**머신러닝 기반 공매도 지표와**

**주식 수익률의 횡단면**

**Machine Learning-Based Short-Selling Indicators and the Cross-Section of Stock Returns**

정단열[[1]](#footnote-1), 이가현[[2]](#footnote-2), 조서현[[3]](#footnote-3), 이민혁[[4]](#footnote-4), 고지훈[[5]](#footnote-5)

Danyeol Jung, Gahyeon Lee, Seohyeon Jo, Minhyuk Lee, and Jihoon Goh

2025년 7월 24일

**머신러닝 기반 공매도 지표와**

**주식 수익률의 횡단면**

**초록**

본 연구는 한국 주식 시장에서 공매도 거래량이 미래 주식 수익률에 미치는 영향을 분석하고, 머신러닝 기법을 활용한 새로운 공매도 지표를 제시한다. 상대적 공매도 거래량이 높은 주식일수록 미래 수익률이 유의하게 낮았으며, 이러한 음의 예측력은 포트폴리오 분석과 Fama-MacBeth 회귀분석을 통해 검증되었다. 기존 공매도 지표가 과거 공매도 거래량에 의존하는 후행적 성격을 지닌다는 한계에 착안하여, 대차잔고 데이터를 활용하여 다음 달 공매도 거래량을 예측하는 선행적 지표를 개발하였다. 이 과정에서 다양한 머신러닝 예측 모형을 비교한 결과, 릿지 회귀 기반의 선형 모형이 가장 우수한 성능을 보였다. 본 연구는 공매도 데이터를 활용하여 예측력이 있는 새로운 지표를 제시함으로써, 기존 연구의 한계를 보완하고 실무적 투자 전략에 시사점을 제공한다.

주제어: 공매도, 주식 수익률, 이상현상, 위험 팩터, 머신러닝

**Ⅰ. 서론**

공매도는 상품의 가격 하락을 예상하여 주식이나 채권을 보유하지 않은 상태로 빌려서 파는 행위이다. 공매도 거래는 시장 참여자, 특히 기관 투자자들의 부정적 시장 전망을 반영한다는 점에서 중요한 의미를 갖는다. 일반적으로 공매도자는 정보에 기반한 거래자로 간주되며, 이들의 거래는 향후 가격 하락에 대한 정보를 내포하고 있을 가능성이 높다. 따라서 특정 주식의 공매도 거래량이 높다는 것은 그만큼 향후 주가 하락 가능성이 크다는 신호로 해석될 수 있으며, 공매도 거래량은 유용한 주가 예측 지표로 활용될 수 있다. 이러한 특성을 바탕으로, 상대적 공매도 거래량이 낮은 주식을 매수하고 높은 주식을 매도하는 롱숏 포트폴리오 전략을 구상할 수 있다.

학문적으로 공매도의 금융 시장에서의 역할에 대한 연구가 활발히 이루어지고 있다. Miller (1977)를 시작으로 많은 연구에서 공매도가 금융 시장에서 중요한 역할을 하고 있음을 밝혔다. Diether, Lee, and Werner (2009)는 상대적 공매도 거래량(*relss*)을 정의하고 이를 주식 수익률과 연관시켜 분석하였다. 연구 결과, 공매도 비율이 10% 증가할 때 월별 수익률이 0.72~0.94% 감소하는 경향성이 나타났으며, 이는 공매도자들이 주가 하락을 타당하게 예견할 수 있음을 시사한다. Boehmer and Wu (2013)는 공매도자들의 활동량이 증가할수록 미래의 주식 가격을 더 정확하게 예측할 수 있다는 결론을 내리고 있다. Desai, Ramesh, Thiagarajan, and Balachandran (2002)는 공매도가 많이 이루어진 주식의 향후 수익률이 낮은 경향이 있다는 실증 결과를 확인하였다. Asquith, Pathak, and Ritter (2005)는 공매도 거래량이 높은 주식의 가격이 하락하는 경향성이 있음을 보였으며, 공매도 거래량이 주식 수익률에 대한 예측력이 있는 지표임을 강조하였다.

Wang and Lee (2015)는 한국 시장을 대상으로 상대적 공매도 거래량을 분석하고, 외국인 보유비중과의 관계를 검토하였다. 한국 시장의 주요 공매도 거래가 외국인 투자자에 의해 발생하며, 공매도자들이 단기적인 반대매매 전략을 사용하고 미래 주가에 대한 정보를 보유하고 있음을 확인하였다. Wang (2023)은 한국 주식 시장에서의 부실기업의 역할에 주목하여 공매도 활동과 다음 달 주식 수익률 사이의 음의 상관관계를 확인하였다. Wang and Lee (2015)와 Wang, Lee, and Woo (2017)에서는 한국 주식 시장에서 투자자 유형에 따른 공매도의 정보 전달 역할의 차이를 분석하였다.

공매도 관련 지표가 수익률 예측력을 가질 수 있는 이유는 주식 시장에서의 공매도 제한과 밀접한 관련이 있다. 일반적으로 차익거래자는 예측력 있는 정보를 활용하여 가격 차이를 이용한 거래를 수행하지만, 공매도가 제한된 환경에서는 이러한 차익거래의 수행에 제약이 발생한다. 이로 인해 정보가 가격에 즉시 반영되지 못하고 일정 기간 왜곡된 상태로 남게 되며, 공매도 관련 지표가 예측력을 가질 수 있는 여지가 생긴다. Shleifer and Vishny (1997), Hong and Stein (2003), Boehmer, Jones, and Zhang (2008) 등의 연구에서는 공매도 제한이 시장 효율성과 가격 형성에 미치는 영향을 분석하였다.

본 연구는 기존 연구에서 사용된 상대적 공매도 거래량(*relss*)의 음의 수익률 예측력을 한국 시장에서 검증하고, 예측력을 향상시킬 수 있는 새로운 공매도 지표를 개발하였다. 먼저 기존의 *relss* 지표를 사용하여 음의 수익률 예측력을 확인하였다. 구체적으로, *relss*를 기준으로 주식을 정렬하여 10개의 포트폴리오를 구성하고, *relss*가 가장 높은 포트폴리오를 매수하고 가장 낮은 포트폴리오를 매도하는 롱숏 포트폴리오 전략을 수행하였다. 그 결과, 평균 수익률(Fama-French 4팩터 알파)이 -0.45%(*t*-통계량 -2.10)로 통계적으로 유의하게 음수임을 확인하였다. 이를 통해, 공매도 거래자들이 미래 주가에 대한 정보를 보유하고 있음을 재확인할 수 있다.

기존 공매도 지표들은 과거 공매도 거래량에 의존하기 때문에 후행적 성격을 지니며, 예측력 측면에서 한계가 존재한다. 본 연구는 이러한 한계를 극복하기 위해 대차잔고 데이터를 활용하여 다음 달 공매도 거래량을 예측하는 선행적 공매도 지표를 머신러닝 기법을 통해 개발하였다. 대차잔고는 공매도 거래를 위해 대기 중인 자금을 의미하므로, 미래의 공매도 거래량을 예측하는 데 중요한 단서가 될 수 있다. 먼저 대차잔고 데이터를 직관적으로 반영한 새로운 공매도 지표()를 개발하고 포트폴리오 분석을 수행한 결과, 롱숏 전략의 평균 수익률(Fama-French 4팩터 알파)은 -0.36%(*t*-통계량 -1.57)로 기존 *relss*에 비해 예측력이 낮았다.

