Bimodal Characteristic Returns and Predictability Enhancement via Machine Learning

Chulwoo Han*

Abstract

This paper documents the bimodality of momentum stocks: both high- and low-momentum stocks have nontrivial probabilities for both high and low returns. The bimodality makes the momentum strategy fundamentally risky and can cause a large loss. To alleviate the bimodality and improve return predictability, this paper develops a novel cross-sectional prediction model via machine learning. By reclassifying stocks based on their predicted financial performance, the model significantly outperforms off-the-shelf machine learning models. Tested on the US market, a value-weighted long-short portfolio earns a monthly alpha of 2.4% (t-statistic = 6.63) when regressed against the Fama-French five factors plus the momentum and short-term reversal factors.

Keywords: Bimodality; deep momentum; machine learning; deep neural network; reclassification.

JEL Classification: G11, G12.

1 Introduction

1.1 Motivation

Price momentum (Jegadeesh and Titman, 1993) is perhaps the most persistent anomaly. While the majority of firm characteristics known to possess predictive power have been revealed to be insignificant in recent studies, *e.g.*, Green et al. (2017); Freyberger et al. (2020), price momentum remains significant. Moreover, recent asset pricing studies employing machine learning find that price momentum is among the most important features (Messmer, 2017; Gu et al., 2020). With the forecasting power of price momentum, a momentum strategy that buys past winners and sells past losers can generate a positive profit. However, it can also experience a significant loss when the market rebounds from a recession, which is known as a momentum crash (Daniel and Moskowitz, 2016).

*Chulwoo Han is with Durham University. Durham Business School, Mill Hill Lane, Durham, DH1 3LB UK; Tel: +44 1913345892; E-mail: chulwoo.han@durham.ac.uk.

A little-known fact about momentum is that high- and low-momentum stocks are cross-sectionally more dispersed and have distinct bimodal cross-sectional distributions of relative return, making the momentum strategy fundamentally risky. Investors who follow a momentum strategy anticipate high-momentum stocks to generate high future returns and low-momentum stocks to lead to low future returns. If this is the case, when stocks are double-sorted independently on momentum and one-month ahead return, high-momentum stocks will tend to belong to high-return quantiles, whereas low-momentum stocks will belong to low-return quantiles, resulting in cross-sectional relative return distributions similar to those shown at the top of Figure 1. In the figure, the horizontal axis represents the future return deciles (high to low), and the vertical axis represents stocks' probability for each decile. 'Hypothetical (H)' and 'Hypothetical (L)' respectively depict the hypothetical distributions of high- and low-momentum stocks.

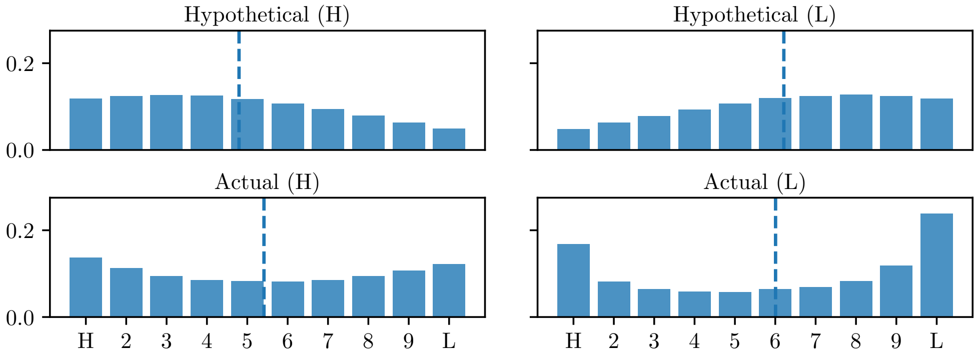


Figure 1: The cross-sectional distribution of momentum stocks: hypothetical vs. actual

This figure compares high- and low-momentum stocks' hypothetical cross-sectional distributions (top panel) with actual distributions (bottom panel). Stocks are double-sorted independently on momentum and one-month ahead return, and the probability mass function is drawn within each momentum decile. The hypothetical distributions are what investors may expect, whereas the actual distributions are obtained from the US stocks for the period spanning 1955.01 to 2017.01. The horizontal axis denotes the future return class (in descending order), and the dotted line represents the average return class.

In contrast, the actual returns of high- and low-momentum stocks form U-shaped bimodal cross-sectional distributions as shown at the bottom of Figure 1. The graphs present the time-series averages of the cross-sectional distributions obtained by double-sorting US stocks on the one-year price momentum and one-month ahead return over the period from 1955 to 2016. 'Actual (H)' and 'Actual(L)' respectively represent the distributions of the highest- and lowest-momentum stocks.

Although past winners are most likely to belong to the highest-return group (H), their likelihood of belonging to the lowest-return group (L) is the second-highest. Likewise, past losers are most likely to belong to the lowest-return group, but their likelihood of belonging to the highest-return group is also significant. The bimodality is more evident among past losers. The bimodality indicates that both past winners and losers are likely to yield either significantly high or low returns and, therefore, are riskier than other stocks.

As detailed in Section 2, the bimodal distribution is not a result of the combination of unimodal distributions at different times: the bimodality is present in many short periods in the sample. Besides, this phenomenon is prevalent when stocks are sorted on past returns of different periods or other firm characteristics. It is also found that the distributions tend to tilt during a recovery period, causing a reversal of probabilities, especially at the low end of momentum. This causes momentum crashes. The bimodality implies that there is a nonlinear relationship between momentum and future returns.

Despite the bimodality, momentum does appear to carry information about future returns and if the nonlinear relationship could be identified, future returns would become more predictable. To address the nonlinearity, one can employ a machine learning method. Machine learning can learn a nonlinear relationship from data, and it offers many classification algorithms that are suitable for stock classification. The empirical analysis in this paper reveals that neural network models using momentum features as input outperform the standard momentum strategy. Nevertheless, they fail to eliminate bimodality: when stocks are classified into return deciles via a neural network, the stocks in the top (highest return) and bottom (lowest return) deciles exhibit bimodality, *i.e.*, have a high probability for the opposite extreme decile.

Fortunately, the neural network models predict the bimodality, although, to my knowledge, this information has never been exploited. To illustrate this point, the reader needs to understand how a neural network classifies stocks. A neural network estimates a stock's probabilities for the return classes and chooses the class with the highest probability as the predicted class. While these probabilities form an estimate of the stock's relative return distribution and provide richer information than the predicted class alone, researchers and practitioners have overlooked them and have made decisions solely based on the predicted class. This practice is not ideal when the estimated return distribution is bimodal, in which case, the predicted class will not reflect the expected return of the stock. The empirical analysis finds that the stocks classified into extreme deciles have predicted probabilities that are bimodal and closely reflect the actual return distributions. That is, a better financial decision can be made by exploiting the entire probability distributions as opposed to the predicted class.

To illuminate the point above, suppose an investor uses a neural network model to classify stocks into high (H), medium (M), and low (L) return classes, and the model predicts stock A's probability for these classes to be 55%, 0%, and 45%, and stock B's probabilities to be 45%, 55%, and 0%. The model will then classify A into H and B into M, and a naïve investor will invest in stock A. However, it is obvious that stock B is a better choice in terms of both mean return and risk, which the investor would have recognized had they paid attention to the probabilities.

1.2 Aim of the Paper

This paper aims to develop a new machine learning-based prediction model for the cross-section of asset returns, using momentum as the primary source of information. A crucial difference from the extant machine learning asset pricing models in the literature is that the proposed model substantially enhances its prediction power by reconciling the aforementioned discrepancy between machine learning prediction and financial performance. The model employs machine learning particularly a deep neural network (DNN), to estimate the return distribution of each stock and predict its financial performance by combining the distributional information with other information. It then ranks the stocks on their predicted financial performance via a process called *reclassification.* This process alleviates the bimodality and renders a significantly superior portfolio performance compared to conventional machine learning models. This modeling framework is henceforth termed Deep Momentum (DM).

Although the DM model uses momentum as input, it is not a mere extension of the momentum strategy. The DM model utilizes the nonlinear information hidden in the momentum features and develops a completely new strategy at the stock level. This approach is distinct from most studies on momentum, which either apply the standard momentum strategy to different asset classes or markets or are a minor variation of the strategy. This paper is not about finding new features but about developing a new methodology that utilizes nonlinear information from existing features.

This paper also addresses another shortcoming of machine learning in the context of portfolio construction. A long-short portfolio strategy derived from machine learning classification faces a data imbalance problem as the number of stocks classified into the highest-return class can be considerably different from the number in the lowest-return class. It is found that machine learning models tend to allocate more stocks to the bottom (lowest-return) class, making the strategy difficult to implement. The reclassification procedure naturally resolves the data imbalance problem by redistributing stocks.

This paper develops five reclassification methods. The first two methods are mainly concerned with redistributing stocks to ensure they are uniformly distributed across classes. The next two methods predict the expected (relative) returns of stocks and use them to reclassify the stocks. One method uses the difference between the probability of high returns and the probability of low returns as a proxy for the expected return. The other method derives stocks' expected returns from the expected returns of the return classes. The last method estimates the Sharpe ratio and uses it as a reclassification criterion. This method is particularly relevant to investors who face short-sale constraints.

1.3 Main Findings

The DM models are tested on the stocks in the US market over the period spanning 1975.01 to 2017.01. When stocks are reclassified based on their predicted mean returns, a value-weighted long-short portfolio that buys the top 10% stocks and sells the bottom 10% stocks earns an annualized mean return above 25% and a Sharpe ratio as high as 1.34. An equal-weighted long-short portfolio performs better and earns an annualized mean return of 40% and a Sharpe ratio as high as 2.57. Remarkably, these long-short strategies do not suffer from momentum crashes, although they use only momentum features and do not explicitly address momentum crashes.

When a size variable is added to the model, both value-weighted and equal-weighted long-short portfolios perform significantly better. With the size variable, the value-weighted portfolio yields an annualized mean return of 35% and a Sharpe ratio above 1.6, and the equal-weighted portfolio yields an annualized mean return around 50% and a Sharpe ratio above 2.8. In contrast, the standard momentum strategy yields a Sharpe ratio of 0.61 (value-weighted) or 0.37 (equal-weighted), and Barroso and Santa-Clara (2015)'s volatility-adjusted momentum strategy yields a Sharpe ratio of 1.00 (value-weighted) or 0.80 (equal-weighted).

A factor model involving Fama and French (2015)'s five factors and the momentum and short-term reversal factors cannot explain the returns of the long-short portfolios. The monthly alpha is 2.4% (t-statistic = 6.63) for the value-weighted portfolio and 3.5% (t-statistic = 9.28) for the equal-weighted portfolio.

The DM strategies remain profitable even when transaction costs are taken into account. The DM strategies incur twice as much turnover (166%) as the momentum strategy (78%), but still earn a higher return: *e.g.*, a value-weighted long-short portfolio earns an annualized mean return of 31% and a Sharpe ratio of 1.46 when subject to 10 basis point transaction costs, and 23% and 1.09 when subject to 30 basis points. These numbers are higher than those of the volatility-adjusted momentum strategy without transaction costs. When more conservative transaction costs are assumed following DeMiguel et al. (2020), the value-weighted long-short portfolio loses most of its profit and has a mean return of 4.8%, but the long-only portfolio remains profitable with a mean excess return of 8.3% and a Sharpe ratio of 0.34. The equal-weighted long-short portfolio is more immune to the transaction costs and earns a mean return of 19.6% and a Sharpe ratio of 1.06 under the most conservative assumption. These results suggest that the portfolios derived from the DM strategy will remain profitable even under a reasonably high transaction cost.

The DM model significantly outperforms a logistic classifier and a linear regressor, suggesting that there is a nonlinear relationship between input features and the future return. A neural network-based ordinal regressor performs comparably to the DM model without reclassification, but it underperforms the DM model augmented with reclassification. This result reveals that the reclassification procedure is key to the superior performance of the strategy.

Experiments with methods to further enhance the performance witness even an annualized mean return of 82% with a Sharpe ratio of 1.88 for the value-weighted long-short portfolio. This tremendous result is achieved when the long-short portfolio is constructed using the top and bottom 1% of the stocks. Compared to the mean return, the Sharpe ratio is found to be more difficult to improve.

The framework can be extended to estimate the covariance matrix of individual stocks. A stock-level portfolio optimization based on the DM model yields an annualized mean return of 59% and a Sharpe ratio of 3.29, which are higher than those from the equal-weighted long-short portfolio.

The profits of the DM strategies are not driven by small firms. The average size of the firms in the long portfolio is larger than the market average and that of the short portfolio is only moderately smaller than the market average. Even the stocks in the top/bottom 1% are larger than those in the top/bottom deciles of the momentum strategy. Excluding small firms reduces the mean return and the Sharpe ratio, but the DM strategy remains profitable with a Sharpe ratio of 1.08 when stocks below the 20% NYSE-size quantile are excluded.

The DM strategies do not time the market aggressively. The model parameters are updated only once a year and remain stable over the test period. A sensitivity analysis suggests that the superior performance of the DM model results from the identification of the nonlinear interaction between the input features. It also reveals that the short-term reversal, size, and overall market return are among the important variables.

1.4 Contribution of the Paper

This paper makes multiple contributions to the literature. It extends the asset pricing literature by introducing a new machine learning-based return prediction model and demonstrating its superior predictive power and profitability. Unlike many extant machine learning asset pricing studies that horse-race different off-the-shelf models, this paper identifies a problem, *i.e.*, the bimodality of momentum stocks, and carefully designs a model to resolve it, employing machine learning where necessary. More specifically, by reclassifying stocks based on their predicted financial performance, the proposed model significantly improves predictability and portfolio performance over a conventional machine learning model. As a result, the model involves only a small number of features, yet the derived portfolio strategy outperforms most strategies reported in the literature. With infinite choices of models offered by machine learning, it is crucial to identify the problem first and design the model accordingly.

The discovery of the bimodality of momentum stocks itself is a significant contribution. It provides a new stock-level perspective on the riskiness of the momentum strategy. As shown in the next section, the bimodality is present in many firm characteristics that are known to have predictive power, such as beta, book-to-market ratio, and size. There appears to be a positive relationship between the bimodality (risk) and return predictability: a group of stocks with higher return predictability, *e.g.*, low-momentum stocks, tend to be cross-sectionally more dispersed, regardless of whether the predicted return is positive or negative. A further investigation into this finding is beyond the scope of this paper but can be an interesting subject for future research.

This paper also highlights the importance of accounting for financial performance in a machine learning-based asset pricing model when the objective of machine learning is not aligned with the investment objective. The objective of a neural network classifier is to predict the most probable return (class), whereas the objective from an investor's perspective should be to predict (the class that corresponds to) the expected return. This paper develops the reclassification procedure to reconcile this discrepancy. Reclassifying stocks based on their predicted financial performance significantly improves portfolio performance compared to a standard neural network classifier.

The empirical fact that an exceedingly high Sharpe ratio can be achieved using only previous returns has a significant implication for future research as it raises the question of which one is more important between features and models. In the traditional asset pricing framework, it is mainly the firm characteristics that matter, but with machine learning playing an increasingly important role in asset pricing, the model itself provides a considerable room to enhance stock return predictability. This paper focuses on modeling rather than identifying important features and provides a new research direction.

Machine learning allows the development of a complex model involving a large number of input features and an arbitrary nonlinear relationship between the input and the output. This flexibility, however, is accompanied by the risk of overfitting. The beauty of the proposed model is in its simplicity. With a small number of input features, it performs robustly while benefiting from the power of machine learning.

1.5 Related Literature

This paper is closely related to momentum as it provides a new stock-level perspective on the riskiness of the momentum strategy and develops a new prediction model based on momentum. Multiple papers document the tail risk of momentum and propose methods to manage it. Daniel and Moskowitz (2016) document momentum crashes and analyze the characteristics of momentum around crashes. They develop a dynamic momentum strategy based on forecasts of momentum's mean and variance, which doubles the alpha and Sharpe ratio of a static strategy. Barroso and Santa-Clara (2015) estimate the risk of momentum by the realized variance of daily returns and show that managing this risk can eliminate most crashes and double the Sharpe ratio. Geczy and Samonov (2016) examine the pre-1927 US stock market and conclude that momentum profits remain positive and significant in this period. They also develop a dynamically hedged momentum strategy that significantly outperforms an unhedged strategy. Daniel et al. (2019) develop a two-state hidden Markov model wherein the process driving market returns transitions between turbulent and calm states; the authors argue that a momentum timing strategy based on this model avoids momentum crashes and achieves superior out-of-sample risk-adjusted performance. Butt and Virk (2019) argue that the variation in market liquidity is an important determinant of momentum crashes and explains the forecasting ability of known predictors for the tail risk of momentum.

While these papers take the static momentum strategy for granted and enhance investment performance by dynamically controlling the exposure to the strategy, this paper uses momentum to predict the financial performance of individual stocks and develops an entirely new strategy at the stock level.

This paper is also related to the proliferating literature on machine learning for asset pricing. Since the resurgence of machine learning and its successful application in various areas, numerous papers on asset return forecast have been published in computer science and operations research journals, and the number is increasing rapidly: Atsalakis and Valavanis (2009); Khashei and Bijari

(2010); Liao and Wang (2010); Wang et al. (2011); Khashei and Bijari (2011); Dai et al. (2012); Kim and Ahn (2012); de Oliveira et al. (2013); Kazem et al. (2013); Adebiyi et al. (2014); Patel et al. (2015); Wang and Wang (2015); Chong et al. (2017) to name a few. A recent survey can be found in Rather et al. (2017).

