2018). Machine learning-based thermal response time ahead energy demand prediction for building heating systems. ***Applied Energy, 221***, 16-27.
14. Han, Q., Guo, B., Ryu, D., & Webb, R. I. (2012). Asymmetric and negative return-volatility relationship: The case of the VKOSPI. ***Investment Analysis Journal, 41***(76), 69-78.
15. Kumar, J., Saxena, D., Singh, A. K., & Mohan, A. (2020). Biphase adaptive learning-based neural network model for cloud datacenter workload forecasting. ***Soft Computing, 24***(19), 14593-14610.
16. Lee, C., & Ryu, D. (2014). The volatility index and style rotation: Evidence from the Korean stock market and VKOSPI. ***Investment Analysts Journal, 43***(79), 29-39.
17. Liang, C., Tang, L., Li, Y., & Wei, Y. (2015). Which sentiment index is more informative to forecast stock market volatility? Evidence from China. ***International Review of Financial Analysis, 71***, 101552.
18. Liu, S. (2015). Investor sentiment and stock market liquidity. ***Journal of Behavioral Finance, 16***(1), 51-67.
19. López-Cabarcos, M. A. et al. (2019). Investor sentiment in the theoretical field of behavioural finance. ***Economic Research, 33***(1), 2101-2228.
20. Lucey, B., & Dowling, M. (2005). The role of feelings in investor decision-making. ***Journal of Economic Surveys, 19***(2), 211-237.
21. Makrehchi, M., Shah, S., & Liao, W. (2013). Stock prediction using event-based sentiment analysis. ***2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)***.
22. Menasce, D., & Almeida, V. (1998). ***Capacity planning for web performance: Metrics, models, and methods***. Prentice Hall.
23. Mozo, A., Ordozgoiti, B., & Gomez-Canaval, S. (2018). Forecasting short-term data center network traffic load

with convolutional neural networks. *PLOS One, 13*(2), e0191939.

24. Muralitharan, K., Sakthivel, R., & Vishnuvarthanc, R. (2018). Neural network based optimization approach for energy demand prediction in smart grid. *Neurocomputing, 273*, 199-208.

25. Nelson, D. M. Q. et al. (2017). Stock market's price movement prediction with LSTM neural networks. *2017 IEEE International Joint Conference on Neural Networks(IJCNN)*.

26. Noh, J., Park, H. J., Kim, J. S., & Hwang, S. J. (2020). Gated recurrent unit with genetic algorithm for product demand forecasting in supply chain management. *Mathematics, 8*(4), 565.

27. Oh, C., & Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *ICIS 2011 Proceedings, 17*.

28. Piccoli, P., & Chaudhury, M. (2018). Overreaction to extreme market events and investor sentiment. *Applied Economics Letters, 25*(2), 115-118.

29. Qiu, L., & Welch, I. (2004). Investor sentiment measures. *Working Paper 10794*, National Bureau of Economic Research.

30. Reis, P. M. N., & Pinho, C. (2020). A new european investor sentiment index (EURsent) and its return and volatility predictability. *Journal of Behavioral and Experimental Finance, 27*, 100373.

31. Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. *Journal of Big Data, 7*, 53.

32. Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality(complete samples). *Biometrika, 52*(3/4), 591-611.

33. Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2017). Divergence of sentiment and stock market trading. *Journal of Banking & Finance, 78*, 130-141.

34. Tugay, R., & Oguducu, S. G. (2020). Demand prediction using machine learning methods and stacked generalization. *6th International Conference on Data Science, Technology and Applications*.

35. Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications, 79*, 153-163.

36. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83.

37. Xiao, G., Wang, R., Zhang, C., & Ni, A. (2021). Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks. *Multimedia Tools and Applications, 80*(15), 22907-22925.

38. Xing, F., Cambria, E., & Welsch, R. (2018). Intelligent asset allocation via market sentiment views. *IEEE Computational Intelligence Magazine, 13*(4), 25-34.

39. Yu, Y., Jindal, V., Bastani, F., Li, F., & Yen, I. L. (2018). Improving the smartness of cloud management via machine learning based workload prediction. *In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 38-44). IEEE.

[URL]

40. 홍승빈 (2020, 7월 3일). 먹통, 또 먹통...비대면 시대 무색한 증권사 거래시스템. **한국금융신문**, https://www.fntimes.com/html/view.php?ud=2020070321221391156c0eb6f11e_18

부록: 파생상품 시장 수집 데이터(변수)