이에 따라 머신러닝 방법론을 도입하여 대차잔고 데이터를 보다 효과적으로 활용하여 지표를 개발하고 예측력을 향상시켰다. 릿지 회귀, 서포트 벡터 머신, 랜덤 포레스트 등 다양한 머신러닝 방법론을 선형 및 지수 형태로 적용하여 비교한 결과, 릿지 회귀를 선형 모형으로 적용했을 때 가장 우수한 성능을 보였다. 해당 머신러닝 기반 지표를 이용하여 구성한 롱숏 포트폴리오의 평균 수익률(Fama-French 4팩터 알파)은 -0.79%(*t*-통계량 -3.13)로 기존 지표보다 개선되었으며, Fama-MacBeth 회귀분석 결과도 일관된 예측력을 확인하여 머신러닝 접근법의 효과성을 입증하였다.

또한, 새롭게 개발한 공매도 지표를 팩터로 만들어 Fama-French 3팩터 모형에 적용하여 추가적인 설명력을 제공함을 보였다. Dong, Wu, Fang, Gozgor, and Yan (2022)의 방법론을 따라 스패닝 테스트를 실시한 결과, 새로운 공매도 팩터의 유의성을 확인하였다.

본 연구는 공매도 데이터를 활용한 자산 수익률 예측 연구에 다음과 같은 두 가지 측면에서 기여한다. 첫째, 기존 공매도 지표가 과거 거래량에 의존하는 후행적 성격을 지닌다는 한계에 착안하여, 본 연구는 대차잔고 데이터를 활용해 향후 공매도 거래량을 예측하는 선행적 지표를 제안하였다. 이 과정에서 다양한 머신러닝 기법을 적용하고, 그 예측 성능을 비교함으로써 적절한 정규화와 해석 가능성을 갖춘 릿지 회귀 모형이 가장 효과적인 예측 도구임을 실증적으로 확인하였다. 이는 공매도 거래량 예측에 있어 모형의 구조적 특성 또한 중요하다는 점을 보여준다. 해당 지표는 실무적 투자 전략 수립에도 활용 가능성이 높다는 점에서 실질적인 의의가 있다.

둘째, 대부분의 선행 연구가 미국 시장을 중심으로 이루어진 반면, 본 연구는 공매도에 대한 제도적 제약과 투자자들의 관심이 특히 높은 한국 주식 시장을 분석 대상으로 설정하여, 공매도 거래의 정보 함의와 예측력을 실증적으로 검증함으로써 기존 문헌과 차별화된다.[[6]](#footnote-6) 또한, 개발된 공매도 예측 지표를 바탕으로 구성한 새로운 팩터는 기존 Fama-French 3팩터 모형의 팩터 대비 유의미한 설명력 개선을 보였으며, 이는 공매도 데이터를 활용한 팩터가 수익률 결정에 실질적으로 기여할 수 있음을 시사한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터와 주요 변수의 구성 방법을 설명한다. 3장에서는 공매도 거래량을 기반으로 한 지표가 주식 수익률에 미치는 영향을 분석한다. 4장에서는 공매도 거래량과 함께 대차잔고 데이터를 활용하여 미래 공매도 거래량을 예측하는 지표를 개발하고 그 지표의 주식 수익률 예측력을 분석한다. 5장에서는 공매도 지표를 기반으로 위험 팩터를 구성하고 Fama-French 3팩터 모형에서 추가적인 설명력을 가지는지 확인한다. 6장에서는 논문의 결론을 제시한다.

**Ⅱ. 데이터 및 변수 설정**

본 연구에서는 한국 주식 시장의 KOSPI와 KOSDAQ에서 상장된 모든 주식을 사용하였다. 또한, 재무 보고 방식 및 자본 구조가 다른 금융업은 데이터의 일관성을 유지하기 위해 제외하였다. DataGuide에서 제공하는 일별 주식 데이터를 월별로 변환하여 사용하였다. <표 1>에 주요 변수들의 정의와 산출 방법을 정리하였다.

<표 1>

한국에서는 2000년 이후 2008년 10월 1일부터 2009년 5월 31일, 2011년 8월 10일부터 2011년 11월 9일, 2020년 3월 16일부터 2021년 5월 2일까지, 그리고 2023년 11월 6일부터 2025년 3월 30일까지 공매도 금지 기간이었다. 이 기간에는 유동성 공급자(LP)에 의한 공매도만 이루어졌기 때문에, 공매도 지표가 시장에 미치는 영향을 적절하게 평가하기 어렵다고 판단하였다. 이에 따라 해당 기간을 제외하고, 2008년 7월부터 2023년 11월까지 총 162개월의 데이터를 사용하여 분석을 진행하였다.

본 연구에서는 머신러닝 방법론을 적용하여 다음 달 공매도 지표를 예측하는 모형을 구축하기 위해 전체 데이터를 훈련 데이터와 테스트 데이터로 분할하였다. 구체적으로 2011년 12월 이전 기간(총 32개월)을 훈련 데이터로 사용하였는데, 이는 전체 기간의 약 20%에 해당한다. 테스트 데이터를 최대한 확보하고 예측 성능을 현실적으로 평가하기 위해 훈련 기간을 상대적으로 짧게 설정하였으며, 2012년 1월 이후 기간(총 130개월)은 테스트 데이터로 활용하였다. 또한, 지표 사이의 비교 가능성과 분석의 일관성을 높이기 위해 모든 실증 분석은 테스트 데이터 기간인 130개월 샘플을 기준으로 통일하여 수행하였다.

공매도 거래량을 측정하기 위해 본 논문에서는 *relss* 지표를 사용하였다. 이 지표는 상대적 공매도 거래량을 나타내며 공매도 전략, 즉 공매도 거래량이 높은 주식을 매수하고 공매도 거래량이 낮은 주식을 매도하는 전략을 시행하기 위한 기준이 된다.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

식 (1)과 같이 각 주식 *i*의 *relss* 값은 월별로 공매도된 주식 수를 전체 거래된 주식 수(거래량)으로 나누어 계산한다. Diether, Lee, and Werner (2009)는 *relss* 지표를 사용하여 NYSE, NASDAQ에서의 공매도 활동과 여러 가지 변수들 간 선형 회귀분석을 통해 공매도자들이 어떤 전략을 사용하는지에 대한 분석을 진행한 바 있다. Wang and Lee (2015)에서 한국 시장, 특히 외국인 공매도 투자자의 전략에 대한 비슷한 분석을 시행할 때도 사용하였다.

**Ⅲ. 공매도와 주식 수익률**

3장에서는 공매도 활동과 미래 주식 수익률 사이의 관계를 실증적으로 분석한다. 먼저 3.1절에서는 상대적 공매도 거래량(*relss*)에 따라 포트폴리오를 구성하고, 이들의 미래 수익률을 비교함으로써 공매도 지표의 예측력을 평가한다. 3.2절에서는 횡단면 회귀분석을 통해 수익률 예측력을 평가한다.