While earlier studies led by scholars from computer science and engineering focus on prediction itself, recent studies by scholars from the finance discipline place more emphasis on asset pricing perspectives. Moritz and Zimmermann (2016) employ a tree-based method to classify stocks and show that their model outperforms traditional factor-based methods. Messmer (2017) employs deep learning with 68 firm characteristics to predict stock returns and finds significant return predictability. Using 94 firm characteristics, 8 macroeconomic variables, and 74 industry dummies, Gu et al. (2020) compare multiple machine learning methodologies for stock return prediction and conclude that neural network models perform best. Feng et al. (2018) utilize machine learning to generate a "long-short" factor and find that it helps explain return anomalies. Bianchi et al. (2020) borrow the research framework of Gu et al. (2020) to test bond return predictability and, consistent with Gu et al. (2020), conclude that a deep neural network (DNN) does improve out-of-sample predicting performance. Gu et al. (2019) propose a new nonlinear latent factor conditional asset pricing model and show that their model delivers out-of-sample pricing errors that are far smaller compared to other leading factor models. Freyberger et al. (2020) use the adaptive group LASSO to study which characteristics provide incremental information for the cross-section of expected returns. They find many of the previously identified return predictors do not provide incremental information and only about ten characteristics have predictive power.

This paper also employs machine learning (deep neural network) to forecast returns and develop portfolio strategies. However, it does not merely rely on an off-the-shelf model but enhances return predictability by addressing the shortcomings of conventional machine learning applications.

2 Bimodal Characteristic Returns

2.1 Cross-Sectional Distribution of Momentum Stock Returns

This section explores the cross-sectional distribution of stock returns conditional on momentum. The stocks in the US market for the period from 1955.01 to 2017.01 are used for all the empirical analyses. Stocks are divided into ten momentum groups (high (H) to low (L)) based on the one-year price momentum. Then, they are labeled 1 (H) to 10 (L) based on the one-month ahead return (in descending order), and a probability mass function is generated for each momentum group. This probability mass function is called a *cross-sectional relative return distribution.* If the price momentum has a forecasting power for future returns, high-momentum stocks will yield high returns and have more masses in high-return groups, whereas low-momentum stocks yield low returns and have more masses in low-return groups.

Before presenting the empirical results, Figure 2 illustrates the relationship between the usual cross-sectional return distribution and the relative return distribution. The light-blue curve in the upper graph of each panel represents the cross-sectional return distribution of a population (*e.g.*, all stocks in a given month), which is assumed to be normal with the mean of 0.1 and standard deviation of 0.2. By definition, the population has a uniform relative return distribution, as described by the light-blue bars in the lower graphs, where the horizontal axis denotes the future return group (high to low). The blue curves represent the cross-sectional return distribution of a sample (*e.g.*, high-momentum stocks in the same month), which is assumed to have the mean and standard deviation given in the title of each graph. The corresponding relative return distributions are depicted by the blue bars in the lower graphs.

The graphs at the top assume that the sample has a different mean but the same variance as the population, whereas those at the bottom assume it has the same mean but different variance. The graphs show that if the sample has a higher (lower) mean, the relative return distribution will slope downward (upward) from left to right. On the other hand, if the sample has a larger (smaller) variance, the relative return distribution will have a "U" (inverted "U") shape.

Figure 3 presents the results from the US market, wherein each graph shows the relative return distribution of each momentum group. The figure reveals the clear bimodality of the stocks in extreme deciles. Bimodality is particularly evident among the lowest-momentum stocks. In contrast, stocks in in-between deciles generally exhibit an inverted U-shaped distribution. This implies that the stocks in the extreme deciles are more dispersed and are therefore more likely to have extreme returns. While high-momentum (H) stocks are more likely to belong to the highest return group and low-momentum (L) stocks to the lowest return group, their likelihood of belonging to the opposite group is also substantial.

Although a large cross-sectional variance does not necessarily imply a large time-series variance of the portfolio returns, it is revealed to be positively correlated. The momentum groups H, L, and the rest, respectively, have a time-series mean of the cross-sectional variance equal to 0.139, 0.205, and 0.117, and a time-series variance of the portfolio returns equal to 0.067, 0.090, and 0.051. This result suggests that the higher volatilities of the high- and low-momentum portfolios are partly due to the significant cross-sectional variances of their constituents.

The bimodality is not a result of the time-varying performance of momentum portfolios. An investigation of yearly distributions during the sample period reveals that bimodality is present in most years. The bimodality is also not a unique phenomenon associated with the one-year price momentum. Figure 4 presents the relative return distributions of the high- and low-momentum stocks formed on different period price momentums, including short-term reversal. Surprisingly, bimodality is prevalent in all momentum features. It is interesting to note how the distribution transforms its shape and the momentum effect shifts to the short-term reversal as the momentum period becomes shorter.

Figure 6 examines the structural change in the distribution during momentum crashes. Three periods of recession-recovery are defined based on the shape of the low-momentum portfolio's

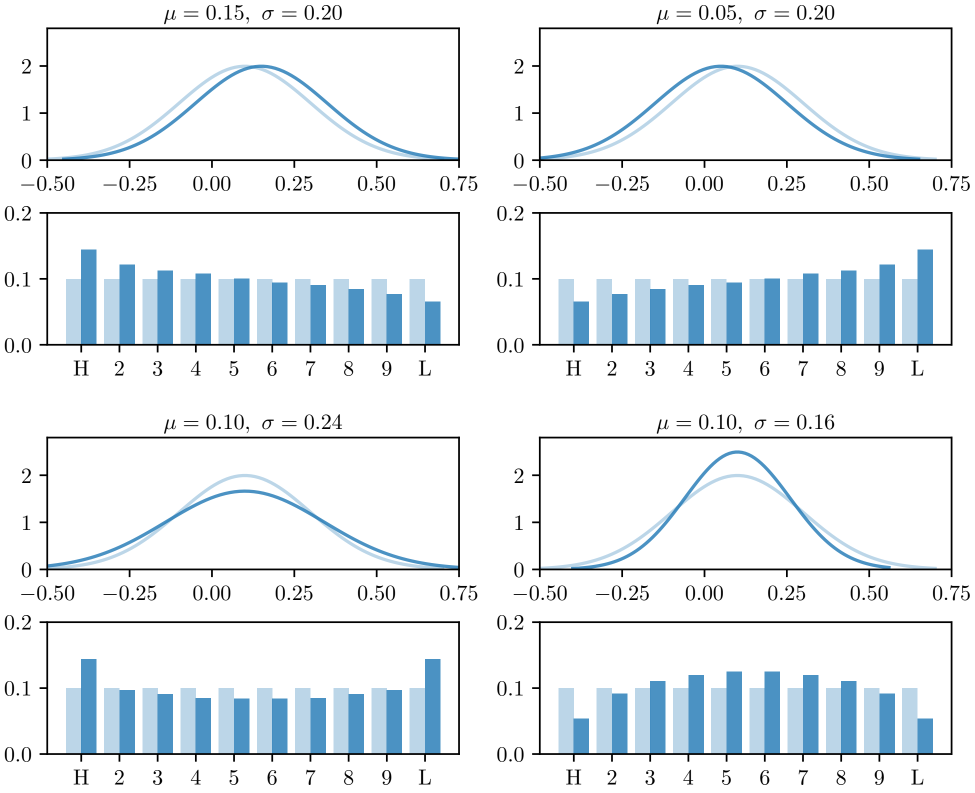


Figure 2: Return distribution vs. relative return distribution

This figure compares cross-sectional return distributions with their corresponding relative return distributions. The light-blue lines in the upper graphs represent the cross-sectional return distribution of the population (assumed $N(0.1, 0.2^2)$), and the blue lines represent the distribution of a sample with the mean and standard deviation given in the title of each graph. The corresponding relative return distributions are presented in the lower graphs using the same color, whereby the horizontal axis denotes the return group (in descending order).

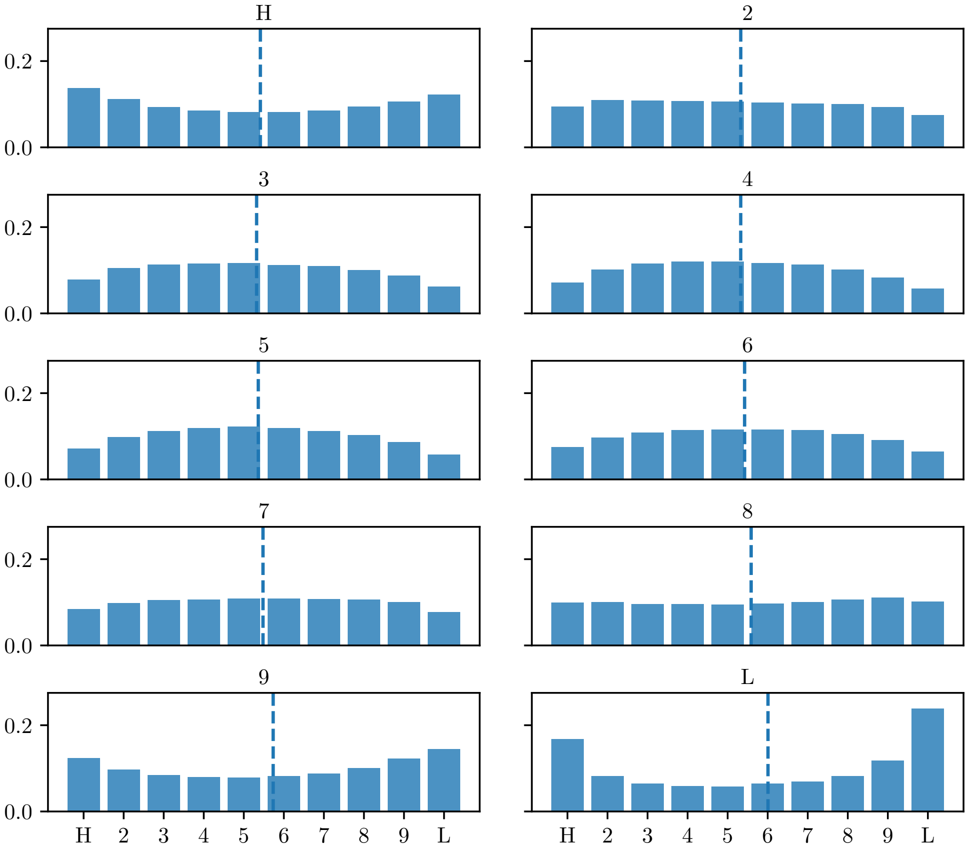


Figure 3: Cross-sectional relative return distribution conditional on momentum

This figure presents the cross-sectional relative return distributions conditional on momentum. Stocks are double-sorted independently on the one-year price momentum and the one-month ahead return, and a probability mass function is drawn within each momentum decile. The title of each graph denotes the momentum decile, and the horizontal axis denotes the return class. The dotted line represents the average return class. The sample period is from 1955.01 to 2017.01.

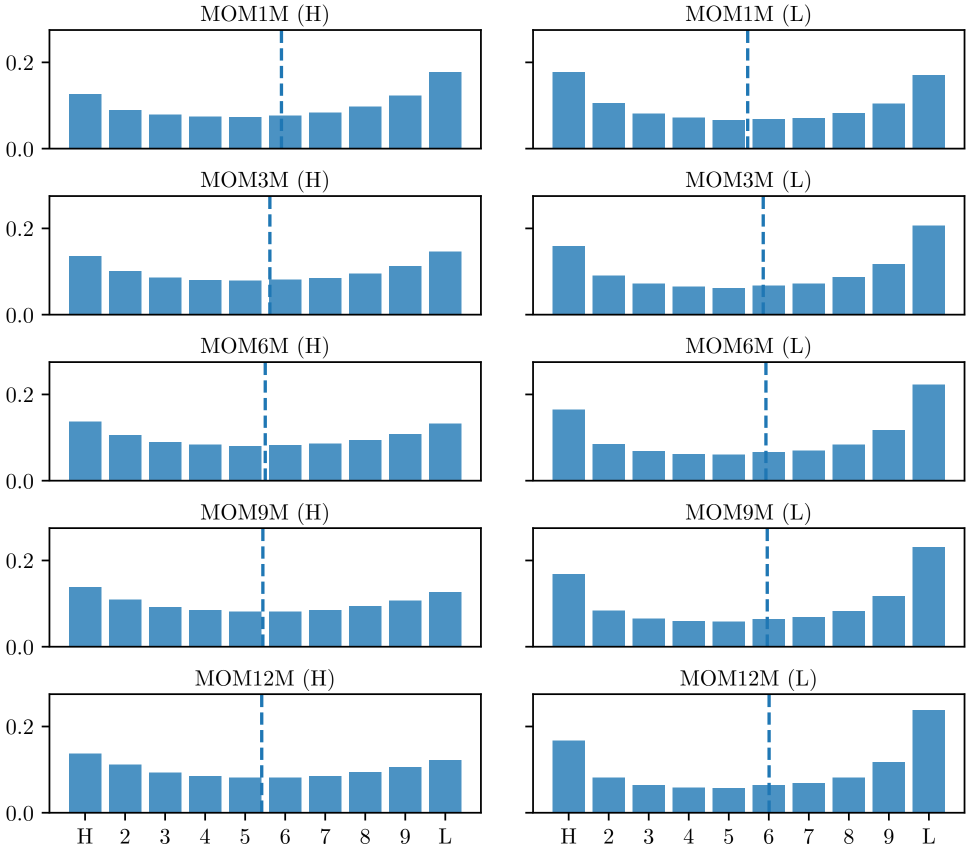


Figure 4: Cross-sectional relative return distribution: various price momentums

This figure presents the cross-sectional relative return distributions in the high- and low-momentum deciles of various momentum features. Stocks are double-sorted independently on the one-year price momentum and the one-month ahead return, and a probability mass function is drawn within each momentum decile. The title of each graph denotes the momentum feature, and the horizontal axis denotes the return class. The dotted line represents the average return class. The sample period is from 1955.01 to 2017.01.

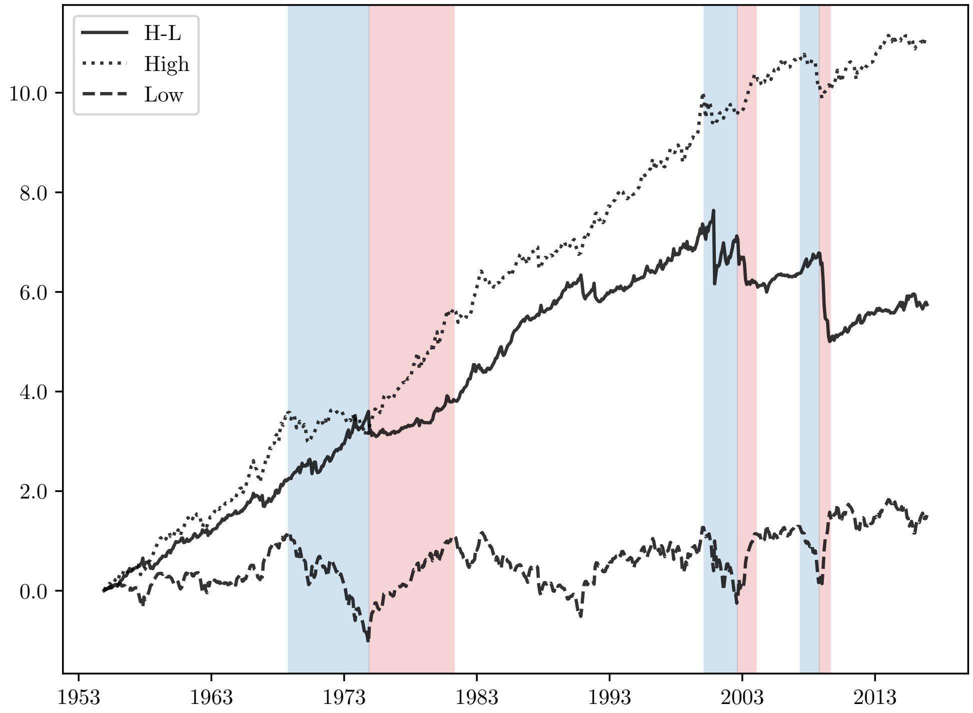


Figure 5: Periods of recession and recovery

This figure presents the cumulative return of the momentum strategy over the period from 1955.01 to 2017.01. Three periods of recession and recovery are identified based on the low-momentum portfolio's cumulative return and are highlighted respectively in blue and red. The identified recession periods are 1968.12-1974.12, 2000.04-2002.09, and 2007.07-2008.11, and the recovery periods are 1975.01-1981.05, 2002.10-2004.02, and 2008.12-2009.09.

cumulative return (Figure 5): 1968.12 - 1974.12 and 1975.01 - 1981.05; 2000.04 - 2002.09 and 2002.10 - 2004.03; and 2007.07 - 2008.11 and 2008.12 - 2009.09. It is not apparent whether there is a momentum crash in the first recovery period, but the period is still included as the cumulative return of the low-momentum portfolio exhibits a distinct down-up pattern.

Figure 6 presents the cross-sectional relative return distributions of the high- and low-momentum stocks in each sub-period. Bimodality is persistent in most periods and particularly distinct among low-momentum stocks. During a recovery period, the distribution of low-momentum stocks has more masses in high-return groups and the highest mode shifts from L to H. This results in a sharp increase in the loser portfolio's value and causes a momentum crash. The substantial change in the low-momentum stocks' distribution indicates that these stocks are more sensitive to the overall market movement, which is consistent with the high beta of the loser portfolio documented in Daniel and Moskowitz (2016).