변수명	설명	변수명	설명
P_Trading_Amount_T	거래대금(합계)	N_Traded_Symbol_T	거래종목수(합계)
P_Trading_Amount_F	거래대금(선물)	N_Traded_Symbol_F	거래종목수(선물)
P_Trading_Amount_O	거래대금(선물)	**N_Traded_Symbol_O**	**거래종목수(옵션)**
P_Trading_Volume_T	거래량(합계)	N_Open_Interest_T	미결제약정(합계)
P_Trading_Volume_F	**거래량(선물)**	N_Open_Interest_F	미결제약정(선물)
P_Trading_Volume_O	거래량(옵션)	N_Open_Interest_O	미결제약정(옵션)
N_Symbol_T	종목수(합계)	N_Active_Account	활동계좌수
N_Symbol_F	종목수(선물)	**N_Investor**	**투자자수**
N_Symbol_O	종목수(옵션)	R_Option_Trading Volume_Org	옵션매매비중(기관)
R_Future_Trading Volume_Org	선물매매비중(기관)	R_ Option _Trading Volume_Ind	옵션매매비중(개인)
R_Future_Trading Volume_Ind	선물매매비중(개인)	R_ Option _Trading Volume_For	옵션매매비중(외국인)
R_Future_Trading Volume_For	선물매매비중(외국인)	R_ Option _Trading Volume_Oth	옵션매매비중(기타)
R_Future_Trading Volume_Oth	선물매매비중(기타)	P_Market_Basis	시장베이시스
N_Order	**호가건수**	P_Theoratical_Basis	이론베이시스
N_Trade	**체결건수**	P_Disparate_Ratio	괴리율
P_Deposit	거래증거금예탁액	B_K200F_Final_Day	K200F 최종거래일여부
P_Margin	**거래증거금필요액**	**P_VKOSPI**	**KOSPI200 변동성 지수**
P_K200S	**KOSPI200 현물가격(종가)**	P_Settle_Amount	결제대금합계
P_K200F	KOSPI200 선물가격(종가)		

주) 굵은 글씨체는 모델에 최종 입력변수로 사용된 변수를 말함.

저 자 소 개

이 국 형 (Kukhyung LEE)

현재 한국거래소에서 과장으로 재직 중이다. 연세대학교 정보대학원에서 정보시스템학(세부전공: 디지털 경영) 석사 학위를 취득하였다. 한국거래소 차세대시스템 구축 등 다수의 대형 프로젝트에 참여하였으며, 주요 관심분야는 지식경영시스템, 지식공유, 기술경영, 머신러닝, 데이터 분석 등이다.

김 미 예 (Miyea KIM)

현재 창원대학교 경영대학 경영학과 조교수로 재직 중이다. 성균관대학교에서 경영학 박사 학위를 받은 이후, 한국연구재단 박사후연수를 수행하고 연세대학교 바른ICT연구소 연구교수로 연구와 교육 활동을 지속하였다. 관심 연구 분야는 온라인 정보 소비와 소비자 심리, ICT와 소비 행동, Information diversity 등이다. 해당 주제로 Asia Pacific Journal of Marketing and Logistics, Technology in Society, Journal of Research in Interactive Marketing, 경영학연구, 소비자학연구 등 국내외 학술지에 다수의 논문을 게재하였다.

박 재 영 (Jaeyoung PARK)

현재 연세대학교 정보대학원 AI-빅데이터 기반 초스마트 사회 구현을 선도하는 교육연구단 박사후연구원으로 재직 중이다. 연세대학교 정보대학원에서 정보시스템학(세부전공: 정보보호) 박사 학위를 취득하였다. 주요 관심분야는 정보보호, 프라이버시, 소셜미디어, ICT 기술 등이다.

김 범 수 (Beomsoo KIM)

현재 연세대학교 정보대학원 원장으로 재직 중이다. 연세대학교 바른ICT연구소 소장, Asia Privacy Bridge Forum 의장으로 ICT 정책, 격차, 과의존, 정보보호등의 이슈 중심으로 관련 연구와 교육 활동을 추진하고 있다. OECD 디지털 거버넌스와 프라이버시 작업반 부의장, 한국대표로 국제기구에서 AI시대 공공데이터 활용과 프라이버시 관련한 국제협력, 정책가이드 등을 마련하고 있다. 관심 연구 분야는 ICT의 효과적 활용, 데이터 거버넌스와 공개된 자료의 활용, 프라이버시, 개인정보보호, 국제협력정책 등이다.

〈 Abstract 〉

Effective Capacity Planning of Capital Market IT System: Reflecting Sentiment Index

Kukhyung Lee*, Miyea Kim**, Jaeyoung Park***, Beomsoo Kim****

Due to COVID-19 and soaring participation of individual investors, large-scale transactions exceeding system capacity limits have been reported frequently in the capital market. The capital market IT systems, which the impact of system failure is very critical, have encountered unexpectedly tremendous transactions in 2020, resulting in a sharp increase in system failures. Despite the fact that many companies maintained large-scale system capacity planning policies, recent transaction influx suggests that a new approach to capacity planning is required. Therefore, this study developed capital market IT system capacity planning models using machine learning techniques and analyzed those performances. In addition, the performance of the best proposed model was improved by using sentiment index that can promptly reflect the behavior of investors. The model uses empirical data including the COVID-19 period, and has high performance and stability that can be used in practice. In practical significance, this study maximizes the cost-efficiency of a company, but also presents optimal parameters in consideration of the practical constraints involved in changing the system. Additionally, by proving that the sentiment index can be used as a major variable in system capacity planning, it shows that the sentiment index can be actively used for various other forecasting demands.

Key Words: Effective Capacity Planning, Sentiment Index, VKOSPI, Capital Market IT Systems, Knowledge Management

* Graduate School of Information, Yonsei University
** College of Business Administration, Changwon National University
*** Graduate School of Information, Yonsei University
**** Graduate School of Information, Yonsei University