**1. 포트폴리오 분석**

실증 분석의 첫 번째 단계는 *relss*에 기반한 포트폴리오를 구성하는 것이다. 데이터에 포함된 모든 주식을 월말의 *relss*를 기준으로 정렬한 후, 10개의 포트폴리오로 구성하여 다음 달 수익률, 시가총액, 가격, 장부가-시장가 비율을 관찰하였으며, 그 수치를 <표 2>에 정리하였다. 여기에서 포트폴리오 1은 *relss*가 가장 작은 포트폴리오를 의미하며, 포트폴리오 10은 *relss*가 가장 큰 포트폴리오를 의미한다.

<표 2>

<표 2>를 살펴보면, 공매도 거래량 *relss*가 높아질수록 시가총액이 커지는 것을 확인할 수 있다. 이는 *relss*와 시가총액 간의 양의 상관관계가 존재한다는 증거이다. 이러한 상관관계는 *SMB* 팩터와의 겹침을 발생시킬 수 있기에, 시가총액의 영향을 제거하고 새로운 공매도 팩터의 독립성을 추출하는 전처리 작업이 필요하다.

다음으로, 동일 가중 포트폴리오, 시가총액 가중 포트폴리오 각각에 대해, CAPM 모형, Fama-French 3팩터 모형, Fama-French 4팩터 모형에 대한 월별 시계열 회귀분석을 진행하고 각 포트폴리오의 초과수익률() 값을 구하였다. 포트폴리오 10과 포트폴리오 1과의 수익률(EW 또는 VW) 차이, 즉 롱숏 포트폴리오의 수익률 또한 회귀분석의 종속변수로 활용하여 초과수익률을 산출하였다.

<표 3>의 결과를 살펴보면, 포트폴리오의 *relss*가높을수록 전반적으로 초과수익률이 감소하는 경향을 확인할 수 있다. 동일 가중 포트폴리오에서 *relss*가 가장 작은 1번째 포트폴리오의 수익률은 *relss*가 가장 큰 10번째 포트폴리오의 수익률보다 평균적으로 0.85%(*t*-통계량 3.03) 더 높은 것을 확인할 수 있다. Fama-French 4팩터 모형에서 또한 0.45%(*t*-통계량 2.10)의 차이를 나타나는 것으로 보았을 때 초과수익률과 *relss*가 음의 상관관계를 가지는 것을 알 수 있다. 롱숏 포트폴리오의 초과수익률과 알파에 대한 *t*-통계량은 1% 유의수준에서 통계적으로 유의하다. 따라서 특정 주식의 공매도 거래량이 높을수록 그 주식의 수익률이 낮아진다는 가설을 인용하는 것이 적합하다고 판단할 수 있다. 하지만 가치 가중 포트폴리오에서는 상대적으로 약한 유의성을 나타냈다.[[7]](#footnote-7)

<표 3>

**2. Fama-MacBeth 회귀분석**

*relss* 지표로 구성한 롱숏 포트폴리오가 기존 Fama-French 4팩터 모형으로 설명할 수 없는 독립적인 수익의 원천을 가진다는 것을 Fama-French 4팩터 알파 테스트를 통해서 직접적으로 확인하였다. 다음으로, 주식의 여러 횡단면 특성을 고려하고자 Fama and MacBeth (1973)가 제시한 Fama-MacBeth 회귀분석을 수행한다. Fama-MacBeth 회귀분석은 주식의 특성 및 지표가 미래의 수익률을 예측할 수 있는지 검증하기 위해, 횡단면 회귀분석을 월별로 진행하고 이렇게 얻은 계수를 시계열로 평균 내어 그 평균값의 통계적 유의성을 평가하는 방법이다. 본 연구에서는 *relss*를 독립변수로 포함하고, 다음 달의 수익률을 종속변수로 놓고 회귀분석을 진행하였다. 미래 주식 수익률을 설명할 수 있다고 알려진 변수인 *BETA*[[8]](#footnote-8), *ME*(시가총액), *PRC*(종가), *BM*(장부가-시장가 비율)을 함께 포함하여 Fama-MacBeth 회귀분석을 수행한다.

<표 4>는 그 결과를 표시하고 있으며, 개별 변수들과 다음 달의 수익률 간의 회귀분석 계수 및 *t*-통계량, 그리고 복합 모형에서의 결과도 포함한다. *relss* 개별로 횡단면 분석을 진행하였을 때(Model 1)와 다른 변수들과 함께 다중선형회귀분석을 진행하였을 때(Model 6)의 *relss* 계수는 통계적으로 유의하다. 특히 Model 6에서 *relss* 계수의 *t*-통계량은 -3.94으로 유의한 음의 값을 가지는데,이는 장부가-시장가 비율, 시가총액 등의 변수들을 통제한 후에도 공매도 거래량이 주식 수익률에 미치는 영향이 유의미하게 남아 있음을 보여준다. 이는 *relss* 계수의 변동성이 작고 다른 변수들과는 독립적인 수익률 예측력을 지니고 있음을 시사한다. 유의한 음의 계수를 통해 주식의 공매도 거래량이 높을수록 미래 주식 수익률이 낮은 경향이 있다는 것을 알 수 있다.

<표 4>

Fama-French 4팩터 초과수익률 검증과 Fama-MacBeth 분석 결과는 모두 상대적 공매도 거래량이 높은 주식의 다음 달 수익률이 감소한다는 가설을 뒷받침한다. 두 분석 모두 *relss* 지표와 이를 기반으로 구성될 공매도 팩터가 기존 위험 팩터와는 독립적인 기여를 할 수 있다는 점을 시사한다. 다시 말해, 투자자들은 공매도 데이터를 분석함으로써 추가적인 투자 인사이트를 얻을 수 있으며, 이는 기존의 시장 팩터들과 별개로 고려할 필요가 있음을 의미한다.

**Ⅳ. 대차잔고를 고려한 지표**

4장에서는 대차잔고 데이터를 활용하여 다음 달의 공매도 거래량을 예측할 수 있는 지표를 개발하고, 그 지표의 수익률 예측력을 평가한다. 기존의 상대적 공매도 거래량(*relss*)은 실제 발생한 거래량을 기반으로 하지만 이번 달의 공매도 거래량에 관한 지표이기 때문에 이를 사용하여 다음 달의 수익률을 예측하는 것에 한계가 있을 수 있다. 이에 따라, 본 연구는 미래 공매도 활동의 신호로 사용될 수 있는 대차잔고 변화율을 활용하여 새로운 지표를 정의하고자 한다. 대차잔고는 향후 공매도 거래에 활용될 수 있는 잠재적인 수요를 반영하므로 예측 지표로서의 확장 가능성이 있다. 본 장에서는 먼저 단순 예측 모형을 통해 대차잔고 기반 지표의 성과를 살펴보고, 머신러닝 방법론을 적용한 다양한 예측 모형을 구성하여 성과를 비교한다. 이를 통해 공매도 정보를 바탕으로 한 지표의 미래 수익률 예측력을 개선할 수 있는 가능성을 탐색한다.