2.2 Bimodality Everywhere

The bimodality is not unique to momentum stocks. Figure 7 demonstrates the cross-sectional relative return distributions derived from various firm characteristics. The characteristics are chosen

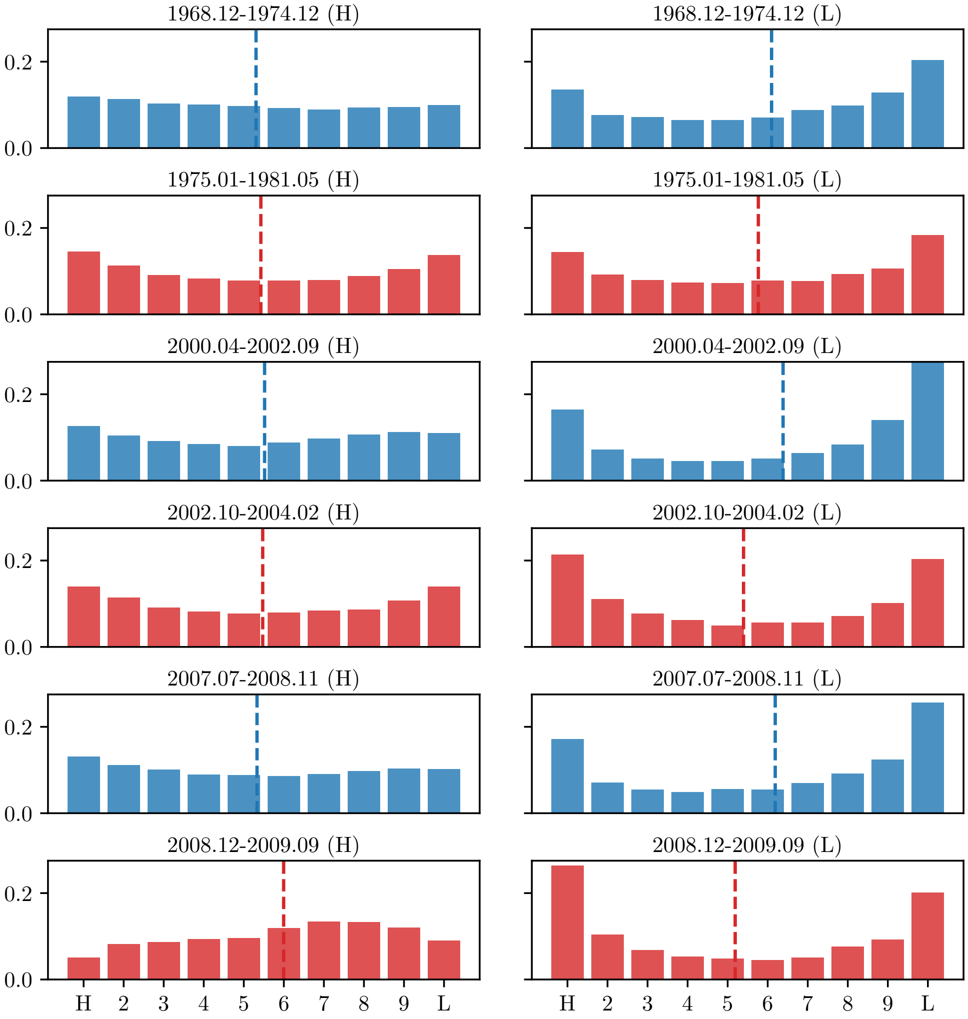


Figure 6: Cross-sectional relative return distribution: recession and recovery periods

This figure presents the cross-sectional relative return distributions in the high- and low-momentum deciles of the one-year price momentum over the recession and recovery periods defined in Figure 5. Stocks are double-sorted independently on the one-year price momentum and the one-month ahead return, and a probability mass function is drawn within each momentum decile. The title of each graph denotes the sample period, and the horizontal axis denotes the return class. The dotted line represents the average return class.

from those used by Green et al. (2017), and their definitions can be found in their appendix. The figure shows that bimodality is prevalent across the characteristics: *e.g.*, high beta stocks, low book-to-market stocks, illiquid stocks, small stocks, and volatile stocks form a bimodal cross-sectional distribution. This finding suggests that the bimodality is associated with return predictability. A study on the relationship between the cross-sectional distribution and risk/return will be interesting future research.

Despite the bimodality, the higher average return of high-momentum stocks and the lower average return of low-momentum stocks suggest that momentum provides valuable information about future returns. If this information could be fully exploited, future returns would become more predictable. The next section explores methods to enhance the predictive power of momentum.

3 Building the Deep Momentum Strategy

3.1 A Little Background on the Model

Given the nonlinear relationship between the momentum features and the future return, machine learning appears to be the right path to improve the predictive power of momentum. Machine learning also allows us to classify stocks exploiting information from multiple features simultaneously. There are two approaches to formulate the return prediction problem via machine learning. One is to frame it as a multiclass classification problem and the other is to frame it as a regression problem.

In the first approach, the objective is to predict a stock's future return class; this is similar to traditional cross-sectional prediction models, in which stocks are sorted on a predictor and classified into quantiles.[1] This approach is suitable for a relative return strategy, such as the long-short strategy, but can also be extended to predict absolute returns, as shown in Section 3.3. As stocks' future returns—and thus their true return classes—are known in the training sample, this problem is categorized as supervised learning, for which machine learning offers a rich set of multiclass classification algorithms, such as random forest and artificial neural network. An example employing this approach is the work of Moritz and Zimmermann (2016), who employ a tree-based method.

The second approach aims to forecast the absolute return of a stock by minimizing the difference between realized returns and predicted returns. The majority of machine learning asset pricing studies, *e.g.*, Messmer (2017); Gu et al. (2020); Bianchi et al. (2020), adopt this approach. Although these studies find machine learning's predictive power for absolute returns, a cross-sectional prediction has been far more successful in the asset pricing literature, and it is likely to be the case even when machine learning methods are employed. Moreover, given the bimodality of stocks' relative return distribution, a point estimate is unlikely to be a sufficient statistic of a stock's future

[1] Predicting a stock's price movement (up or down) is another common classification problem but is different from this approach as it involves absolute return prediction.

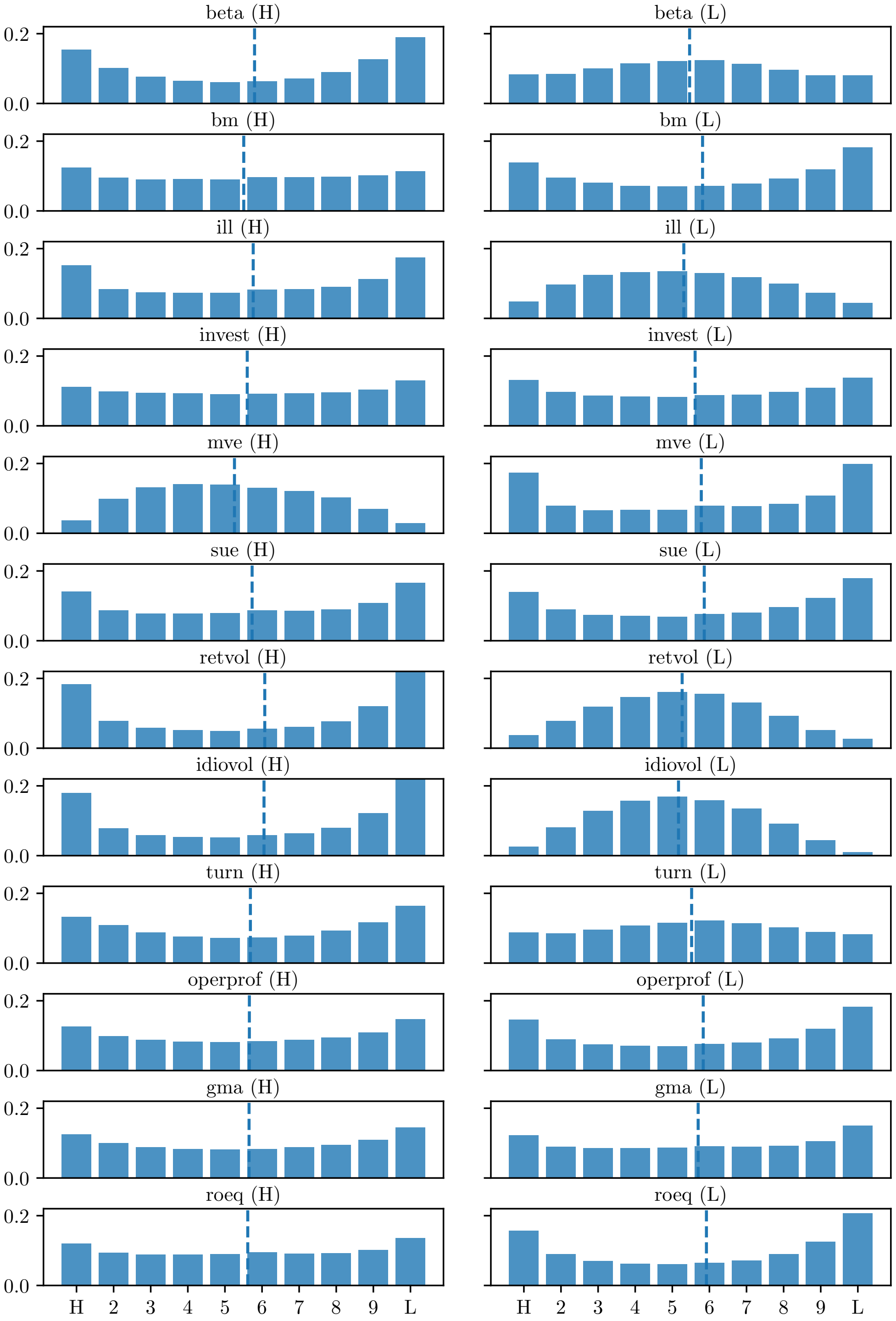


Figure 7: Cross-sectional relative return distribution: various factors

This figure presents the cross-sectional relative return distributions in the high and low deciles of various firm characteristics. Stocks are double-sorted independently on a characteristic and the one-month ahead return, and a probability mass function is drawn within each characteristic decile. The definitions of the characteristics can be found in the appendix of Green et al. (2017). The horizontal axis denotes the return class. The dotted line represents the average return class. The sample period is from 1980.01 to 2017.01.

performance.[2] Therefore, this study adopts the first approach.

A straightforward method to implement the first approach is to employ an off-the-shelf classifier readily available from many machine learning software packages. However, applying an off-the-shelf model to asset pricing poses two potential problems: 1) the predicted class may not correspond to the mean return of the stock, and 2) the number of stocks in each class can vary considerably across classes.

The first problem arises when the predicted return distribution is not unimodal, in which case the predicted class, *i.e.*, the class with the highest probability can be different from the class that corresponds to the mean return. Indeed, it is found that stocks classified into an extreme (either highest- or lowest-return) decile often have the second-highest probability in the opposite class.

Even when a model is trained using a balanced set of data—this is the case of stock classification where each decile contains 10% of the sample—the predicted classes can be imbalanced. Neural network models using momentum features tend to allocate more stocks to the lowest-return class, making the implementation of a long-short portfolio strategy arduous as it requires selling short many stocks.

With the problems above, it is crucial to consider the entire probability distribution (the probabilities for all classes) and to reclassify stocks so that the predicted class better reflects financial performance, and the stocks are distributed uniformly across classes. This paper offers five reclassification methods to achieve this goal.

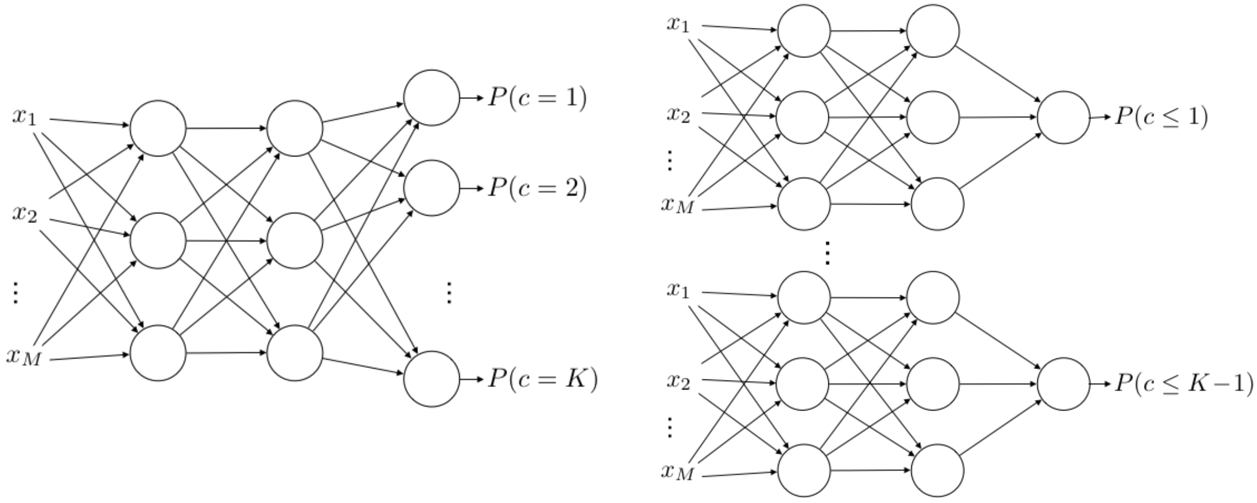
The remainder of this section consists of two parts. The first part describes the neural network models used to estimate the cross-sectional distribution, and the second part develops the five reclassification methods using the results from the first step.

3.2 Stock Classification via Deep Neural Network

In the first step, a deep neural network (DNN) is employed for stock classification.[3] A DNN is chosen because it provides a stock's probability for each class as well as its predicted class. These probabilities essentially form a discretized version of the stock's return distribution, which acts as a vital element in the second step. Moreover, DNNs have shown superior performance for asset pricing compared to other machine learning methods: see *e.g.*, Gu et al. (2020); Bianchi et al. (2020).

[2] In principle, a regression model does not suffer from the bimodality problem. If a stock has high probabilities for both extreme returns, the regressor will predict the return to be somewhere in the middle. However, the high risk cannot be inferred from the point estimate of the return.

[3] The reader is assumed to be familiar with the basics of neural networks, and the description is limited to the necessary elements to understand the proposed methodology. For more details on neural networks and machine learning in general, the reader is referred to LeCun et al. (2015); Goodfellow et al. (2016).



(a) Nominal classifier

(b) Ordinal classifier

Figure 8: Nominal classifier vs ordinal classifier

This figure describes two types of classifiers: nominal (a) and ordinal (b). Nominal classification is performed in a single step, whereas ordinal classification requires $(K-1)$ binary classifications, where K is the number of classes.

3.2.1 Nominal Classifier

Let $x = \{x_1, \ldots, x_M\}$ denote the input features (explanatory variables), and c and y the target variable (output class) and its one-hot encoding, respectively: if possible outcomes are $1, \ldots, K$, and the true class is k, then $c = k$ and y is a K-dimensional vector with the k-th element equal to 1 and 0 otherwise. A DNN estimates the probability for each class, $P(c = k),\ k = 1, \ldots, K$, and chooses the class with the highest probability as the predicted class. A typical multi-layered neural network for multiclass classification is illustrated in Figure 8(a).

Given a sample of N observations, $\{x^i, y^i\}_{i=1}^N$, a DNN can be trained to minimize the cost function:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, \hat{y}^i\left(x^i; \theta\right)\right), \tag{1}$$

where θ is a set of model parameters, and $\hat{y}(x;\theta)$ is the output of the DNN, whose k-th element is given by the softmax function:

$$\hat{y}_k(x;\theta) = \frac{e^{z_k(x;\theta)}}{\displaystyle\sum_{k=1}^{K} e^{z_k(x;\theta)}}, \quad k = 1, \ldots, K, \tag{2}$$

for some function $z_k(x;\theta)$. The exact form of $z_k(x;\theta)$ is determined by the network specification. Note that $\hat{y}_k(x;\theta)$ can be interpreted as the probability of a sample with features x belonging to class k, $P(c = k|x)$. The loss function is defined as the cross-entropy

$$L\left(y, \hat{y}(x; \theta)\right) = -\sum_{k=1}^{K} y_k \log \hat{y}_k(x; \theta). \tag{3}$$

The multiclass classifier defined above is a nominal classifier meaning that it ignores the order between classes: it only cares whether a prediction is correct or not and is indifferent to the distance between the predicted class and the true class. In contrast, there exists ordinality between the return classes: if the true class of a stock is 1 (highest return class), predicting its class to be 2 (second-highest) is better than predicting it to be 10 (lowest). Therefore, it is important to take the ordinality into account when classifying stocks.