**1. 포트폴리오 분석**

새로운 지표 *reldc*를 식 (2)와 같이 대차잔고 변화량를 상장주식수로 나누는 것으로 정의한다. 이 지표는 대차잔고 변화율로 상장된 주식 수에 대한 대차잔고의 변화율을 의미한다.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

시장 민감도 및 미래 수익률에 대한 예측력이 검증된 *relss*와 대차잔고 변화율 *reldc*를 조합하여, 다음 달의 공매도 거래량, 더 나아가 다음 달의 수익률을 예측할 수 있는 최적의 지표를 만드는 것을 목표로 한다. 직관적으로 보았을 때 대차잔고가 늘어났다는 것은 그 다음 달에 공매도를 할 목적으로 주식을 대여했다는 뜻이기에, 상대적 공매도 거래량에 대차잔고 변화율을 곱하여 대략적인 예측을 시행한다. 미래 공매도 거래량을 대략적이고 직관적으로 예측하기 위해 이번 달 공매도 거래량과 대차잔고 변화율 *reldc*에 1을 더한 값을 곱한다. 즉, 이번 대차잔고 변화율만큼 다음 상대적 공매도 거래량이 변화할 것이라고 예측하는 것이다. 이러한 방식으로 새로운 지표 를 식 (3)과 같이 정의하고, 이를 기준으로 포트폴리오 분석을 진행한다.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

를 기준으로 롱숏 포트폴리오를 구성하고, Fama-French 4팩터 알파를 계산한 값을 <표 5>의 ‘기존 결과’에 나타냈다. 롱숏 포트폴리오의 동일 가중 Fama-French 4팩터 알파 -0.36%(*t*-통계량 -1.57)로 유의성이 약해진 것을 확인하였다. *t*-통계량이 줄어든 이유는 *reldc*의 정보가 효과적으로 반영되지 않았기 때문으로 추측된다. 이에 따라, 상대적 공매도 거래량과 대차잔고 변화율을 더 효과적으로 조합하는 방법을 찾을 필요가 있으며 이는 3절에서 검토한다.

<표 5>

**2. Fama-MacBeth 회귀분석**

*relss*에 대해 진행했던 Fama-MacBeth 회귀분석을 에 대해서 진행하였으며, 모든 변수에 정규화와 윈저화를 적용하였다. <표 6> ‘기존 결과’의 복합 모형 계수를 보면, 값이 평균에서 1 표준편차만큼 증가했을 때 다음 달의 주식 수익률이 0.188% 감소한다고 해석할 수 있다. 미래 주식 수익률에 대한 의 계수가 -0.188로 음의 상관관계를 보이는 것으로 미루어 보아, 공매도 거래량이 높은 주식이 확연하게 다음 달에 낮은 수익률을 보인다는 것을 확인할 수 있다. 이는 가 중요한 예측 변수로 작용할 수 있음을 시사한다. *t*-통계량은 *ME*와 *BM*, *BETA*를 컨트롤하기 전에는 -2.58, 후에도 -2.23으로 유의하게 유지되었다.

<표 6>

**3. 머신러닝 방법론**

Gu, Kelly, and Xiu (2020)를 비롯해 금융 분야에서도 주식 수익률을 설명하기 위해 머신러닝 방법론을 도입하는 연구가 활발하게 이루어지고 있다.[[9]](#footnote-9) 본 연구에서도 여러 머신러닝 방법론을 활용하여 미래의 공매도 거래량을 예측하고 이를 주식 수익률의 횡단면을 설명하는 데 사용하였다. 식 (3)의 공식을 사용한 의 직관적 정의는 예측력 및 수익성에 한계가 있으며 개선이 필요하므로, 머신러닝 모형을 활용하여 개선된 지표를 산출한다. *relss*와 *reldc*를 입력변수로, 다음 달의 *relss*를 출력변수로 설정하여 회귀분석 및 머신러닝 방법론을 여러 모형으로 적용하였다. 본 연구에서는 2011년 이전 데이터는 훈련 데이터, 2011년 이후 데이터는 테스트 데이터로 사용하였다. 그리고 머신러닝 방법론을 구현할 때 기본 하이퍼파라미터 세팅을 사용하였다.

|  |  |  |
| --- | --- | --- |
|  |  | (4) |
|  |  | (5) |
|  |  | (6) |

식 (4)와 같은 입력변수와 출력변수 사이의 선형 관계와 식 (5), (6)과 같은 지수 관계를 모두 고려하였다. 선형 회귀, 릿지 회귀, 서포트 벡터 머신(Support Vector Machine, SVM) 모형을 각각 식 (4)와 식 (6)의 형태로 적용하였다. 추가로 랜덤 포레스트(Random Forest, RF) 모형을 식 (4)의 형태로 적용하였다. 총 7개의 모형을 추가로 고려하여 기존의 결과와 비교하였다. 각 모형에서 산출된 예측 지표는 구분을 위해 윗첨자를 사용하여 표기하였다. 구체적으로 는 식 (4) 기반 선형 회귀 모형, 는 식 (5) 기반 선형 회귀 모형, 는 식 (4) 기반 릿지 회귀 모형, 는 식 (5) 기반 릿지 회귀 모형, 는 식 (4) 기반 서포트 벡터 머신 모형, 는 식 (5) 기반 서포트 벡터 머신 모형, 는 식 (4) 기반 랜덤 포레스트 모형을 각각 의미한다.

선형 회귀는 선형 관계를 단순하고 직관적으로 파악할 수 있는 기본 모형으로, 다른 복잡한 모형의 성능을 비교하고 평가하는 기준선 역할을 한다. 또한 계산이 빠르고 구현이 용이하여 다양한 데이터 환경에서 예측력 비교를 위한 참조 모델로 적합하다. 릿지 회귀는 선형 회귀에 L2 정규화[[10]](#footnote-10)를 적용하여 회귀계수의 크기를 제약함으로써 다중공선성을 완화하고 추정의 안정성을 높인다. 이를 통해 과적합을 방지하고, 제한된 변수 환경에서도 안정적인 예측 성능을 제공한다. 서포트 벡터 머신은 마진을 최적화하는 원리를 기반으로 예측 오차를 최소화하며, 기본 하이퍼파라미터로도 안정적인 회귀 결과를 제공할 수 있다. 특히 이상치나 잡음이 포함된 데이터에서도 일정 수준의 강건성을 유지한다는 장점이 있다. 랜덤 포레스트는 다수의 의사결정나무를 앙상블하여 비선형 관계와 변수 간 상호작용을 효과적으로 포착한다. 또한 개별 트리의 과적합 위험을 줄여 다양한 데이터 패턴에 대한 예측력을 높인다.

<표 5>에서 볼 수 있듯이, 동일 가중 포트폴리오에서 가장 높은 유의성을 보인 것은 식 (4) 기반 릿지 회귀 모형에서 산출한 이다. 이 모형의 Fama-French 4팩터 알파 결과는 동일 가중 포트폴리오에서 -0.79%(*t*-통계량 -3.13), 가치 가중 포트폴리오에서 -0.30%(*t*-통계량 -1.22)로 나타났다. 이는 기존 *relss* 및 의 정의보다 개선된 결과이다. 가치 가중 포트폴리오에서 통계적 유의성이 여전히 낮게 나타나는 이유는 대형주의 비중이 커지면서 공매도 지표의 예측력이 희석되기 때문이다. 공매도 지표의 예측력은 주로 소형주에서 강하게 나타나며, 동일 가중 포트폴리오는 소형주의 영향을 상대적으로 더 크게 반영한다. 비록 예측력이 소형주 중심의 효과에서 비롯된 측면이 있지만, 동일 가중 포트폴리오에서 확인된 유의한 결과는 공매도 변수의 예측 가능성을 검증한다는 점에서 충분히 의미가 있다.