3.2.2 Ordinal Classifier

Ordinality can be taken into account by reformulating the problem as $(K-1)$ binary classification problems as illustrated in Figure 8(b). The k-th binary classifier classifies stocks into two groups: a group consisting of the classes $1, \ldots, k$ and the other group consisting of the rest. Hence, the logistic function of the k-th binary classifier has the interpretation of the cumulative probability, $P(c \leq k)$, and can be used to derive the probability mass function as follows:

$$\begin{aligned} P(c = 1) &= P(c \leq 1) \\ P(c = 2) &= P(c \leq 2) - P(c \leq 1) \\ &\vdots \\ P(c = K) &= 1 - P(c \leq K - 1). \end{aligned} \tag{4}$$

The k-th binary classifier has the following loss function:

$$L^k\left(\nu, \hat{\nu}(x; \theta)\right) = -\nu \log \hat{\nu}(x; \theta) - (1 - \nu) \log(1 - \hat{\nu}(x; \theta)), \tag{5}$$

where $\nu = 1$ if $c \leq k$, and 0 otherwise, and $\hat{\nu}(x; \theta) = \frac{1}{1 + e^{z^k(x; \theta^k)}}$ for some function $z^k(\cdot)$ and its associated model parameters θ^k.

Data Imbalance Problem The binary classifiers in the ordinal classifier face a data imbalance problem, *e.g.*, when $k = 1$, one class ($c \leq 1$) contains only 10% of the sample, whereas the other class contains 90%, thereby rendering 90% accuracy when all stocks are predicted to be in the second class. This causes poor classification of the minority class.

In many problems such as cancer diagnosis or fraud detection, missing one in the minority (false negative or type II error) is more harmful than predicting a benign sample to be in the minority (false positive or type I error), and recall (true positive/positive) is a more relevant metric than precision (true positive/positive prediction). Therefore, it is crucial to preprocess an imbalanced data sample so that the data are balanced before training. There exist several preprocessing techniques, such as undersampling from the over-represented class, oversampling from the underrepresented class, and overweighting the underrepresented class.

In stock classification, in contrast, false negative (missing a high return stock) is not so harmful, whereas false positive (predicting a low return stock to be in a high return class) can cause a significant loss; *i.e.*, precision is a more relevant metric than recall. Therefore, the data imbalance in the ordinal classifier is indeed beneficial as it has the effect of emphasizing precision, and preprocessing becomes unnecessary.

3.2.3 Input Features

Two sets of input features are tested in the empirical studies, namely one set derived from the price momentums and the other set derived from past monthly returns. As the momentum features are constructed from past monthly returns, the latter carries the information of the former. The objective is to examine whether the DNNs can extract the information contents of momentum and beyond from the raw monthly returns. The input features are described below and summarized in Table 1(a).

Price Momentum Features Five price momentum features are chosen for the momentum-based feature set. The m-month price momentum, MOM_m, is defined as the cumulative return over the period from $t-m$ to $t-2$ for $m = 3, 6, 9, 12$, and the previous one-month return for $m = 1$ (Jegadeesh and Titman, 1993):

$$MOM_m = \prod_{j=t-m}^{t-2} (r_j + 1) - 1, \quad m = 3, 6, 9, 12, \tag{6}$$

$$MOM_1 = r_{t-1}, \quad m = 1, \tag{7}$$

where r_j denotes the return in month j.

Since a DNN is trained with many years' data spanning different market conditions, market trends need to be removed from the momentum features; otherwise, a past return will have the same effect on the output, although it can be considered high in a bear market and low in a bull market. To eliminate the overall market trend from the momentum features, they are normalized in each month by the cross-sectional mean and standard deviation:[4]

$$nMOM_m = \frac{MOM_m - M_{MOM_m}}{S_{MOM_m}}, \tag{8}$$

where M_{MOM_m} and S_{MOM_m} are respectively the cross-sectional mean and standard deviation of MOM_m.

[4] This is different from the normalization commonly employed in a neural network model, whereby inputs are normalized across the entire sample.

The final feature set consists of the normalized momentum features, $nMOM_m$, and the cross-sectional means, M_{MOM_m}. The latter is included to take the macroeconomic status into account.

Past Monthly Returns The second feature set consists of past twelve monthly returns denoted by RET_m:

$$RET_m = r_{t-m}, \quad m = 1, \ldots, 12. \tag{9}$$

Monthly returns up to one year are used to be consistent with the period of the price momentum features. For the same reason as discussed above, the monthly returns are normalized by their cross-sectional mean and standard deviation:

$$nRET_m = \frac{RET_m - M_{RET_m}}{S_{RET_m}}, \tag{10}$$

where M_{RET_m} and S_{RET_m} are respectively the cross-sectional mean and standard deviation of RET_m. The final feature set consists of the normalized past monthly returns, $nRET_m$, and the cross-sectional means, M_{RET_m}.

Size Dummies The price momentum or monthly return features can generate significant returns for equal-weighted portfolios, but they are not as effective for value-weighted portfolios. This implies that stocks of different sizes follow different price dynamics. In order to account for the size effect, size dummies are added to the feature set. In each month, stocks are divided into deciles on the market capitalization at the end of the previous month and assigned one of ten size dummies, $D_s, \ s = 1, \ldots, 10$.

Table 1(b) summarizes the combinations of input features and classifiers tested in the empirical analysis.

3.2.4 Estimation and Model Calibration

The architecture of a DNN is determined by the number of layers, the number of neurons in each layer, and the activation function of each neuron, and these hyperparameters need to be specified before estimating the model parameters. A neural network with many hidden layers and neurons is likely to overfit the data, whereas a shallow neural network may not have sufficient power to decompose and analyze data effectively. A mix of random and grid search algorithms are used to span the space of the hyperparameters and the final values are chosen via cross-validation. The final network architecture is reported in Table 1(c).

Regularization Besides the hyperparameters associated with the architecture, regularization parameters need to be specified to prevent overfitting. Combinations of L1-norm, L2-norm, dropout, and early stopping are tested, and early stopping is found to be most effective. The other regularization techniques also help, especially when the size dummies are included, but the gain appears

Table 1: Model specifications

This table presents the specifications of the test models. Panel (a) lists the input features of the neural network models, panel (b) lists the models tested in the empirical study, and panel (c) describes the network architecture.

(a) Input features

Price Momentum Features	
$nMOM_m$	m-month price momentum normalized by the cross-sectional mean and variance. $m \in \{1, 3, 6, 9, 12\}$.
M_{MOM_m}	Cross-sectional mean of MOM_m.
Past Monthly Returns	
$nRET_m$	$(t - m)$ month return normalized by the cross-sectional mean and variance. $m \in \{1, \ldots, 12\}$.
M_{RET_m}	Cross-sectional mean of RET_m.
Size Dummies	
D_s	Size dummy. $s = 1, \ldots, 10$.

(b) Test models

Model	Input features	Classifier
MOM-NOM	$nMOM_m + M_{MOM_m}$	Nominal
MOM-ORD	$nMOM_m + M_{MOM_m}$	Ordinal
RET-NOM	$nRET_m + M_{RET_m}$	Nominal
RET-ORD	$nRET_m + M_{RET_m}$	Ordinal
MOM-SZ-NOM	$nMOM_m + M_{MOM_m} + D_s$	Nominal
MOM-SZ-ORD	$nMOM_m + M_{MOM_m} + D_s$	Ordinal
RET-SZ-NOM	$nRET_m + M_{RET_m} + D_s$	Nominal
RET-SZ-ORD	$nRET_m + M_{RET_m} + D_s$	Ordinal

(c) Network architecture

Hyperparameter	Selected Value
Number of hidden layers	5
Number of neurons	64 for each hidden layer
Activation function	ReLU

to be minimal. In order to evaluate the models from a conservative perspective, this paper employs only early stopping for all model specifications.

3.2.5 Model Evaluation

The DNN models are evaluated using conventional classification performance metrics. For the overall classification performance, the average loss and accuracy are employed:[5]

$$\text{Average Loss} = \frac{1}{N}\sum_{i=1}^{N} L\left(y^i, \hat{y}^i\right), \tag{11}$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{N}. \tag{12}$$

To assess the performance in each class, the following metrics are employed:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{13}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{14}$$

$$\text{F1-score} = \frac{2}{1/\text{Precision} + 1/\text{Recall}}, \tag{15}$$

where TP, FP, and FN respectively denote true positive, false positive, and false negative.

The following loss measures designed to account for ordinality are also employed:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(c^i - k)^2 \hat{y}_k^i, \tag{16}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}|c^i - k| \hat{y}_k^i. \tag{17}$$

Both MSE and MAE have their minimum value of 0 when the predicted probability for the true class is 1 and increase as the probabilities for distant classes increase. These measures can also be defined at the class level: they are respectively denoted by MSEP and MAEP when stocks are grouped on predicted classes, and by MSER and MAER when grouped on true classes. MSEP and MAEP measure the prediction error within a predicted class and is similar to precision, whereas MSER and MAER measure the prediction error within a true class and is similar to recall. Like precision, MSEP and MAEP are deemed more relevant to stock classification.

3.3 Reclassification

The standard method to predict the class of a stock is to choose the class with the maximum probability, *i.e.*, $\hat{c} = \operatorname{argmax}_k \hat{y}_k$. This method, however, poses two problems when applied to

[5] The loss is defined as in Equation (3) for both nominal and ordinal classifiers.

portfolio strategy development.

The first problem is that the class of the highest mode can be different from the class corresponding to the mean return. In such a case, stocks classified into the top (bottom) class will not be the ones with the highest (lowest) expected return. The second problem is that the number of stocks can vary considerably across classes. This can make a long-short portfolio strategy difficult to implement. The classifiers tend to allocate more stocks to the bottom class, thereby requiring large short sales.[6]

This section develops five reclassification methods to address these problems. The first two methods are mainly concerned with redistributing stocks to ensure they are uniformly distributed across classes. The next two methods predict the expected (relative) returns of stocks and use them to reclassify the stocks. The last method predicts the Sharpe ratio and uses it as a reclassification criterion. This method is particularly relevant to investors who face short-sale constraints. In the description of each method below, the number of classes, K, is assumed to be an even number.

3.3.1 Reclassification on Ranking (Rank)

The first method is to reclassify stocks based on their ranking (probability) within each class. The procedure is as follows.

1. Let $U_k,\ k = 1, \dots, K$, denote the number of remaining positions in class k. In the beginning, $U_k = \frac{1}{K}N$ for all classes.

2. For an unclassified stock i, find its ranking among unclassified stocks within each class based on the probability for that class. Denote the rank of stock i in class k $R_k(i)$: *e.g.*, if stock i has the highest $P(c = k)$ in class k, $R_k(i) = 1$.

3. The predicted class of stock i is the class with the highest $P(c^i = k)$ among the classes that satisfy $R_k(i) \le U_k$, *i.e.*,
$$\hat{c}^i = \underset{k}{\operatorname{argmax}}\, \hat{y}^i_k \quad \text{s.t. } R_k(i) \le U_k.$$

4. Repeat steps 2 and 3 until all stocks are assigned a class.

3.3.2 Reclassification on Probability (Prob)

This method is similar to the standard classification method, except it redistributes stocks between adjacent classes to ensure they are uniformly distributed across classes. Let F_k denote the number of stocks allocated to class k by a classifier in Section 3.2.

1. If $F_1 \ge \frac{1}{K}N$, keep the first $\frac{1}{K}N$ stocks (based on $P(c = 1)$) for class 1. Use the rest to fill class 2.

[6] This phenomenon is associated with the empirical fact that the momentum profit is mainly driven by the short positions.

2. If $F_1 < \frac{1}{K}N$, borrow $\left(\frac{1}{K}N - F_1\right)$ stocks from class 2 to fill class 1. Choose stocks with the highest $P(c = 1)$. If the number of stocks in class 2 is not sufficient, *i.e.*, $F_2 < \frac{1}{K}N - F_1$, borrow the difference from class 3.

3. Apply the same procedure moving down the classes until the $K/2$-th class is filled.

4. Repeat the above steps backwards starting from the last class ($k = K$) until the remaining classes ($k = K/2 + 1, \dots, K$) are filled.

3.3.3 Reclassification on Probability Difference (PrDf)

If the expected returns of the classes were known, the expected return of a stock could be obtained using the law of total expectation:

$$\begin{aligned} \mu = E[r] = E\left[E[r|c]\right] &= \sum_{k=1}^{K} P(c = k)E[r|c = k] \\ &= \sum_{k=1}^{K} P(c = k)\mu_k, \end{aligned} \tag{18}$$

where μ_k is the expected return of class k. Stocks could then be reclassified based on the expected return.

Without the knowledge of the true mean returns of the classes, they might be assumed to decrease linearly from top to bottom:

$$\begin{aligned} \mu_1 &= \bar{\mu} + \frac{K}{2}\eta, & \mu_K &= \bar{\mu} - \frac{K}{2}\eta, \\ \mu_2 &= \bar{\mu} + \left(\frac{K}{2} - 1\right)\eta, & \mu_{K-1} &= \bar{\mu} - \left(\frac{K}{2} - 1\right)\eta, \\ &\vdots \\ \mu_{K/2} &= \bar{\mu} + \eta, & \mu_{K/2+1} &= \bar{\mu} - \eta, \end{aligned} \tag{19}$$

where $\bar{\mu}$ is the average of μ_k and η is a constant. Then the mean return of a stock becomes proportional to

$$\sum_{k=1}^{K/2} \left(P(c = k) - P(c = K + 1 - k)\right)\left(\frac{K}{2} + 1 - k\right). \tag{20}$$

This idea is generalized to develop the following sorting criteria:

$$PD(h) = \sum_{k=1}^{h} \left(\hat{y}_k - \hat{y}_{K+1-k}\right)\left(\frac{K}{2} + 1 - k\right), \quad h = 1, \dots, \frac{K}{2}. \tag{21}$$

For the empirical analysis, $PD(1)$ (PrDf1) and $PD(5)$ (PrDf5) are employed.

3.3.4 Reclassification on Mean Return (Return)

In this method, the mean return of each class is estimated by the sample analogue over the past twenty years, and the mean stock return is estimated from Equation (18). Denoting the sample mean of class k $\hat{\mu}_k$, the estimate of stock i's mean return is given by

$$\hat{\mu}^i = \sum_{k=1}^{K} \hat{y}_k^i \hat{\mu}_k. \tag{22}$$

3.3.5 Reclassification on Sharpe Ratio (Sharpe)

If short sales are not allowed, it is better to sort stocks on the Sharpe ratio than the expected return. Using the law of total variance, the variance of a stock return can be written as follows:

$$\begin{aligned}\sigma^2 &= V[r] = E\left[V[r|c]\right] + V\left[E[r|c]\right] \\ &= \sum_{k=1}^{K} P(c=k)V[r|c=k] + \sum_{k=1}^{K} P(c=k)E[r|c=k]^2 - \left(E\left[E[r|c]\right]\right)^2 \\ &= \sum_{k=1}^{K} P(k)\left(\sigma_k^2 + \mu_k^2\right) - \mu^2,\end{aligned} \tag{23}$$

where σ_k^2 is the variance of class k's return. Substituting μ_k and σ_k^2 with their sample analogues $\hat{\mu}_k$ and $\hat{\sigma}_k^2$, the variance of stock i can be estimated as follows:

$$\hat{\sigma}^{i2} = \sum_{k=1}^{K} \hat{y}_k^i \left(\hat{\sigma}_k^2 + \hat{\mu}_k^2\right) - \hat{\mu}^{i2}. \tag{24}$$

The Sharpe ratio of the stock is then estimated by $SR^i = \hat{\mu}^i/\hat{\sigma}^i$.

4 Empirical Analysis

4.1 Data

The sample for the empirical studies consists of the US equity market data available from the Center for Research in Security Prices (CRSP). All stocks with common shares (share code 10 or 11) listed on NYSE, Amex, and Nasdaq (exchange code 1, 2, and 3) are included. The sample period is from 1955.01 to 2017.01. The Treasury Bill rate is used as the risk-free rate.

In month t (at the end of $t-1$) during the sample period, stocks are collected with the following conditions to ensure the input features are available.[7] To be included in the sample for month t, a stock must have a price at the end of month $t-13$ and a return for $t-2$. Besides, any missing

[7] These conditions are adopted from the K. French's website: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_10_port_form_pr_12_2.html`.

returns from $t-12$ to $t-3$ must be -99.0. Each included stock should also have market equity at the end of month $t-1$. If a stock is delisted during the holding period, its return is calculated following the method of Beaver et al. (2007): if a delisted stock has a delist return with dividend in the CRSP delist file, it is used; otherwise, the delist return is assumed to be -30%.

After filtering, a total of 22,919 firms remain in the sample, and the monthly average, minimum, and maximum number of firms are 3,837 (1955.10), 994 (1955.11), and 6,777 (1997.11), respectively.

The first twenty years (1955.01 to 1974.12) of the sample are used to train the models, and the test period starts at 1975.01. The models are retrained every year stacking the sample while holding the last ten years of the data for validation: *e.g.*, in 1975.01, the training set consists of the stocks in the period from 1955.01 to 1964.12, and the validation set is from 1965.01 to 1974.12; in 1976.01, these sets respectively consist of the stocks from 1955.01 to 1965.12 and 1966.01 to 1975.12.

In every month during the test period (1975.01 to 2017.01), the stocks are classified into deciles using the latest model parameter estimates, and decile portfolios are constructed via either the equal-weight or value-weight scheme. These portfolios are held for one month.

In order to account for the randomness of machine learning training, all experiments are repeated fifty times, and the results presented in this paper are the averages of the fifty trials unless otherwise noted.

4.2 Classification Performance

4.2.1 Overall Classification Performance

This section evaluates the classification performance of the classifiers in Section 3.2. Table 2 presents the overall classification performance of the models in the training, validation, and test sets. For the performance evaluation of the Jegadeesh and Titman (1993)'s momentum strategy (JT), the predicted class is assumed to have the probability 1 and other classes 0: *e.g.*, the highest-momentum stocks have $\hat{y}_1 = 1$ and $\hat{y}_k = 0$ for $k \neq 1$.

The accuracy in the training set is around 15%, and it remains above 13.5% in the test set for all DNN models. JT has a lower accuracy of 13.07% in the test set, but it is still considerably higher than 10%, the accuracy of random prediction.

In terms of loss and accuracy, a nominal classifier always outperforms an ordinal classifier. This is expected since a nominal classifier minimizes the loss of the entire network, whereas an ordinal classifier minimizes the loss of each binary classifier individually.

In contrast, ordinal classifiers usually have smaller MSEs and MAEs. These measures account for ordinality and are more relevant to financial performance. Nevertheless, the error reduction seems rather trivial for the increased computational cost.[8] Surprisingly, JT has the lowest overall

[8] Training an ordinal classifier requires training nine binary classifiers. This takes more than five times the computational cost required to train a nominal classifier.

MSE and MAE. However, a closer investigation shows that JT has a significantly larger MSE and MAE in the extreme classes (H and L) compared to the neural network models (Table 3).

The performance of the classifiers in the validation or test set is comparable to the performance in the training set, suggesting that the classifiers do not suffer from overfitting, even though they only employ early stopping. The robustness of the models can be attributed to the small number of features compared to the size of the sample.

The return-based classifiers (RET-NOM and RET-ORD) perform comparably to the momentum-based classifiers (MOM-NOM and MOM-ORD). Although the monthly returns can reproduce the momentum features and potentially carry more information, they do not appear to enhance the classification performance.

4.2.2 Class-Level Classification Performance

Table 3 reports the class-level classification performance of the momentum-based classifiers. The classification performance of JT is also presented for comparison. The return-based classifiers perform comparably and their results are omitted to save space.

Remarkably, all models including JT have a significantly higher precision in the extreme classes. In particular, the precision of the lowest-return class is above 20% in all models, which means more than 20% of the stocks predicted to be in this class will end up in the class. This finding is consistent with the empirical fact that the profit of a momentum strategy results mainly from the short positions. Extreme classes also have high recall values indicating that these models can identify high and low return stocks.[9]

As mentioned earlier, JT performs much worse than the DNN models in the extreme classes in terms of most metrics. The difference is particularly evident in class H, suggesting that the DNN models are better than JT at identifying winners. Still, all models can identify losers better than winners.

4.2.3 Reclassification Performance

Table 4 compares the effects of the reclassification methods applied to momentum-based classifiers. The return-based classifiers yield similar results and their results are omitted from the paper. All metrics are computed during the test period.

When stocks are reclassified on Rank or Prob, precision increases in both extreme classes, whereas recall decreases. This is desirable as the correct identification of high- and low-return stocks is more important than missing stocks in these classes. Reclassifying stocks on Rank or Prob, however, does not reduce the classification error measured by MSEP or MAEP.

In contrast, when stocks are reclassified on a measure of expected financial performance (PrDf, Return, or Sharpe), both MSEP and MAEP decrease significantly, especially in the highest-return

[9] The small recall value of MOM-ORD in class H (12.71%) is due to the small number of stocks allocated to this class (6.93%).

class. Unexpectedly, precision also decreases, suggesting that it does not reflect financial performance adequately. The results in Section 4.3 support this claim: the portfolios derived from PrDf, Return, or Sharpe outperform those derived from Rank or Prob. This finding highlights the importance of accounting for financial performance during classification and the risk of solely relying on conventional evaluation metrics.

Figures 9 and 10 compare the probability distributions after reclassification. The orange bars represent the average predicted probabilities within each class, and the blue bars represent the actual frequencies obtained from realized returns. The predicted probability distributions before reclassification (Org) show that the stocks predicted to be in H (L) have a high probability for L (H), implying that the predicted class is unlikely to reflect the expected return.

Nevertheless, the predicted probabilities are notably similar to the actual frequencies. The ordinal classifier performs particularly well, as evidenced by the narrow gap between the mean values. The similarity between the predicted and actual distributions suggests that exploiting the entire distribution can offer a better classification that mirrors stocks' future performance.

The reclassification results reveal that this is indeed the case. When stocks are reclassified on a measure of expected financial performance, the bimodality in class H almost disappears. For instance, when stocks are reclassified on Return, the stocks in H have a significantly smaller predicted probability for L and do end up in L less frequently, although they are also less likely to yield high returns. Compared to the H stocks of Org, these stocks have a similar average return but much lower risk, resulting in a considerably higher Sharpe ratio (see Section 4.3). The stocks predicted to be in L by Return reclassification still exhibit a bimodal distribution, but they are less likely to belong to H and more likely to belong to L, and have a lower average return compared to the L stocks of Org.

Figures 9 and 10 only report the probability distributions in class H and L, but the observations from the in-between classes are similar to those discussed here and confirm the effectiveness of reclassification, especially via a measure of financial performance.

4.3 Financial Performance

This section first demonstrates the performance of equal-weighted portfolios and then shows how the size dummies can improve the performance of value-weighted portfolios. Table 5 reports the mean excess returns (MR) and Sharpe ratios (SR) of the long (H), short (L), and long-short (H-L) portfolios constructed by the DM models using only momentum or return features, and Figure 11 presents the cumulative returns of these portfolios for the MOM-NOM classifier. All results are from the test period 1975.01 to 2017.01 and are averages of fifty random trials unless otherwise noted.

The level of the mean returns and Sharpe ratios are phenomenal, especially when stocks are reclassified on a financial performance measure. When stocks are reclassified on Return, the annualized mean return of the long-short portfolio (H-L) is above 40% in all models, and the annualized

Sharpe ratio is well above 2.0 and as high as 2.49 under the return-based ordinal classifier (RET-ORD). The corresponding minimum and maximum values out of the fifty trials are respectively 2.33 and 2.69. This result is a stark contrast to the poor performance of the standard momentum strategy, namely a mean return of 10% and a Sharpe ratio of 0.37. Recall that these models only use the price momentum features or monthly returns.

To put this result into perspective, Gu et al. (2020) achieve a Sharpe ratio of 2.35 using a neural network regression model with over a hundred features: 920 covariates derived from 94 characteristics, 74 industry dummies, and 8 macroeconomic variables. As shown later in the robustness check, the DM models outperform the models of Gu et al. (2020) in their sample period.

All reclassification criteria improve the standard neural network model (Org), but the Return criterion appears to offer the best and most robust performance. It yields the highest mean return in all cases and generates the highest Sharpe ratio in the majority of the cases. The portfolio obtained from the Return criterion also performs more consistently across the classifiers.

Compared to the standard neural network model, reclassifying stocks on a financial performance reduces the risk of H stocks and generates a higher Sharpe ratio. It also lowers the average return of L stocks, resulting in a higher return and Sharpe ratio of the long-short portfolio.

Although the DM models perform well even before reclassification (Org), their performance is less consistent across models and random trials. Moreover, since the long and short portfolios of Org do not consist of 10% of the stock universe, any comparison between Org and other reclassification criteria should be interpreted with caution.

Investors who are not allowed to short-sell stocks will be more attracted to high Sharpe ratio stocks than high return stocks and can be served better by the Sharpe reclassification criterion. When stocks are reclassified on Sharpe, the long and short portfolios do earn the highest and lowest Sharpe ratios (about 1.20 and -0.43, respectively) in most models. Remarkably, reclassifying on the Sharpe ratio does not compromise profitability: the mean excess return of the long portfolio is slightly lower than that from the Return criterion but is still above 25% in all models.

Figure 11 presents the cumulative returns of the portfolios obtained from the NOM-MOM classifier. The value of the long-short portfolio grows steadily without much fluctuation over the entire test period. This figure—and those that follow—clearly shows that reclassifying stocks on financial performance enhances portfolio performance considerably. Another remarkable finding from the figures is that momentum crashes disappear in the DM models even though these models do not deploy any special device to prevent a crash. The DM models appear to learn from the past and exploit the opportunity to generate even higher returns.

4.3.1 Nominal vs. Ordinal Classifiers

Section 3.2 argues that the order between return classes matters and introduces ordinal classifiers. The ordinal classifiers indeed exhibit superior classification performance and yield higher returns before reclassification. However, the difference between the nominal and ordinal classifiers becomes almost indistinguishable after reclassification. Although it is disappointing that the ordinal classifiers no longer outperform the nominal classifiers, it also means that reclassification can save a considerable amount of the computational time required for ordinal classification. Accounting for ordinality gives a rather limited gain, and redistributing stocks via reclassification appears to override it. The rest of the paper drops the ordinal classifiers and focuses on the nominal classifiers. Nevertheless, it may be worth revisiting ordinal classifiers for future research, given their superior classification performance.

4.4 Performance of Value-Weighted Portfolios

When stocks are value-weighted, the long-short portfolio performs considerably worse than when they are equal-weighted. Both the mean return and Sharpe ratio are halved, indicating that the decrease in the Sharpe ratio is mainly due to the decrease in the mean return. This result is consistent with the finding of Gu et al. (2020), whereby the value-weighted long-short portfolio achieves the maximum Sharpe ratio of 1.25, about half of the equal-weighted portfolio's maximum Sharpe ratio. The poor performance of the value-weighted portfolios can be ascribed in part to the way the neural network models are trained, in which each stock carries the same weight in the cost function.

The performance of the value-weighted portfolio can be improved significantly by adding a categorical size variable (size dummies). Table 6 reports the performance of the DM models including the size variable. The results are phenomenal. The Sharpe ratio of the value-weighted long-short portfolio is often above 1.60 and as high as 1.86, and the annualized mean return is generally above 30%. The long portfolio alone has an annualized mean excess return close to 25% and a Sharpe ratio as high as 1.11. Recall that the presented results are averages of fifty trials, not the best case.

The size variable also improves the performance of the equal-weighted portfolio. The mean return of the equal-weighted long-short portfolio is above 40% in most cases and often exceeds 50% per annum. The Sharpe ratio is usually above 2.6 and as high as 2.90. The long portfolio alone earns an annualized mean excess return of 35% and a Sharpe ratio above 1.35 in several cases. The Sharpe ratio of the long portfolio is particularly high under the Sharpe criterion, confirming its effectiveness.

Figure 12 shows that the value-weighted portfolios grow less consistently than the equal-weighted portfolios, but the long-short portfolio does not experience a momentum crash. Compared with Figure 11, it is obvious that the value-weighted long-short portfolio performs better with the size dummies.

The momentum strategy performs better when stocks are value-weighted: the annualized mean return is 19% and the Sharpe ratio is 0.61. Consistent with the findings of Barroso and Santa-Clara (2015), the volatility-adjusted momentum strategy outperforms the standard momentum strategy with a Sharpe ratio of 1.00 for the value-weighted portfolio and 0.80 for the equal-weighted portfolio. However, it underperforms the DM strategy by a large margin.

4.4.1 Size Effects

A good performance of a value-weighted portfolio does not imply that the strategy can be applied to large firms: if the top and bottom deciles consist of small firms, the value-weighted portfolio is merely a value-weighted portfolio of those small firms. In order to ensure that performance is not driven by the size effect, the firm sizes in each decile need to be examined.

The average firm size in each decile reveals that the momentum strategy tends to buy smaller stocks (average market capitalization of \$1,279M) than the market average (\$1,709M) and sell significantly smaller stocks (\$201M), which implies that many small stocks should be sold short to implement the strategy.

The DM strategy before reclassification buys and sells even smaller stocks with the average size of the long portfolio stocks equal to \$159M and that of the short portfolio stocks equal to \$153M. However, the size of the stocks contained in the DM portfolio changes dramatically after reclassification. When stocks are reclassified via Return, the average size of the stocks in the long portfolio becomes \$2,495M, larger than the market average, and the average size of the stocks in the short portfolio becomes \$1,589M, close to the market average. This result confirms that the performance of the DM strategy is not driven by small firms and also suggests that the DM strategy is more implementable than the momentum strategy.

4.5 Factor Regression

The returns of the decile portfolios are regressed on the Fama-French five factors plus the momentum and short-term reversal factors to examine whether these factors can explain the returns. The decile portfolios are obtained from the MOM-SZ-NOM classifier with Return reclassification. Table 7 reports the regression results.

The results show that the alphas of both value-weighted and equal-weighted long-short portfolios are economically and statistically significant—2.4% per month (t-statistic=6.63) and 3.5% per month (t-statistic=9.28), respectively, implying that the factors cannot explain the returns of these portfolios. When stocks are equal-weighted, the size factor (SMB) plays a more important role compared to when they are value-weighted.

The momentum (MOM) and the short-term reversal (STR) factors cannot explain much of the returns and the signs of their coefficients often contradict the momentum and short-term reversal effects. The coefficient on the momentum factor (MOM) is insignificant for the value-weighted H portfolio and negative (weakly significant) for the equal-weighted H portfolio, whereas it is significantly negative for the L portfolios. The coefficient on the short-term reversal factor (STR) is insignificant for both value-weighted H and L portfolios, whereas it is significantly positive for the equal-weighted H portfolio and significantly negative for the L portfolio. These results suggest that the momentum features in the DM models do not predict the future returns in a linear fashion and the DM models must exploit the nonlinear information contained in the past returns to enhance return predictability.

4.6 Turnover and the Effects of Transaction Costs

The high return and the low rebalancing frequency of the DM strategy imply that transaction costs are unlikely to erase its profits. To confirm this, Table 8 presents the performance of a DM strategy under three different transaction cost assumptions. The first two cases assume a constant transaction cost of 10 and 30 basis points, respectively. The last case follows DeMiguel et al. (2020) and Freyberger et al. (2020) and assumes that the transaction cost decreases with time and market capitalization. Under the last assumption, the transaction cost ranges from 35 basis points (when trading large-cap stocks in 2002.01 and onward) to 198 basis points (when trading small-cap stocks in 1980.01 or before). For the exact formula, the reader is referred to the above references.

When subject to 10 basis point transaction costs, the value-weighted DM strategy earns a mean return of 31% and a Sharpe ratio of 1.46, which are considerably higher than those of the standard momentum strategy and the volatility-adjusted momentum strategy before transaction costs. The strategy remains profitable even under a more conservative assumption of 30 basis point transaction costs, under which it earns a mean return of 23% and a Sharpe ratio of 1.09. Under the varying transaction costs, the value-weighted long-short strategy becomes barely profitable with a mean return of 4.8% and a Sharpe ratio of 0.22. The benefit from the short position is overshadowed by the transaction costs and the long-only portfolio outperforms the long-short portfolio with a mean excess return of 8.3% and a Sharpe ratio of 0.34. The equal-weighted long-short portfolio is more immune to the transaction costs due to its higher return, and it remains significantly profitable even under the varying transaction costs, with a mean return of 19.6% and a Sharpe ratio of 1.06. These results suggest that the long-short portfolio derived from the DM strategy will remain profitable even under a reasonably high transaction cost.

4.7 Sensitivity Analysis

Where does the superior performance originate from? The nonlinear nature of a neural network model makes it difficult to measure the effects of the input features on the prediction. The same value of a feature can predict a completely different outcome depending on the values of the other features. Nevertheless, a sensitivity analysis can help us better understand the role of each feature.

In Figure 13, the top panel demonstrates the predicted probability for each return class given a value of the one-month momentum (MOM1M), one-year momentum (MOM12M), or size (SIZE) feature, and the bottom panel demonstrates the predicted class given the values of a pair of features. All features are standardized except for the size dummies, and the values of the other features are assumed to be 0.

When the one-month momentum is either very small or large, both extreme deciles (H and L) have high probabilities. It is not clear from the figure, but the predicted return class is L when MOM1M is extremely small, becomes H as MOM1M increases, and remains as H when MOM1M is large. This result is in contrast with the short-term reversal effect, which would predict H when MOM1M is small and L when it is large.

When the one-year momentum has a small value, class L has a considerably higher probability than the other classes, which is consistent with the empirical fact that the short positions are easier to predict. As the value of MOM12M increases, the probabilities across return classes become less distinguishable. Contrary to the momentum effect, a large value of MOM12M predicts the return class to be L, although the probability for H is also high.

The return class probabilities hardly change as the size of the firm changes, but the probabilities for the extreme classes increase as the firm size reduces, which is consistent with the small-firm effect.

When a pair of features are investigated simultaneously, more interesting patterns emerge. When MOM1M is small, a large MOM12M (winner) predicts a high return as does the momentum effect, but when MOM1M is large, a large MOM12M predicts a low return. On the contrary, a small MOM12M predicts a high return when MOM1M is high, and a low return when MOM1M is low. This partly explains why the DM strategy shows superior performance during a momentum crash period. When the market recovers, losers perform better than winners and have higher short-term returns (MOM1M). This leads the losers to be classified into a high return group and the winners into a low-return group in the following months, resulting in the superior performance of the model over this period.

The middle graph shows that the short-term reversal effect is evident among small firms, but MOM1M predicts the opposite when firms are large. In contrast, the last graph suggests that there is only a weak interaction between SIZE and MOM12M.

The sensitivity analysis suggests that the DM model captures the nonlinear interaction between the input features, which cannot be identified by a linear model.

Figure 14 demonstrates another sensitivity analysis, where important features are identified by training the DM model while excluding some of the input features. Consistent with the previous analysis, it is clear that the past one-month return is the most important feature followed by the size dummies. Excluding the one-year momentum does not appear to affect the performance significantly, but it is perhaps because the mid-term momentum features such as MOM9M carries some of the information contained in MOM12M. The figure also reveals that the macroeconomic variables play a nontrivial role in the model. When all cross-sectional means, M_{MOM_m}, are excluded (Ex-Mmom), both the mean return and Sharpe ratio are reduced considerably.