또한, 이러한 의 정의들에 대해 Fama-MacBeth 회귀분석도 진행하였다. <표 6>의 결과를 보면, Fama-MacBeth 회귀분석의 결과 또한 식 (4) 기반 릿지 회귀()에서 가장 뛰어난 유의성을 보임을 알 수 있다. 복합 모형(Panel B)에서 다른 지표들을 조정하고 나서도 유의성을 보이고 있기에, 지표의 독립적인 예측력이 존재함을 알 수 있다. 복합 모형의 계수를 해석하면, 값이 평균에서 1 표준편차만큼 증가했을 때 다음 달의 주식 수익률이 0.394% 감소한다고 할 수 있다. 이는 기존 결과와 비교했을 때 경제학적으로도 두 배 이상 큰 영향력을 나타내며, 통계적으로도 더 유의한 결과(*t*-통계량 -3.95)이다. 추가적으로 <표 2>와 같이 지표를 기반으로 구성한 포트폴리오의 요약 통계량을 <표 7>에 제시하였다.

<표 7>

실증 분석 결과를 통해 적절한 정규화와 해석 가능성을 갖춘 릿지 회귀 모형이 가장 효과적인 예측 기법임을 확인하였다. 본 연구의 모형은 상대적 공매도 거래량과 대차잔고 변화율의 두 입력변수를 사용하며, 이들 사이에는 일정 수준의 상관관계가 존재할 수 있다. 릿지 회귀는 L2 정규화를 통해 계수의 분산을 줄여 추정의 안정성을 높이고, 입력변수 사이의 상관관계로 인한 과적합을 완화함으로써 예측 성능을 향상시킨다. 따라서 릿지 회귀의 정규화가 예측 성능 개선에 기여한 것으로 해석할 수 있다. 해당 지표는 실무적 투자 전략 수립에도 활용 가능성이 높다는 점에서 실질적인 의의가 있다.

**Ⅴ. 팩터 모형**

5장에서는 공매도 지표를 기반으로 새로운 팩터를 구성하고, 기존의 Fama-French 3팩터 모형에 추가할 수 있는지를 실증적으로 분석한다. 앞에서 수행한 포트폴리오 분석, Fama-MacBeth 회귀분석을 통해 공매도 지표가 미래 수익률을 예측할 수 있는 정보를 담고 있음을 확인했다. 본 장에서는 해당 정보를 기반으로 공매도 팩터를 구성하고 스패닝 테스트를 진행하였다. 공매도 팩터 *SSHML*은 가 상위 10% 인 주식을 매수하고, 하위 10% 주식을 매도하는 롱숏 포트폴리오의 수익률로 정의하였다.

Dong, Wu, Fang, Gozgor, and Yan (2022)은 Fama-French 3팩터 모형에서 4개의 팩터를 추가하고 이를 검증하기 위해 스패닝 테스트를 진행하였다. 스패닝 테스트는 개별 팩터를 번갈아가며 종속변수로, 나머지 팩터들을 독립변수로 설정하고 회귀분석을 수행하여 결정계수() 수치를 확인하는 것이다. 만약 새로 만든 팩터가 유의미하다면 다른 팩터들로 설명이 되지 않아야 하고, 새로운 팩터가 종속변수일 때의 모형 결정계수()가 비교적 낮은 값을 가져야 한다.

<표 8>는 스패닝 테스트의 결과를 나타낸다. 종속변수가 *SSHML*일 때의 *SMB* 계수와 종속변수가 *SMB*일 때의 *SSHML* 계수를 보면, 두 팩터 사이의 상관관계가 높음을 확인할 수 있다. 또한, 두 모형 모두에서 결정계수가 유사하게 높고, 앞서 언급한 두 계수가 각각 통계적으로 유의하므로 두 팩터 모두 의미 있는 설명력을 지닌다. 종속변수가 *HML*이나 *MKTRF*일 때의 결과를 보면 *SSHML* 계수가 *SMB* 계수보다 크고 통계적 유의성도 높게 나타난다. 이는 공매도 팩터가 *SMB* 팩터에 비해 더 우수한 예측력을 가지며, *SSHML*이 *SMB*와 상관성이 크지만 단순히 *SMB*의 변형이 아닌 독립적인 정보도 추가로 포함하고 있음을 시사한다.

<표 8>

**Ⅵ. 결론**

본 연구는 기존 공매도 관련 지표들이 과거 공매도 거래량에 의존하는 후행적 성격을 지닌다는 한계를 극복하기 위해 대차잔고 데이터를 활용하여 다음 달 공매도 거래량을 예측하는 선행적 지표를 개발하였다. 기존 연구에서 사용되던 *relss* 지표와, 대차잔고 변화율을 대표하는 *reldc* 지표에 릿지 회귀, 서포트 벡터 머신, 랜덤 포레스트 등 여러 가지 머신러닝 모형을 적용하여 높은 수익률 예측력을 가지는 최적의 를 구성하였다. 가장 뛰어난 성능을 보인 것은 선형 모형에 릿지 회귀분석을 적용한 것이다. 이 지표는 독자적 수익의 원천이 존재하는지 판단하는 Fama-French 4팩터 초과수익률 검증과, 횡단면 수익률에 대한 예측력을 검증하는 Fama-MacBeth 회귀분석에서 가장 월등한 유의성을 보였다. 이에 따라 수익률 예측에 유용하게 활용될 수 있을 것으로 예상한다.

또한, 해당 지표를 바탕으로 구성한 팩터(롱숏 포트폴리오 수익률)를 Fama-French 3팩터 모형에 추가했을 때 유의미한 결과를 확인했다. 본 연구는 해당 팩터를 기존 Fama-French 3팩터 모형에 추가하여, 각 팩터를 번갈아가며 종속변수로 설정한 뒤 모형의 설명력을 비교해보는 스패닝 테스트를 진행하였다. 결과를 통해 정의한 공매도 팩터(*SSHML*)가 *SMB* 팩터와 유사하지만 단순히 *SMB* 팩터의 변형이 아닌 고유한 팩터이며, 새로운 팩터로 추가될 경우 설명력의 개선을 이루어낼 수 있음을 밝혀냈다. 결론적으로 공매도 활동은 단순한 거래량 통계가 아닌 자산 가격 결정에서 중요한 위험 팩터로 기능하며, 공매도 기반 팩터가 추가된 위험 팩터 모형은 실질적인 가치를 제공할 수 있을 것으로 기대된다.

**References**

Asquith, P., Pathak, P. A. & Ritter, J. R. 2005. Short interest, institutional ownership, and stock returns. *Journal of Financial Economics,* 78**,** 243-276.

Bali, T. G., Cakici, N. & Whitelaw, R. F. 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics,* 99**,** 427-446.

Boehmer, E., Jones, C. M. & Zhang, X. 2008. Which Shorts Are Informed? *The Journal of Finance,* 63**,** 491-527.

Boehmer, E. & Wu, J. 2013. Short Selling and the Price Discovery Process. *The Review of Financial Studies,* 26**,** 287-322.

Chung, J. M. & Wang, S.-F. 2020. Short selling and stock price crash risk. *Journal of Derivatives and Quantitative Studies: 선물연구,* 28**,** 63-76.