4.8 Comparison with Alternative Models

Table 9 compares a DM model (MOM-SZ-NOM with Return reclassification) against a linear (logistic) model. The logistic classifier employs the same input features as those used in the DM model and its regularization parameters are determined via hyperparameter tuning. The logistic classifier also benefits from reclassification, and the results in the table are from Return reclassification. The table shows that the DM model significantly outperforms the logistic classifier with an almost twice-higher mean return and Sharpe ratio when stocks are value-weighted. This result further confirms the earlier finding that the nonlinear effects of the input features are an important contributor to the superior portfolio performance.

The table also contrasts the DM model against three regressors: a linear regressor and two nonlinear regressors. The linear regressor (Linear) is simply an ordinary least squares regressor with the dependent variable defined as the ranking of the stocks based on the one-month ahead return. The ranking is normalized so that the value lies in $[-1, 1]$. The first nonlinear regressor (Ordinal1) is a feed-forward neural network regressor that predicts the ranking. This type of regression is known as the ordinal regression.[10] The second nonlinear regressor (Ordinal2) is the same as Ordinal1 except that the target variable is the return class, *i.e.* grouped ranking. The regressors predict stocks' rankings or classes using the same input features and classify stocks into ten groups based on the magnitude of the predicted value. A long-short portfolio is then constructed by buying the stocks in the first group and selling those in the last group. The parameters of the neural network regressors are determined via hyperparameter tuning.

The linear regressor significantly underperforms the DM model and performs comparably to the linear classifier, reaffirming the importance of nonlinearity. The ordinal regressors perform impressively and their performances are comparable to each other. The value-weighted long-short portfolios derived from the ordinal regressors achieve relatively lower returns and Sharpe ratios compared to the portfolio from the DM model, but still earn annualized mean returns around 25% and Sharpe ratios above 1.2. The equal-weighted portfolios also perform very well with mean returns around 40% and Sharpe ratios around 2.5. Overall, the ordinal regressors perform comparably to the DM model without reclassification but underperform against the DM model augmented with reclassification. This result highlights the importance of reclassification in the DM models.

4.9 Performance Enhancement and Robustness Check

Appendix A examines several methods to enhance portfolio performance. Below is a summary of the findings and the details can be found in the appendix. When stocks are chosen from the top/bottom 5% or 1%, the annualized mean return and Sharpe ratio of the value-weighted long-short portfolio respectively increase to 45% and 1.75 or 82% and 1.88. When stocks are chosen from the intersection of ten or fifty repeated trainings, the annualized mean return and Sharpe ratio increase to 46% and 1.88 or 58% and 1.64. A stock-level portfolio optimization based on the DM model yields an annualized mean return of 59% and a Sharpe ratio of 3.29. These results suggest that there is room to improve portfolio performance even further.

Appendix B runs a host of robustness checks. It shows that the DM model is robust across random trials, and the performance remains little changed when the training set is rolled over instead of being expanded. The long-short portfolio strategy remains profitable even when the

[10] I am grateful to the anonymous reviewer who suggested the ordinal regressor.

stocks below the NYSE-size 20% quantile are excluded, with a mean return of 20% and a Sharpe ratio of 1.08. The DM model also exhibits consistently superior performance in an extended sample period. Increasing the number of classes helps improve the mean return but has a limited effect on the Sharpe ratio. More details can be found in the appendix.

5 Conclusion

This paper documents the bimodality of high- and low-momentum stocks' cross-sectional relative return distributions, which indicates that both past winners and losers have nontrivial probabilities for both high and low returns. The bimodality makes the momentum strategy fundamentally risky. In order to eliminate the bimodality and improve return predictability, this paper develops a novel cross-sectional prediction model via machine learning.

Tested on the US market over the the 1975.01 to 2017.01 period, a value-weighted long-short portfolio derived from the model earns an annualized mean return of 35% and a Sharpe ratio above 1.6, while an equal-weighted portfolio earns an annualized mean return around 50% and a Sharpe ratio above 2.8. Remarkably, these long-short strategies do not suffer from momentum crashes. The performance can be further improved to a tremendous annualized mean return of 82% and a Sharpe ratio of 1.88 when the value-weighted long-short portfolio is constructed from the top and bottom 1% of the stocks. Furthermore, a stock-level optimal portfolio derived from the model earns an annualized mean return of 59% and a Sharpe ratio of 3.29.

One major contribution of the paper is its discovery of bimodality. This bimodality is prevalent not only among momentum features but also across well-known firm characteristics, such as beta, book-to-market ratio, and size. The bimodality provides a new stock-level perspective on the riskiness of characteristic-based portfolio strategies.

Another contribution is the introduction of a new machine learning-based return prediction model and the demonstration of its significant predictive power and profitability. Unlike many extant studies that horse-race different off-the-shelf models, this paper identifies a problem in an existing model, *i.e.*, the bimodality of momentum stocks, and carefully designs a model to address it, employing machine learning where necessary. In particular, it introduces reclassification methods to reconcile the discrepancy between machine learning prediction and financial performance. Machine learning permits the development of a complex model involving a large number of input features and an arbitrary nonlinear relationship between the input and the output. This flexibility, however, is accompanied by the risk of overfitting. With a small number of input features, the proposed model performs robustly while benefiting from the power of machine learning.

This paper offers several topics for future research. Investigating the existence of the bimodality in other firm characteristics or asset classes and developing a suitable asset pricing model would be interesting. The results from different firm characteristics may also be combined to develop an ensemble model for return prediction. Applying the modeling framework of this paper to the estimation of the mean return and covariance matrix of a large dimensional system seems another promising area of research. The methodology may develop into a robust portfolio rule for large dimensional systems by accounting for the estimation errors in the class moments or the relative return distributions. Finally, different approaches can be considered to account for financial performance in the classification. Improving the ordinal classifier seems a good starting point.

References

Adebiyi, A.A., Adewumi, A.O., Ayo, C.K., 2014. Comparison of arima and artificial neural networks models for stock price prediction. Journal of Applied Mathematics 2014.

Atsalakis, G.S., Valavanis, K.P., 2009. Surveying stock market forecasting techniques–part II: Soft computing methods. Expert Systems with Applications 36, 5932–5941.

Barroso, P., Santa-Clara, P., 2015. Momentum has its moments. Journal of Financial Economics 116, 111–120.

Beaver, W., McNichols, M., Price, R., 2007. Delisting returns and their effect on accounting-based market anomalies. Journal of Accounting and Economics 43, 341–368.

Bianchi, D., Büchner, M., Tamoni, A., 2020. Bond risk premia with machine learning. Review of Financial Studies *forthcoming*.

Butt, H., Virk, N., 2019. Momentum crashes and variations to market liquidity. Working Paper.

Chong, E., Han, C., Park, F.C., 2017. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications 83, 187–205.

Dai, W., Wu, J.Y., Lu, C.J., 2012. Combining nonlinear independent component analysis and neural network for the prediction of asian stock market indexes. Expert systems with applications 39, 4444–4452.

Daniel, K., Jagannathan, R., Kim, S., 2019. A hidden markov model of momentum. Working Paper.

Daniel, K., Moskowitz, T.J., 2016. Momentum crashes. Journal of Financial Economics 122, 221–247.

DeMiguel, V., Martin-Utrera, A., Nogales, F.J., Uppal, R., 2020. A transaction-cost perspective on the multitude of firm characteristics. The Review of Financial Studies 33, 2180–2222.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of financial economics 116, 1–22.

Feng, G., Polson, N., Xu, J., 2018. Deep learning factor alpha. Available at SSRN 3243683 .

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. The Review of Financial Studies 33, 2326–2377.

Geczy, C.C., Samonov, M., 2016. Two centuries of price-return momentum. Financial Analysts Journal 72, 32–56.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. `http://www.deeplearningbook.org`.

Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average us monthly stock returns. The Review of Financial Studies 30, 4389–4436.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Gu, S., Kelly, B.T., Xiu, D., 2019. Autoencoder asset pricing models. Journal of Econometrics *forthcoming*.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of finance 48, 65–91.

Kazem, A., Sharifi, E., Hussain, F.K., Saberi, M., Hussain, O.K., 2013. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Applied soft computing 13, 947–958.

Khashei, M., Bijari, M., 2010. An artificial neural network (p, d, q) model for timeseries forecasting. Expert Systems with applications 37, 479–489.

Khashei, M., Bijari, M., 2011. A novel hybridization of artificial neural networks and arima models for time series forecasting. Applied Soft Computing 11, 2664–2675.

Kim, K.J., Ahn, H., 2012. Simultaneous optimization of artificial neural networks for financial forecasting. Applied Intelligence 36, 887–898.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Ledoit, O., Wolf, M., et al., 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics 40, 1024–1060.

Liao, Z., Wang, J., 2010. Forecasting model of global stock index by stochastic time effective neural network. Expert Systems with Applications 37, 834–841.

Messmer, M., 2017. Deep learning and the cross-section of expected returns .

Moritz, B., Zimmermann, T., 2016. Tree-based conditional portfolio sorts: The relation between past and future stock returns .

de Oliveira, F.A., Nobre, C.N., Zárate, L.E., 2013. Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index–case study of petr4, petrobras, brazil. Expert Systems with Applications 40, 7596–7606.

Patel, J., Shah, S., Thakkar, P., Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications 42, 259–268.

Rather, A.M., Sastry, V., Agarwal, A., 2017. Stock market prediction and portfolio selection models: a survey. OPSEARCH 54, 558–579.

Wang, J., Wang, J., 2015. Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks. Neurocomputing 156, 68–78.

Wang, J.Z., Wang, J.J., Zhang, Z.G., Guo, S.P., 2011. Forecasting stock indices with back propagation neural network. Expert Systems with Applications 38, 14346–14355.

A Further Performance Enhancement

This section explores several methods to enhance performance. Some are revealed to be effective, while others are less so.

A.1 Extreme Long-Short Portfolios

In the previous analyses, long-short portfolios are constructed going long on the top 10% stocks and short on the bottom 10% stocks. If the classifiers can identify future winners and losers at a granular level, constructing portfolios using the best of the best and the worst of the worst will generate even better performance. This question is answered by constructing portfolios using the top and bottom 5% or 1% stocks. The results are presented in Table 10 (row 5% and 1%) and Figure 15 and 16.

The level of the mean return achieved by the stocks at the extreme ends is enormous. When stocks are reclassified on a measure of financial performance, both the mean return and the Sharpe ratio keep increasing as stocks are chosen from more extreme ends: the annualized mean return and Sharpe ratio of the value-weighted long-short portfolio are, respectively, 45% and 1.75 at the 5% level and 82% and 1.88 at the 1% level when stocks are reclassified via Return. The equal-weighted long-short portfolio even earns about 100% at the 1% level.

The stocks in the top/bottom 1% are smaller than those in the top/bottom 10%, but they are still larger than those in the top and bottom deciles of the standard momentum strategy: the average market capitalizations of the stocks in the top and bottom 1% are respectively 1,347 and 315 million dollars, whereas those of the top and bottom deciles of the momentum strategy are 1,279 and 201 million dollars.

The Rank and Prob criteria underperform the financial performance-based criteria at the extreme ends. The Sharpe ratio decreases considerably while the increase in the mean return is moderate. The differences among the reclassification criteria are more evident in the cumulative return graphs. Rank and Prob lose their ability to identify losers at the extreme ends. The standard momentum strategy performs much worse at the extreme ends when stocks are equal-weighted. While it appears to perform better when stocks are value-weighted, it suffers enormous losses during the momentum crash in 2009.

Overall, a reclassification method based on a measure of financial performance maintains its ability to identify winners and losers even at the extreme ends, whereas other methods start to reveal their limitations.

A.2 Optimal Long-Short Strategy

Instead of the zero-investment long-short portfolio strategy, an optimal long-short portfolio strategy is examined. A mean-variance optimal portfolio of the H and L portfolios is given by

$$[w_H, w_L] = \left[1, -\frac{\mu_L \sigma_H^2 - \sigma_{HL}\mu_H}{\mu_H \sigma_L^2 - \sigma_{HL}\mu_L}\right], \tag{25}$$

where σ_{HL} is the covariance between the two portfolios, which is estimated from the monthly class returns over the previous twenty years. For simplicity, the weight on the long portfolio is fixed to 1. Over the test period, the mean, standard deviation, maximum, and minimum of $-w_L$ are respectively 2.267, 0.337, 3.169, and 1.509, indicating that the optimal strategy always sells more stocks in class L compared to the zero-investment strategy.

The results from the optimal strategy are presented in Table 10 (row Opt) and Figure 17. The performance of the optimal strategy is rather disappointing. While the mean return increases under Return and Sharpe, the Sharpe ratio is considerably compromised. The performance deteriorates more significantly under Rank, Prob, and JT. Due to the leverage of the optimal strategy, the cumulative return of the JT long-short portfolio falls into a deep negative territory during the momentum crash around 2009.

Although all models perform worse under the optimal portfolio rule, it is unambiguous that the DM models outperform JT, reclassification helps improve portfolio performance and its robustness, and the financial performance-based criteria are the most reliable reclassification criteria.

A.3 Intersection Portfolios

In machine learning, there is randomness in model parameter estimation, and the stocks chosen for each class vary every time the model is retrained. Therefore, the portfolio is expected to perform more robustly if it is constructed using common stocks across multiple trials. This method is a type of ensemble method.

Common stocks out of ten and fifty random trials are examined. When stocks are chosen from ten random trials, about 50-60% of the stocks are commonly included in H and 60-75% in L. When stocks are chosen from fifty random trials, the corresponding values are 20-35% and 40-55%, respectively. The criteria based on financial performance (PrDf1, PrDf5, Return, and Sharpe) tend to have smaller intersections compared to Rank and Prob. When no common stocks exist, only the other side portfolio is constructed.

The results are presented in Table 10 (row I10 and I50) and Figure 18 and 19. When stocks are chosen from ten random trials, the performance (mean return in particular) improves under most reclassification criteria. When stocks are chosen from fifty random trials, the mean return increases further under all reclassification criteria, but the Sharpe ratio starts to decrease under the financial performance-based criteria. This can be partly explained by the fact that the intersections are smaller under these criteria, and therefore the portfolios are less diversified. The results suggest that using the intersection of many trials improves the performance but only to the point where the adverse effect of reduced diversification starts to overwhelm the benefit.

A.4 Stock-Level Optimal Portfolio

Beginning with the cross-sectional prediction, the DM models derive the estimates of the mean and standard deviation of individual stock returns via the law of total expectation and variance. The same approach can be applied to estimate the covariance matrix of stock returns. From the law of total covariance, the covariance between the returns of stock i and j, σ^{ij}, is given by

$$\sigma^{ij} = Cov[r^i, r^j] = E\left[Cov[r^i, r^j | c^i, c^j]\right] + Cov\left[E[r^i|c^i, c^j], E[r^j|c^i, c^j]\right]. \tag{26}$$

Assuming that stocks' classes are determined independently of each other, *i.e.*, $P(c^i = k, c^j = l) = P(c^i = k)P(c^j = l),\ i \neq j$, it can be shown that σ^{ij} is given by

$$\sigma^{ij} = \sum_{k=1}^{K}\sum_{l=1}^{K} P(c^i = k)P(c^j = l)\sigma_{kl}, \quad i \neq j, \tag{27}$$

where σ_{kl} is the covariance between class k and l. The covariance, σ^{ij}, can then be estimated by

$$\hat{\sigma}^{ij} = \sum_{k=1}^{K}\sum_{l=1}^{K} \hat{y}_k^i \hat{y}_l^j \hat{\sigma}_{kl}, \quad i \neq j, \tag{28}$$

where $\hat{\sigma}_{kl}$ is a sample analogue of σ_{kl}.

The covariance estimates, together with the mean and variance estimates in Section 3.3, allow us to perform a stock-level portfolio optimization instead of using the arbitrary equal- or value-weight scheme. The mean-variance optimal portfolio of the following form is considered:

$$w^* = c\hat{\Sigma}^{-1}\hat{\mu}, \quad c = 1 \text{ or } \frac{0.02}{\sqrt{\hat{\mu}'\hat{\Sigma}^{-1}\hat{\mu}}}, \tag{29}$$

where $\hat{\Sigma}$ and $\hat{\mu}$ are respectively the covariance matrix and mean return estimates of individual stocks. Setting $c = 1$ assumes a mean-variance investor with the risk aversion parameter equal to 1, and setting $c = \frac{0.02}{\sqrt{\hat{\mu}^T\hat{\Sigma}^{-1}\hat{\mu}}}$ aims to set the variance of the portfolio equal to 0.02^2.

When $c = 1$, the optimal portfolio yields an annualized mean return of 13864.8% and a Sharpe ratio of 3.023, and when the risk is targeted to 2% per month, it yields an annualized mean return of 59% and a Sharpe ratio of 3.29. These values are higher than those of the equal-weight long-short portfolio under the Return criterion, which are 51% and 2.84. This result is astonishing considering that the optimization involves the estimation of the covariance matrix of thousands of stocks.

Of course, the optimal portfolio uses a high level of leverage, especially when $c = 1$, and is not implementable. A stock-level optimization with thousands of stocks is also unlikely to occur in practice. However, the result suggests that the mean and covariance estimates provide valuable information and demonstrates how the DM models can be used to build an optimal portfolio. For example, one might build an optimal long-only portfolio using stocks in the H class with suitable short-sale constraints.

The estimation method proposed here offers another way to estimate the covariance matrix of a large dimensional system, and it would be interesting to compare it with other methods for large dimensional systems such as the one in Ledoit et al. (2012). This and other topics related to input parameter estimation and portfolio optimization via the deep momentum are left for future research.

B Robustness check

This section examines the robustness of the proposed models from various perspectives.

B.1 Variation across Random Trials

The performance of a machine learning model varies each time it is trained with a new sequence of random numbers. Therefore, it is important to repeat the same experiment many times and check the robustness of the model. To identify the potential best- and worst-case scenarios, Table 11 reports the basic statistics of the mean return and Sharpe ratio obtained from fifty random trials.

The models are surprisingly robust to random trials. The standard deviations are remarkably small, particularly for the mean return: *e.g.*, the standard deviation of the mean return of the long-short portfolio is only about 2% of the average. The standard deviation of the Sharpe ratio is higher but still around 4% of the average. The small variation makes the performance under the worst-case scenario still impressive. For instance, the worst-case mean return and Sharpe ratio of the equal-weighted long-short portfolio are respectively 49% and 2.63 under Return and 45% and 2.45 under Sharpe. The corresponding best-case values are 54% and 3.01 and 49% and 2.98, respectively. This result suggests that the DM models are robust to the sequence of random numbers.

B.2 Stacking vs. Rolling Windows

The size of the estimation sample is a critical factor that needs to be determined prior to model estimation. While a larger sample is considered better for machine learning, it is not necessarily the case in asset pricing applications, where there is no guarantee that the training, validation, and test sets have the same distribution. In order to examine the effect of the sample size, previous experiments are repeated under the same conditions except for the training sample, whose size is now fixed at twenty years.

Table 12 (row Rolling) reports the results from one model (MOM-SZ-NOM). The mean returns and Sharp ratios from the rolling window are comparable to those from the stacking window, suggesting that twenty years of data is sufficient to train the models. Nevertheless, it is recommended to include all available data to span different market states as broadly as possible.

B.3 Small Firm Effects

The same experiments are repeated excluding micro firms. Micro firms are defined as the firms at the bottom 5% in terms of market capitalization. As shown in Table 12 (row Ex-Micro), the mean returns and Sharpe ratios decrease only slightly when micro firms are excluded. When more firms are removed from the sample by excluding stocks below NYSE-size 10% quantile (row Ex-NYSE10) or 20% quantile (row Ex-NYSE20), the financial performance deteriorates further. Still, the long-short strategy remains profitable: the annualized mean return and Sharpe ratio are respectively 24% and 1.20 in the first case, and 20% and 1.08 in the latter case. This result suggests that the strategy is implementable even for a large-sized fund.

B.4 Sample Period of Gu et al. (2020) and Extended Sample Period

Table 12 (row GKX) also reports the results from the sample period of Gu et al. (2020), namely 1987.01 to 2016.12. The DM models perform slightly worse in this period compared to the test period of this paper. Nevertheless, both the mean return and Sharpe ratio are generally higher than those reported in Gu et al. (2020) for both equal-weighted and value-weighted portfolios: *e.g.*, the mean return and Sharpe ratio under the Return criterion are respectively 52% and 2.57 for the equal-weighted portfolio and 34% and 1.49 for the value-weighted portfolio, whereas the corresponding values from Gu et al. (2020) are 39% and 2.35 (DNN4), and 25% and 1.25 (DNN3), respectively. This result is particularly noteworthy considering the smaller number of the input features used in the DM models.

The DM models continue to perform superbly when tested from 1947.05. The mean return and Sharpe ratio are respectively 36% and 1.77 in this period, higher than those from the original test period (see row From1947 of Table 12).

B.5 Number of Classes

The number of classes can affect the performance of the DM model. Too few classes will make it difficult to distinguish stocks at a granular level, whereas too many classes will result in insufficient samples in each class for training. To examine the impact of the number of classes, Table 13 compares the portfolio performances obtained from the DM model with the number of classes varying from 5 to 100. In each panel, the upper part reports the performance of the portfolios constructed from the top and bottom quantiles: *e.g.*, when $K = 100$, the H portfolio is made of the top 1% stocks. The lower part reports the performance of the portfolios constructed from the top and bottom deciles, *i.e.*, the portfolios are made of top/bottom 10% regardless of the number of classes. The DM model used is MOM-SZ-NOM with Return reclassification.

As the number of classes increases, the mean return of the top (bottom) quantile increases (decreases) monotonically, resulting in a monotonic increase in the mean return of the long-short portfolio. The Sharpe ratio also increases with the number of classes when it is small, but remains stable when $K > 20$. It is noteworthy that the performances of the portfolios when $K = 20$ and 50 are comparable to the results from top/bottom 5% and 1% in Table 10, respectively. When stocks are regrouped to construct decile portfolios, the mean return tends to increase with the number of classes, but the Sharpe ratio does not improve. Overall, increasing the number of classes is found to help improve the mean return but have a limited effect on the Sharpe ratio.

B.6 Correlation between Models

Tabel 14 reports the correlations between the portfolio returns obtained from the variations of the DM model. It shows that they have a very high correlation with the correlation coefficient close to or above 0.9, supporting the previous claim that there is no discernible difference between two sets of input features and between nominal and ordinal classifiers.

Table 2: Classification performance

This table reports the classification performance of the four classifiers listed in Table 1. The performance metrics are defined in Section 3.2.5. The training set (Train) and validation set (Valid) performances are the average performances obtained from the yearly retraining over the period from 1974.12 to 2016.12, and the test set performance (Test) is obtained from the test period from 1975.01 to 2017.01. For the performance measures of the Jegadeesh and Titman (1993)'s momentum strategy (JT), the predicted class is assumed to have the probability 1 and other classes 0.

	Train				Valid				Test			
	Loss	Acc	MSE	MAE	Loss	Acc	MSE	MAE	Loss	Acc	MSE	MAE
MOM-NOM	2.24	15.30	16.05	3.20	2.25	14.70	16.15	3.22	2.25	14.80	16.08	3.21
MOM-ORD	2.42	14.84	15.99	3.20	2.56	13.71	16.04	3.21	2.51	13.81	15.99	3.20
RET-NOM	2.23	15.95	15.91	3.18	2.25	14.84	16.17	3.21	2.24	15.01	16.05	3.19
RET-ORD	2.78	15.35	15.76	3.16	3.32	13.43	15.90	3.19	3.23	13.56	15.88	3.18
JT									18.02	13.07	15.50	3.08

Table 3: Class-level classification performance

This table reports the class-level classification performance of the momentum-based classifiers in the test period, 1975.01 to 2017.01. The performance metrics are defined in Section 3.2.5. The last column is the percentage of the stocks predicted to be in each class, and the last row (wavg) is the average performance weighted by this percentage. For the performance measures of the Jegadeesh and Titman (1993)'s momentum strategy (JT), the predicted class is assumed to have the probability 1 and other classes 0.

	Precision	Recall	f1-score	MSEP	MAEP	Support	Predicted	(%)
MOM-NOM								
H	16.48	22.52	19.03	20.40	3.64	244459	333994	13.68
2	11.93	10.52	11.18	15.62	3.20	244413	215567	8.83
3	11.89	7.00	8.80	13.30	2.95	244365	143966	5.89
4	13.20	15.12	14.09	11.80	2.77	244311	279783	11.46
5	13.83	17.83	15.57	11.40	2.72	244250	314946	12.90
6	12.96	7.25	9.28	12.04	2.80	244214	136628	5.59
7	11.84	6.76	8.60	13.24	2.94	244166	139485	5.71
8	11.32	6.41	8.18	14.69	3.11	244102	138229	5.66
9	12.08	11.78	11.92	16.85	3.33	244046	238009	9.75
L	20.83	42.83	28.02	21.43	3.70	243991	501709	20.54
avg	13.64	14.80	13.47	15.08	3.12	2442317	2442317	100.00
wavg	14.80	19.57	16.13	16.08	3.21	2442317	2442317	100.00
MOM-ORD								
H	18.77	12.71	15.15	21.68	3.74	244459	165629	6.78
2	11.78	5.29	7.29	16.84	3.32	244413	109730	4.49
3	11.29	7.65	9.09	14.30	3.05	244365	165481	6.78
4	11.93	12.82	12.33	13.40	2.95	244311	262579	10.75
5	12.35	17.80	14.56	12.97	2.90	244250	352066	14.42
6	11.77	16.60	13.76	13.33	2.94	244214	344504	14.11
7	11.09	13.64	12.21	14.11	3.03	244166	300145	12.29
8	10.79	9.50	10.08	15.70	3.20	244102	214849	8.80
9	12.44	8.72	10.23	18.04	3.43	244046	171102	7.01
L	22.88	33.41	27.16	22.14	3.75	243991	356232	14.59
avg	13.51	13.81	13.19	16.25	3.23	2442317	2442317	100.00
wavg	13.81	15.98	14.44	15.99	3.20	2442317	2442317	100.00
JT								
H	13.64	13.64	13.64	29.14	4.43	244459	244459	10.01
2	10.99	10.99	10.99	18.76	3.52	244413	244413	10.01
3	11.41	11.41	11.41	12.33	2.83	244365	244365	10.01
4	12.18	12.18	12.18	8.41	2.37	244311	244311	10.00
5	12.44	12.44	12.44	6.78	2.18	244250	244250	10.00
6	11.61	11.61	11.61	7.28	2.27	244214	244214	10.00
7	10.80	10.80	10.80	9.83	2.59	244166	244166	10.00
8	10.61	10.61	10.61	14.28	3.05	244102	244102	9.99
9	12.36	12.36	12.36	20.43	3.56	244046	244046	9.99
L	24.66	24.66	24.66	27.78	3.99	243991	243991	9.99
avg	13.07	13.07	13.07	15.50	3.08	2442317	2442317	100.00
wavg	13.07	13.07	13.07	15.50	3.08	2442317	2442317	100.00

Table 4: Reclassification performance

This table reports the class-level classification performance of the momentum-based classifiers after reclassification. 'Org' denotes the neural network classifier without reclassification and other rows denote reclassification criteria defined in Section 3.3. The results are from the test period, 1975.01 to 2017.01.

		MOM-NOM					MOM-ORD				
		Precision	Recall	f1-score	MSEP	MAEP	Precision	Recall	f1-score	MSEP	MAEP
Org	H	16.48	22.52	19.03	20.40	3.64	18.77	12.71	15.15	21.68	3.74
	L	20.83	42.83	28.02	21.43	3.70	22.88	33.41	27.16	22.14	3.75
	avg	13.64	14.80	13.47	15.08	3.12	13.51	13.81	13.19	16.25	3.23
Rank	H	18.57	18.62	18.59	21.89	3.76	16.99	16.96	16.98	20.64	3.65
	L	25.88	25.95	25.92	23.02	3.80	25.04	25.07	25.06	22.58	3.77
	avg	14.22	14.23	14.23	16.07	3.21	13.62	13.62	13.62	15.99	3.20
Prob	H	17.67	17.64	17.66	20.98	3.68	16.22	16.19	16.20	19.77	3.57
	L	25.93	25.94	25.93	23.01	3.80	25.45	25.45	25.45	22.73	3.78
	avg	13.99	13.99	13.99	16.08	3.21	13.32	13.32	13.32	15.99	3.20
PrDf1	H	13.25	13.25	13.25	17.37	3.34	13.42	13.42	13.42	17.38	3.35
	L	24.54	24.54	24.54	22.04	3.72	24.32	24.32	24.32	21.85	3.71
	avg	13.03	13.03	13.03	16.08	3.21	13.01	13.01	13.01	15.99	3.20
PrDf5	H	8.49	8.49	8.49	13.61	2.94	9.09	9.09	9.09	14.06	3.00
	L	22.50	22.50	22.50	20.88	3.62	22.51	22.51	22.51	20.80	3.62
	avg	11.73	11.73	11.73	16.08	3.21	11.84	11.84	11.84	15.99	3.20
Return	H	13.17	13.17	13.17	17.26	3.32	13.81	13.81	13.81	17.50	3.35
	L	22.58	22.58	22.58	20.96	3.62	23.09	23.09	23.09	21.02	3.64
	avg	12.54	12.54	12.54	16.08	3.21	12.77	12.77	12.77	15.99	3.20
Sharpe	H	9.86	9.86	9.86	14.60	3.05	10.38	10.38	10.38	14.82	3.07
	L	22.11	22.11	22.11	20.70	3.60	22.79	22.79	22.79	20.86	3.62
	avg	11.89	11.89	11.89	16.08	3.21	12.12	12.12	12.12	15.99	3.20

Table 5: Financial performance: without size dummies

This table reports the annualized mean excess returns (MR) and Sharpe ratios (SR) of the long (H), short (L), and long-short (H-L) portfolios obtained from the DM models without size dummies. The returns are calculated over the test period from 1975.01 to 2017.01.

(a) Value-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
MOM-NOM	H	0.16	0.57	0.17	0.52	0.17	0.56	0.17	0.70	0.14	0.76	0.17	0.73	0.16	0.78
	L	0.01	0.05	-0.04	-0.07	-0.05	-0.15	-0.09	-0.34	-0.10	-0.36	-0.09	-0.39	-0.09	-0.39
	H-L	0.16	1.09	0.21	1.13	0.22	1.11	0.26	1.34	0.24	1.23	0.26	1.32	0.25	1.27
MOM-ORD	H	0.19	0.56	0.15	0.51	0.16	0.59	0.17	0.70	0.15	0.74	0.18	0.71	0.17	0.76
	L	-0.03	-0.10	-0.04	-0.11	-0.05	-0.15	-0.09	-0.30	-0.09	-0.39	-0.11	-0.42	-0.11	-0.42
	H-L	0.22	1.03	0.19	1.02	0.21	1.03	0.26	1.27	0.24	1.18	0.29	1.34	0.27	1.28
RET-NOM	H	0.15	0.52	0.15	0.45	0.16	0.52	0.16	0.69	0.15	0.80	0.16	0.71	0.15	0.78
	L	-0.01	-0.02	-0.08	-0.25	-0.08	-0.28	-0.09	-0.26	-0.07	-0.30	-0.07	-0.26	-0.06	-0.34
	H-L	0.16	1.10	0.23	1.28	0.24	1.20	0.25	1.15	0.22	1.07	0.23	1.08	0.21	1.07
RET-ORD	H	0.14	0.40	0.12	0.44	0.12	0.51	0.13	0.57	0.13	0.73	0.15	0.65	0.14	0.73
	L	-0.05	2.42	-0.06	-0.17	-0.06	-0.22	-0.07	-0.26	-0.06	-0.28	-0.09	-0.37	-0.09	-0.34
	H-L	0.20	0.84	0.18	1.04	0.19	0.94	0.21	1.08	0.19	1.01	0.24	1.20	0.23	1.16
JT	H	0.14	0.62												
	L	-0.04	-0.07												
	H-L	0.19	0.61												

(b) Equal-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
MOM-NOM	H	0.29	0.93	0.30	0.86	0.32	0.95	0.31	1.13	0.25	1.19	0.32	1.14	0.28	1.19
	L	0.00	0.02	-0.05	-0.14	-0.05	-0.15	-0.10	-0.33	-0.11	-0.44	-0.11	-0.44	-0.12	-0.43
	H-L	0.28	2.45	0.34	2.19	0.37	2.23	0.41	2.36	0.37	2.11	0.43	2.41	0.40	2.25
MOM-ORD	H	0.35	0.96	0.28	0.85	0.31	0.96	0.31	1.08	0.25	1.12	0.33	1.13	0.30	1.16
	L	-0.03	-0.07	-0.05	-0.15	-0.05	-0.16	-0.10	-0.35	-0.11	-0.42	-0.12	-0.46	-0.13	-0.47
	H-L	0.38	2.05	0.33	1.99	0.36	2.03	0.41	2.15	0.37	1.94	0.46	2.28	0.42	2.11
RET-NOM	H	0.26	0.89	0.28	0.82	0.29	0.90	0.29	1.13	0.24	1.22	0.30	1.16	0.26	1.23
	L	0.01	0.05	-0.04	-0.07	-0.04	-0.11	-0.09	-0.26	-0.10	-0.38	-0.10	-0.38	-0.10	-0.40
	H-L	0.25	2.57	0.32	2.55	0.34	2.39	0.38	2.29	0.34	1.82	0.40	2.32	0.36	2.00
RET-ORD	H	0.31	0.87	0.24	0.79	0.26	0.89	0.26	1.02	0.23	1.11	0.30	1.13	0.26	1.18
	L	-0.04	-0.07	-0.04	-0.07	-0.04	-0.07	-0.08	-0.28	-0.08	-0.33	-0.11	-0.41	-0.11	-0.42
	H-L	0.36	1.96	0.28	2.28	0.30	2.10	0.34	2.20	0.31	1.76	0.41	2.49	0.38	2.10
JT	H	0.17	0.71												
	L	0.07	0.21												
	H-L	0.10	0.37												

Table 6: Financial performance: with size dummies

This table reports the annualized mean excess returns (MR) and Sharpe ratios (SR) of the long (H), short (L), and long-short (H-L) portfolios obtained from the DM models including size dummies. The returns are calculated over the test period from 1975.01 to 2017.01.