Desai, H., Ramesh, K., Thiagarajan, S. R. & Balachandran, B. V. 2002. An Investigation of the Informational Role of Short Interest in the Nasdaq Market. *The Journal of Finance,* 57**,** 2263-2287.

Diether, K. B., Lee, K.-H. & Werner, I. M. 2009. Short-Sale Strategies and Return Predictability. *The Review of Financial Studies,* 22**,** 575-607.

Dong, D., Wu, K., Fang, J., Gozgor, G. & Yan, C. 2022. Investor attention factors and stock returns: Evidence from China. *Journal of International Financial Markets, Institutions and Money,* 77**,** 101499.

Fama, E. F. & MacBeth, J. D. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy,* 81**,** 607-636.

Gu, S., Kelly, B. & Xiu, D. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies,* 33**,** 2223-2273.

Hong, H. & Stein, J. C. 2003. Differences of Opinion, Short-Sales Constraints, and Market Crashes. *The Review of Financial Studies,* 16**,** 487-525.

Lee, D. K., Chun, D. & Cho, H. 2019. Short Interest and Market Risk Premium: The Case of the Korean Market. *Korean J Financ Stud,* 48**,** 541-566.

Lee, K.-H. & Wang, S.-F. 2019. Time-Varying Aggregate Short-Selling in Korea. *Asia-Pacific Journal of Financial Studies,* 48**,** 690-720.

Miller, E. M. 1977. Risk, Uncertainty, and Divergence of Opinion. *The Journal of Finance,* 32**,** 1151-1168.

Shleifer, A. & Vishny, R. W. 1997. The Limits of Arbitrage. *The Journal of Finance,* 52**,** 35-55.

Wang, S.-F. 2023. Return predictability of short-selling and financial distress firms: Evidence from Korean stock market. *Pacific-Basin Finance Journal,* 82**,** 102198.

Wang, S.-F. & Lee, K.-H. 2015. Do foreign short-sellers predict stock returns? Evidence from daily short-selling in Korean stock market. *Pacific-Basin Finance Journal,* 32**,** 56-75.

Wang, S.-F., Lee, K.-H. & Woo, M.-C. 2017. Do individual short-sellers make money? Evidence from Korea. *Journal of Banking & Finance,* 79**,** 159-172.

Woo, M. C. & Kim, M. A. 2017. Motivation of Short-Selling in KOSPI and KOSDAQ Market. *Korean J Financ Stud,* 46**,** 159-186.

이효정 2022. 공매도잔고와 공매제한. *재무관리연구,* 39**,** 71-101.

박주미, 정민수, 김홍선, 김성문.(2024).뉴스 감성 분석을 반영한 포트폴리오 선정 모형의 투자 성과 분석: 한국 주식시장을 중심으로.한국경영과학회지,49(4),57-72.

이익선.(2020).딥러닝기법과 최소자승법을 활용한 미국 주식시장 예측 연구.경영과학,37(2),19-31.

**<표 1> 변수 목록**

이 표는 분석을 위해 사용한 변수 목록을 정리한 것이다.

|  |  |
| --- | --- |
| **변수 이름** | **변수 정의** |
| *relss* | 상대적 공매도 거래량, 주식의 공매도 거래량을 전체 거래량으로 나눈 값 |
| *reldc* | 대차잔고 비율, 주식의 대차잔고 변화량을 전체 상장주식수로 나눈 값 |
|  | 대차잔고 비율을 반영하여 다음 달 공매도 거래량을 직관적으로 예측한 지표 |
|  | 식 (4) 기반 선형 회귀 모형으로 산출된 예측 지표 |
|  | 식 (5) 기반 선형 회귀 모형으로 산출된 예측 지표 |
|  | 식 (4) 기반 릿지 회귀 모형으로 산출된 예측 지표 |
|  | 식 (5) 기반 릿지 회귀 모형으로 산출된 예측 지표 |
|  | 식 (4) 기반 서포트 벡터 머신 모형으로 산출된 예측 지표 |
|  | 식 (5) 기반 서포트 벡터 머신 모형으로 산출된 예측 지표 |
|  | 식 (4) 기반 랜덤 포레스트 모형으로 산출된 예측 지표 |
| *ME* | 시가총액, 주식의 종가와 상장주식수의 곱 |
| *PRC* | 주식의 종가 |
| *BM* | 장부가-시장가 비율, 주식의 장부가를 시가총액으로 나눈 값 |

**<표 2> 요약 통계량**

이 표는 *relss*를 기준으로 구성한 포트폴리오의 요약 통계량을 나타낸다. rank가 1인 포트폴리오는 *relss*가 가장 낮은 포트폴리오, rank가 10인 포트폴리오는 *relss*가 가장 높은 포트폴리오이다. 여기에서 *ME*는 시가총액(억 원), *PRC*는 가격, *BM*은 장부가-시장가 비율을 의미한다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| rank | *relss* | *ME* | *PRC* | *BM* |
| 1 | 0.0002 | 1159 | 10315 | 1.200 |
| 2 | 0.0010 | 1287 | 10771 | 1.116 |
| 3 | 0.0022 | 1494 | 11276 | 1.054 |
| 4 | 0.0038 | 1811 | 12619 | 1.040 |
| 5 | 0.0061 | 2184 | 15892 | 1.022 |
| 6 | 0.0091 | 3120 | 18359 | 1.042 |
| 7 | 0.0134 | 4819 | 24504 | 1.039 |
| 8 | 0.0246 | 24988 | 39215 | 1.012 |
| 9 | 0.0456 | 29247 | 68051 | 1.008 |
| 10 | 0.0962 | 39371 | 105018 | 0.997 |

**<표 3> Fama-French 4팩터 알파 포트폴리오 분석**

이 표는 *relss*를 기준으로 포트폴리오를 구성했을 때의 초과수익률 및 Fama-French 4팩터 알파를 나타낸다. Panel A는 동일 가중으로 구성한 포트폴리오에 대한 결과이며, Panel B는 가치 가중으로 구성한 포트폴리오에 대한 결과이다. raw는 단순 수익률, capm, ff3, ff4는 각각 자산가격결정모형, Fama-French 3팩터 모형, 그리고 4팩터 모형에 대한 초과수익률과 *t*-통계량(아래에 괄호로 표시되어 있는 값)이다. 10-1은 포트폴리오 10을 매수하고 포트폴리오 1을 매도한 롱숏 포트폴리오의 수익률이다. \*\*\*, \*\*, \*는 각각 1%, 5%, 10% 유의수준에서의 통계적 유의성을 나타낸다.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 10-1 |
| Panel A: 동일 가중 포트폴리오 | | | | | | | | | | | |
| raw | 0.39 | 0.02 | 0.30 | 0.38 | 0.10 | -0.07 | 0.02 | -0.08 | -0.28 | -0.47 | -0.85\*\*\* |
|  | (0.70) | (0.04) | (0.62) | (0.85) | (0.27) | (-0.16) | (0.07) | (-0.19) | (-0.86) | (-1.64) | (-3.03) |
| capm | 0.61 | 0.26 | 0.54 | 0.63 | 0.34 | 0.17 | 0.26 | 0.16 | -0.05 | -0.25\* | -0.86\*\*\* |
|  | (1.24) | (0.61) | (1.40) | (1.59) | (1.08) | (0.54) | (0.97) | (0.56) | (-0.28) | (-1.69) | (-3.06) |
| ff3 | 0.08 | -0.25 | 0.03 | 0.12 | -0.10 | -0.31\*\* | -0.16 | -0.16 | -0.25\* | -0.38\*\*\* | -0.46\*\* |
|  | (0.26) | (-1.09) | (0.13) | (0.80) | (-0.74) | (-2.29) | (-1.24) | (-1.08) | (-1.88) | (-2.98) | (-2.16) |
| ff4 | 0.11 | -0.24 | 0.04 | 0.14 | -0.09 | -0.31\*\* | -0.14 | -0.16 | -0.24\* | -0.35\*\*\* | -0.45\*\* |
|  | (0.35) | (-1.05) | (0.19) | (0.99) | (-0.68) | (-2.17) | (-1.17) | (-1.06) | (-1.84) | (-2.89) | (-2.10) |
| Panel B: 가치 가중 포트폴리오 | | | | | | | | | | | |
| raw | -0.24 | -0.50 | -0.14 | -0.42 | -0.34 | -0.42 | -0.52 | -0.47 | -0.01 | -0.50 | -0.26 |
|  | (-0.51) | (-1.10) | (-0.29) | (-0.88) | (-0.82) | (-0.92) | (-1.25) | (-1.38) | (-0.04) | (-1.45) | (-0.83) |
| capm | -0.02 | -0.26 | 0.08 | -0.18 | -0.11 | -0.21 | -0.31 | -0.27 | 0.19 | -0.28\* | -0.25 |
|  | (-0.05) | (-0.65) | (0.23) | (-0.49) | (-0.31) | (-0.66) | (-1.05) | (-1.27) | (0.88) | (-1.85) | (-0.80) |
| ff3 | -0.47\* | -0.68\*\*\* | -0.37\* | -0.54\*\* | -0.42 | -0.51\*\*\* | -0.50\* | -0.29\* | 0.22 | -0.22 | 0.25 |
|  | (-1.94) | (-3.17) | (-1.74) | (-2.38) | (-1.62) | (-2.09) | (-1.87) | (-1.67) | (1.02) | (-1.39) | (1.16) |
| ff4 | -0.47\* | -0.70\*\*\* | -0.38\* | -0.56\*\*\* | -0.43\* | -0.51\*\* | -0.49\* | -0.29\* | 0.19 | -0.21 | 0.26 |
|  | (-1.95) | (-3.07) | (-1.79) | (-2.63) | (-1.68) | (-2.17) | (-1.84) | (-1.73) | (0.81) | (-1.31) | (1.21) |

**<표 4> Fama-MacBeth 회귀분석**

이 표는 Fama-MacBeth 회귀분석에 대한 결과이다. Model 1~5까지의 결과는 수익률과 각 지표에 대한 단순선형회귀의 계수이며, *t*-통계량을 괄호 안에 나타냈다. Model 6는 모든 지표와 다음 달 수익률 간의 다중선형회귀분석의 결과이며, 계수와 *t*-통계량을 나타냈다. \*\*\*, \*\*, \*는 각각 1%, 5%, 10% 유의수준에서의 통계적 유의성을 나타낸다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
| *relss* | -8.031\*\* |  |  |  |  | -13.031\*\*\* |
|  | (-2.20) |  |  |  |  | (-3.94) |
| *BETA* |  | -0.626\*\*\* |  |  |  | -0.626\*\*\* |
|  |  | (-2.87) |  |  |  | (-2.95) |
| *ME* |  |  | -0.000 |  |  | -0.000\*\*\* |
|  |  |  | (-1.27) |  |  | (-3.09) |
| *PRC* |  |  |  | 0.000\*\*\* |  | 0.000\*\*\* |
|  |  |  |  | (3.27) |  | (5.33) |
| *BM* |  |  |  |  | -0.001\*\*\* | -0.001\*\*\* |
|  |  |  |  |  | (-4.30) | (-4.56) |

**<표 5> Fama-French 4팩터 알파 포트폴리오 분석: 머신러닝 방법론 적용**

이 표는 머신러닝 방법론을 적용한 에 대한 포트폴리오 분석의 결과이다. OLS는 선형 회귀, Ridge는 릿지 회귀, SVM은 서포트 벡터 머신, RF는 랜덤 포레스트 모형이며 지수 관계를 반영한 모형은 ‘(지수)’로 표시하였다. 각 결과값은 개별 지표로 구성한 10-1 롱숏 포트폴리오의 수익률에 대한 Fama-French 4팩터 알파와, 이에 대한 *t*-통계량을 나타낸다. Panel A는 동일 가중, Panel B는 가치 가중 포트폴리오에 대한 결과이다. \*\*\*, \*\*, \*는 각각 1%, 5%, 10% 유의수준에서의 통계적 유의성을 나타낸다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 기존 결과 | OLS | OLS (지수) | Ridge | Ridge (지수) | SVM | SVM (지수) | RF |
| Panel A: 동일 가중 포트폴리오 | | | | | | | | |
| alpha | -0.36 | -0.70\*\* | -0.64\*\* | -0.79\*\*\* | -0.64\*\* | -0.22 | -0.71\*\*\* | -0.37\*\* |
| *t* | (-1.57) | (-2.47) | (-2.47) | (-3.13) | (-2.44) | (-1.47) | (-2.71) | (-2.04) |
| Panel B: 가치 가중 포트폴리오 | | | | | | | | |
| alpha | 0.06 | -0.05 | -0.05 | -0.30 | -0.05 | -0.46\* | -0.09 | 0.30 |
| *t* | (0.26) | (-0.22) | (-0.22) | (-1.22) | (-0.22) | (-1.75) | (-0.37) | (1.48) |

**<표 6> Fama-MacBeth 회귀분석: 머신러닝 방법론 적용**

이 표는 머신러닝 방법론을 적용한 에 대한 Fama-MacBeth 회귀분석의 결과이다. OLS는 선형 회귀, Ridge는 릿지 회귀, SVM은 서포트 벡터 머신, RF는 랜덤 포레스트 모형이며 지수 관계를 반영한 모형은 ‘(지수)’로 표시하였다. Panel A는 단순선형회귀, 즉 *ret* ~ 모형에 대한 계수 및 *t*-통계량(아래에 괄호로 표시되어 있는 값)이며, Panel B는 다중선형회귀, 즉 *ret* ~ + *BETA* + *ME* + *PRC* + *BM*에서의 의 계수 및 *t*-통계량을 나타낸다. \*\*\*, \*\*, \*는 각각 1%, 5%, 10% 유의수준에서의 통계적 유의성을 나타낸다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 기존 결과 | OLS | OLS (지수) | Ridge | Ridge (지수) | SVM | SVM (지수) | RF |
| Panel A: 개별 모형 | | | | | | | | |
| coef. | -0.281\*\*\* | -0.257\*\* | -0.304\*\* | -0.256\*\* | -0.303\*\* | -0.127\*\* | -0.296\*\* | -0.208\*\* |
| *t* | (-2.58) | (-2.14) | (-2.17) | (-2.09) | (-2.16) | (-2.35) | (-2.17) | (-1.97) |
| Panel B: 복합 모형 | | | | | | | | |
| coef. | -0.188\*\* | -0.419\*\*\* | -0.458\*\*\* | -0.394\*\*\* | -0.456\*\*\* | -0.169\*\*\* | -0.451\*\*\* | -0.315\*\*\* |
| *t* | (-2.23) | (-4.67) | (-3.71) | (-3.95) | (-3.70) | (-4.11) | (-3.78) | (-3.91) |

**<표 7> 요약 통계량: 머신러닝 방법론 적용**

이 표는 를 기준으로 구성한 포트폴리오의 요약 통계량을 나타낸다. rank가 1인 포트폴리오는가 가장 낮은 포트폴리오, rank가 10인 포트폴리오는 가 가장 높은 포트폴리오이다. 여기에서 *ME*는 시가총액(억 원), *PRC*는 가격, *BM*은 장부가-시장가 비율을 의미한다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| rank | *relss* | *ME* | *PRC* | *BM* |
| 1 | 0.0025 | 1259 | 10942 | 1.190 |
| 2 | 0.0040 | 1467 | 11533 | 1.067 |
| 3 | 0.0056 | 1681 | 12287 | 1.016 |
| 4 | 0.0075 | 2101 | 14367 | 1.017 |
| 5 | 0.0099 | 2669 | 17268 | 1.001 |
| 6 | 0.0126 | 4607 | 22309 | 1.031 |
| 7 | 0.0167 | 7083 | 28341 | 1.033 |
| 8 | 0.0313 | 28945 | 45893 | 0.995 |
| 9 | 0.0491 | 32001 | 74829 | 1.017 |
| 10 | 0.0856 | 40913 | 108147 | 1.000 |

**<표 8> 스패닝 테스트**

이 표는 스패닝 테스트 결과이며, 각 모형에 대한 선형 회귀 분석 결과를 나타낸다. 가장 왼쪽 열에 적힌 팩터들이 좌변에 있는 변수이며, 각 열에 적힌 변수들 아래 값들은 회귀 분석 계수와 *t*-통계량(괄호)을 의미한다. 가장 오른쪽 열은 모형에 대한 결정계수() 값이다. \*\*\*, \*\*, \*는 각각 1%, 5%, 10% 유의수준에서의 통계적 유의성을 나타낸다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Intercept | *SSHML* | *SMB* | *HML* | *MKTRF* |  |
| *SSHML* | -0.004 |  | -0.626\*\*\* | -0.101 | -0.074 | 0.375 |
|  | (-1.60) |  | (-8.81) | (-1.55) | (-1.28) |  |
| *SMB* | 0.002 | -0.558\*\*\* |  | -0.034 | 0.040 | 0.364 |
|  | (0.74) | (-10.55) |  | (-0.51) | (0.68) |  |
| *HML* | 0.002 | -0.139 | -0.053 |  | -0.110 | 0.029 |
|  | (0.79) | (-1.36) | (-0.52) |  | (-1.21) |  |
| *MKTRF* | -0.004 | -0.153 | 0.093 | -0.166 |  | 0.049 |
|  | (-1.10) | (-1.26) | (0.67) | (-1.08) |  |  |

**Machine Learning-Based Short-Selling Indicators and the Cross-Section of Stock Returns**

Danyeol Jung[[11]](#footnote-11), Gahyeon Lee[[12]](#footnote-12), Seohyeon Jo[[13]](#footnote-13), Minhyuk Lee[[14]](#footnote-14), and Jihoon Goh[[15]](#footnote-15)

**Abstract**

This study examines the impact of short-selling volume on future stock returns in the Korean stock market and proposes new short-selling indicators using machine learning techniques. Stocks with higher relative short-selling volume tend to exhibit significantly lower future returns, and this negative predictive power is confirmed through portfolio analysis and Fama-MacBeth regressions. Noting the limitation that existing indicators rely solely on past short-selling volume and thus have a lagging nature, this study develops leading indicators by predicting next month’s short-selling volume using stock loan balance data. In this process, various machine learning models are compared, and the ridge regression-based linear model demonstrates the best predictive performance. By presenting new predictive indicators derived from short-selling data, this study complements the existing literature and offers practical implications for investment strategies.

Keywords: Short Selling, Stock Returns, Anomalies, Risk Factors, Machine Learning

1. \* 이 과제는 2025년 부산대학교 특성화사업에 의하여 연구되었음.

   제1저자, 한국과학영재학교, 부산, 대한민국; Email: 23-101@ksa.hs.kr [↑](#footnote-ref-1)
2. 공동저자, 한국과학영재학교, 부산, 대한민국; Email: 23-079@ksa.hs.kr [↑](#footnote-ref-2)
3. 공동저자, 한국과학영재학교, 부산, 대한민국; Email: 23-107@ksa.hs.kr [↑](#footnote-ref-3)
4. 공동저자, 경영학과, 부산대학교, 부산, 대한민국; Email: minhyuk.lee@pusan.ac.kr [↑](#footnote-ref-4)
5. 교신저자, 경영학과, 부산대학교, 부산, 대한민국; Tel: 051-510-2566;   
   Email: jihoongoh@pusan.ac.kr [↑](#footnote-ref-5)
6. 한국 시장의 공매도 데이터를 활용한 연구는 Woo and Kim (2017), Lee and Wang (2019), Lee, Chun, and Cho (2019), Chung and Wang (2020), 이효정 (2022) 등이 있다. [↑](#footnote-ref-6)
7. 추가적으로 KOSPI 시장과 KOSDAQ 시장을 구분하여 예측력을 비교한 결과, KOSPI 시장에 상장된 주식으로 구성한 *relss* 롱숏 포트폴리오는 평균 초과수익률이 -1.03%(*t*-통계량 -3.05), KOSDAQ 시장에서는 -0.74%(*t*-통계량 -2.37)로 나타났다. 두 시장 모두 통계적으로 유의한 음의 수익률을 보이며, 이는 공매도 지표의 예측력이 두 시장 모두에서 일관되게 작용하고 있음을 시사한다. [↑](#footnote-ref-7)
8. Bali, Cakici, and Whitelaw (2011)의 정의를 따른다. [↑](#footnote-ref-8)
9. 박주미, 정민수, 김홍선, 김성문 (2024), 이익선 (2020) 등은 머신러닝 방법론을 주식 시장 분석에 활용하였다. [↑](#footnote-ref-9)
10. L2 정규화는 회귀계수의 제곱합을 손실함수에 패널티 항으로 추가하는 방식으로, 계수의 크기를 줄여 과적합을 방지하고 예측 성능을 향상시키는 역할을 한다. [↑](#footnote-ref-10)
11. \* This work was supported by 2025 Specialization Project of Pusan National University.

    Korea Science Academy, Busan, Republic of Korea; Email: 23-101@ksa.hs.kr [↑](#footnote-ref-11)
12. Korea Science Academy, Busan, Republic of Korea; Email: 23-079@ksa.hs.kr [↑](#footnote-ref-12)
13. Korea Science Academy, Busan, Republic of Korea; Email: 23-107@ksa.hs.kr [↑](#footnote-ref-13)
14. Department of Business Administration, Pusan National University, Busan, Republic of Korea; Email: minhyuk.lee@pusan.ac.kr [↑](#footnote-ref-14)
15. Department of Business Administration, Pusan National University, Busan, Republic of Korea; Tel: +82 515102566; Email: jihoongoh@pusan.ac.kr [↑](#footnote-ref-15)