(a) Value-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
MOM-SZ-NOM	H	0.21	0.67	0.23	0.64	0.24	0.72	0.25	0.95	0.18	0.94	0.23	0.94	0.20	0.95
	L	-0.01	-0.02	-0.08	-0.25	-0.10	-0.27	-0.13	-0.47	-0.13	-0.54	-0.12	-0.48	-0.12	-0.48
	H-L	0.22	1.39	0.32	1.47	0.34	1.51	0.38	1.77	0.31	1.61	0.35	1.66	0.31	1.61
MOM-SZ-ORD	H	0.26	0.76	0.25	0.75	0.24	0.82	0.25	0.97	0.18	0.93	0.25	1.02	0.23	1.11
	L	-0.02	-0.05	-0.05	-0.13	-0.06	-0.11	-0.11	-0.35	-0.13	-0.49	-0.14	-0.53	-0.13	-0.57
	H-L	0.28	1.36	0.30	1.50	0.29	1.37	0.36	1.68	0.30	1.55	0.39	1.80	0.36	1.86
RET-SZ-NOM	H	0.18	0.58	0.20	0.56	0.20	0.62	0.20	0.83	0.15	0.85	0.18	0.79	0.16	0.85
	L	-0.01	-0.02	-0.11	-0.30	-0.11	-0.34	-0.11	-0.34	-0.09	-0.37	-0.08	-0.35	-0.07	-0.30
	H-L	0.19	1.31	0.30	1.54	0.32	1.52	0.30	1.47	0.24	1.23	0.26	1.22	0.23	1.21
RET-SZ-ORD	H	0.22	0.61	0.18	0.58	0.17	0.65	0.20	0.82	0.16	0.86	0.22	0.93	0.20	1.03
	L	-0.05	-0.13	-0.07	-0.23	-0.08	-0.25	-0.08	-0.30	-0.07	-0.33	-0.11	-0.42	-0.10	-0.42
	H-L	0.27	1.26	0.25	1.44	0.25	1.25	0.29	1.49	0.23	1.20	0.33	1.58	0.30	1.57
JT	H	0.14	0.62												
	L	-0.04	-0.07												
	H-L	0.19	0.61												

(b) Equal-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
MOM-SZ-NOM	H	0.30	1.05	0.36	1.06	0.37	1.12	0.35	1.32	0.26	1.28	0.36	1.33	0.31	1.35
	L	-0.01	-0.02	-0.05	-0.15	-0.06	-0.11	-0.12	-0.46	-0.15	-0.60	-0.15	-0.63	-0.16	-0.65
	H-L	0.30	2.60	0.41	2.61	0.43	2.56	0.48	2.66	0.41	2.39	0.51	2.84	0.47	2.64
MOM-SZ-ORD	H	0.38	1.13	0.35	1.08	0.37	1.16	0.35	1.31	0.27	1.26	0.38	1.34	0.33	1.38
	L	-0.03	-0.10	-0.05	-0.15	-0.05	-0.17	-0.12	-0.45	-0.14	-0.57	-0.17	-0.65	-0.17	-0.69
	H-L	0.40	2.44	0.40	2.56	0.42	2.54	0.47	2.58	0.41	2.41	0.54	2.77	0.50	2.65
RET-SZ-NOM	H	0.28	1.01	0.32	0.96	0.34	1.07	0.33	1.32	0.24	1.30	0.33	1.33	0.28	1.37
	L	0.00	0.01	-0.04	-0.07	-0.05	-0.14	-0.11	-0.39	-0.13	-0.52	-0.13	-0.54	-0.14	-0.54
	H-L	0.28	2.83	0.36	2.90	0.39	2.85	0.44	2.69	0.37	2.13	0.46	2.81	0.41	2.45
RET-SZ-ORD	H	0.39	1.10	0.30	1.00	0.31	1.06	0.31	1.26	0.24	1.21	0.34	1.31	0.30	1.38
	L	-0.04	-0.14	-0.05	-0.14	-0.05	-0.15	-0.11	-0.35	-0.11	-0.46	-0.15	-0.58	-0.16	-0.61
	H-L	0.43	2.43	0.35	2.79	0.36	2.64	0.42	2.62	0.35	2.12	0.49	2.88	0.45	2.61
JT	H	0.17	0.90												
	L	0.07	0.34												
	H-L	0.10	0.37												

Table 7: Factor regression

This table reports the factor regression results of the DM portfolio: MOM-SZ-MOM with Return reclassification. The portfolio returns are regressed on the Fama-French five factors (MKT, SMB, HML, RMW, CMA) plus the momentum (MOM) and short-term reversal (STR) factors over the test period from 1975.01 to 2017.01. The *t*-statistics are the Newey-West adjusted *t*-statistics.

(a) Value-weighted

	Estimates								*t*-statistics							
	α	MKT	SMB	HML	RMW	CMA	MOM	STR	α	MKT	SMB	HML	RMW	CMA	MOM	STR
H	0.012	1.100	0.639	-0.034	-0.383	-0.215	0.057	-0.082	5.764	19.443	6.598	-0.308	-2.738	-1.453	0.713	-0.728
2	0.008	1.034	0.536	-0.037	-0.141	-0.019	0.015	-0.114	6.410	35.458	7.119	-0.598	-1.526	-0.207	0.295	-1.373
3	0.006	1.038	0.406	-0.024	-0.085	0.001	0.027	-0.127	4.414	38.019	8.045	-0.494	-0.962	0.023	0.889	-1.366
4	0.003	1.020	0.362	0.007	0.022	0.017	-0.016	-0.070	3.515	37.071	7.602	0.183	0.432	0.278	-0.478	-1.090
5	0.002	1.011	0.302	0.019	0.027	0.011	-0.018	-0.006	1.837	31.739	6.905	0.424	0.488	0.163	-0.502	-0.084
6	-0.000	1.023	0.208	0.091	0.028	-0.072	-0.021	0.060	-0.442	30.554	4.842	2.238	0.582	-1.084	-0.561	1.058
7	-0.002	1.032	0.186	0.056	-0.015	-0.094	-0.054	0.074	-2.046	26.116	3.574	0.982	-0.231	-1.221	-1.015	1.125
8	-0.005	1.060	0.167	0.058	-0.025	-0.085	-0.072	0.102	-4.290	27.130	3.139	1.044	-0.361	-1.163	-1.351	1.868
9	-0.009	1.074	0.191	0.051	-0.082	0.021	-0.121	0.096	-6.357	24.069	2.626	0.841	-0.901	0.222	-2.035	1.651
L	-0.016	1.093	0.447	-0.091	-0.335	0.050	-0.205	-0.010	-7.636	21.809	5.527	-1.078	-2.819	0.321	-2.367	-0.094
H-L	0.024	0.010	0.197	0.046	-0.044	-0.256	0.254	-0.077	6.631	0.102	1.280	0.296	-0.191	-1.103	1.670	-0.381

(b) Equal-weighted

	Estimates								*t*-statistics							
	α	MKT	SMB	HML	RMW	CMA	MOM	STR	α	MKT	SMB	HML	RMW	CMA	MOM	STR
H	0.023	0.937	0.929	-0.052	-0.379	0.029	-0.282	0.324	6.971	11.689	9.361	-0.454	-3.630	0.141	-2.191	2.254
2	0.010	0.922	0.841	0.043	-0.045	0.075	-0.153	0.196	5.759	18.321	14.561	0.612	-0.623	0.569	-1.820	2.550
3	0.007	0.912	0.823	0.121	0.057	0.093	-0.105	0.137	6.652	32.742	17.580	2.233	1.082	1.159	-2.077	3.364
4	0.005	0.906	0.823	0.128	0.079	0.080	-0.093	0.098	6.796	44.845	17.948	2.929	1.759	1.274	-2.601	3.781
5	0.003	0.902	0.796	0.128	0.054	0.070	-0.092	0.088	5.510	48.821	19.178	3.436	1.405	1.278	-3.484	3.784
6	0.001	0.909	0.758	0.145	0.034	0.020	-0.105	0.066	1.745	45.139	18.443	3.580	0.779	0.401	-5.095	2.848
7	-0.001	0.901	0.763	0.120	-0.004	-0.009	-0.140	0.077	-1.052	37.816	15.564	2.559	-0.070	-0.159	-6.338	2.467
8	-0.003	0.919	0.770	0.119	-0.057	-0.028	-0.182	0.073	-4.636	32.210	13.772	1.953	-0.791	-0.400	-6.576	1.715
9	-0.007	0.917	0.782	0.053	-0.233	-0.015	-0.219	0.010	-7.550	29.878	15.026	0.801	-3.239	-0.181	-8.815	0.239
L	-0.016	0.910	0.878	-0.157	-0.718	0.089	-0.348	-0.150	-10.692	23.330	16.743	-2.167	-10.014	0.869	-8.044	-2.702
H-L	0.035	0.030	0.056	0.094	0.343	-0.051	0.059	0.469	9.278	0.408	0.512	0.820	2.552	-0.234	0.444	3.348

Table 8: Effects of transaction costs

This table reports the performance of the long (H), short (L), and long-short (H-L) portfolios obtained from the MOM-SZ-NOM classifier with Return reclassification criterion. The transaction cost is assumed to be constant at 10 or 30 basis points, or vary with time and market capitalization. The exact formula for the varying transaction cost can be found in DeMiguel et al. (2020).

(a) Value-weighted

	Gross			Cost = 10bp			Cost = 30bp			Varying cost		
	H	L	H-L	H	L	H-L	H	L	H-L	H	L	H-L
Mean	0.235	-0.116	0.350	0.215	-0.136	0.310	0.175	-0.176	0.230	0.083	-0.268	0.050
Stddev	0.249	0.240	0.210	0.250	0.240	0.210	0.250	0.240	0.210	0.250	0.240	0.210
Sharpe	0.944	-0.483	1.660	0.863	-0.567	1.460	0.706	-0.733	1.090	0.335	-1.112	0.220
Cumulative	10.437	-4.220	13.660	9.620	-5.120	12.000	7.960	-6.930	8.680	4.180	-11.170	0.950
MDD	0.585	0.990	0.350	0.600	1.000	0.420	0.640	1.000	0.550	0.660	1.000	0.890
Turnover	0.832	0.820	1.660	0.830	0.820	1.660	0.830	0.820	1.660	0.830	0.820	1.660

(b) Equal-weighted

	Gross			Cost = 10bp			Cost = 30bp			Varying cost		
	H	L	H-L	H	L	H-L	H	L	H-L	H	L	H-L
Mean	0.365	-0.150	0.520	0.346	-0.167	0.480	0.310	-0.201	0.410	0.200	-0.293	0.200
Stddev	0.274	0.240	0.180	0.270	0.240	0.180	0.270	0.240	0.180	0.270	0.250	0.190
Sharpe	1.332	-0.615	2.840	1.263	-0.684	2.650	1.131	-0.824	2.270	0.730	-1.196	1.060
Cumulative	15.595	-5.690	20.620	14.870	-6.350	19.290	13.420	-7.670	16.620	9.060	-11.950	8.230
MDD	0.407	1.000	0.100	0.430	1.000	0.100	0.460	1.000	0.140	0.490	1.000	0.560
Turnover	0.763	0.700	1.440	0.760	0.700	1.440	0.760	0.700	1.440	0.760	0.700	1.440

Table 9: Comparison with alternative models

This table compares a DM model (MOM-SZ-NOM with Return reclassification) with a linear classifier (Logistic) and three regressors. The linear regressor (Linear) is simply an ordinary least squares regressor with the dependent variable defined as the ranking of the stocks based on the one-month ahead return. The ranking is normalized so that the value lies in $[-1, 1]$. 'Ordinal1' is a feed-forward neural network regressor that predicts the ranking. 'Ordinal2' is the same as Ordinal1 except that the target variable is the return class, *i.e.* grouped ranking. The regressors predict stocks' rankings (classes) and use them to classify the stocks into ten groups. A long-short portfolio is constructed from the stocks in the first and the last groups. The logistic classifier is augmented with Return reclassification. All models use the same input features, *i.e.*, the normalized momentums, cross-sectional means of the momentums, and size dummies.

(a) Value-weighted

	Classifiers						Regressors								
	MOM-SZ-NOM			Logistic			Linear			Ordinal1			Ordinal2		
	H	L	H-L	H	L	H-L	H	L	H-L	H	L	H-L	H	L	H-L
Mean	0.235	-0.116	0.350	0.168	-0.019	0.187	0.167	0.029	0.138	0.185	-0.092	0.277	0.172	-0.077	0.249
Stddev	0.249	0.240	0.212	0.281	0.259	0.212	0.312	0.205	0.236	0.243	0.230	0.211	0.211	0.257	0.206
Sharpe	0.944	-0.483	1.656	0.598	-0.073	0.880	0.535	0.141	0.585	0.761	-0.400	1.317	0.815	-0.300	1.212
Cumulative	10.437	-4.215	13.655	7.239	-0.274	6.859	6.859	2.271	4.604	8.402	-3.094	10.650	8.160	-2.730	9.481
MDD	0.585	0.992	0.345	0.649	0.912	0.593	0.734	0.783	0.719	0.786	0.995	0.417	0.545	0.989	0.424
Turnover	0.832	0.821	1.655	0.695	0.882	1.577	0.792	0.899	1.692	0.842	0.835	1.672	0.809	0.842	1.647

(b) Equal-weighted

	Classifiers						Regressors								
	MOM-SZ-NOM			Logistic			Linear			Ordinal1			Ordinal2		
	H	L	H-L	H	L	H-L	H	L	H-L	H	L	H-L	H	L	H-L
Mean	0.365	-0.150	0.515	0.282	-0.034	0.317	0.321	-0.036	0.357	0.298	-0.119	0.417	0.265	-0.122	0.387
Stddev	0.274	0.244	0.181	0.285	0.248	0.174	0.289	0.239	0.185	0.249	0.243	0.161	0.215	0.253	0.159
Sharpe	1.332	-0.615	2.842	0.989	-0.137	1.816	1.111	-0.151	1.929	1.197	-0.490	2.584	1.233	-0.482	2.430
Cumulative	15.595	-5.694	20.624	12.012	-0.786	12.549	13.608	-0.757	14.147	13.051	-4.360	16.747	11.972	-4.603	15.522
MDD	0.407	0.999	0.095	0.577	0.928	0.415	0.560	0.929	0.298	0.488	0.999	0.194	0.574	0.997	0.263
Turnover	0.763	0.699	1.444	0.705	0.767	1.464	0.638	0.839	1.466	0.740	0.712	1.437	0.764	0.703	1.458

Table 10: Methods to enhance performance

This table reports the annualized mean excess returns (MR) and Sharpe ratios (SR) of the long (H), short (L), and long-short (H-L) portfolios obtained from the performance enhancement methods described in Section A. The base model is MOM-SZ-NOM. The returns are calculated over the test period from 1975.01 to 2017.01.

(a) Value-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe		JT	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
5%	H	0.21	0.67	0.27	0.65	0.28	0.68	0.28	0.93	0.19	0.95	0.26	0.90	0.22	0.97	0.16	0.60
	L	-0.01	-0.02	0.03	0.09	0.02	0.06	-0.22	-0.70	-0.20	-0.74	-0.19	-0.70	-0.18	-0.69	-0.12	-0.30
	H-L	0.22	1.39	0.24	0.83	0.25	0.86	0.50	1.89	0.39	1.77	0.45	1.75	0.40	1.76	0.28	0.77
1%	H	0.21	0.67	0.46	0.75	0.46	0.76	0.44	0.96	0.25	0.91	0.43	0.97	0.34	1.01	0.18	0.50
	L	-0.01	-0.02	0.06	0.12	0.06	0.13	-0.39	-0.94	-0.37	-1.06	-0.39	-1.04	-0.34	-1.02	-0.18	-0.26
	H-L	0.22	1.39	0.40	0.75	0.40	0.75	0.83	1.85	0.62	1.86	0.82	1.88	0.68	1.83	0.38	0.62
Opt	H	0.21	0.67	0.23	0.64	0.24	0.72	0.25	0.95	0.18	0.94	0.23	0.94	0.20	0.95	0.14	0.62
	L	-0.01	-0.02	-0.08	-0.25	-0.10	-0.27	-0.13	-0.47	-0.13	-0.54	-0.12	-0.48	-0.12	-0.48	-0.04	-0.07
	H-L	0.20	0.39	0.37	0.63	0.41	0.67	0.50	0.93	0.42	0.85	0.45	0.95	0.41	0.89	0.22	0.29
I10	H	0.22	0.71	0.26	0.69	0.25	0.72	0.29	1.08	0.19	0.93	0.30	1.06	0.23	1.03	0.14	0.62
	L	-0.06	-0.22	-0.12	-0.33	-0.13	-0.36	-0.19	-0.62	-0.18	-0.65	-0.16	-0.62	-0.17	-0.65	-0.04	-0.07
	H-L	0.28	1.49	0.38	1.52	0.39	1.52	0.47	1.91	0.37	1.59	0.46	1.85	0.40	1.82	0.19	0.61
I50	H	0.30	0.87	0.33	0.82	0.35	0.89	0.32	0.99	0.27	0.69	0.34	0.91	0.30	0.82	0.14	0.62
	L	-0.09	-0.26	-0.17	-0.47	-0.18	-0.51	-0.24	-0.73	-0.27	-0.88	-0.24	-0.78	-0.24	-0.84	-0.04	-0.07
	H-L	0.39	1.58	0.51	1.79	0.53	1.80	0.56	1.75	0.54	1.27	0.58	1.64	0.55	1.50	0.19	0.61

(b) Equal-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe		JT	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
5%	H	0.30	1.05	0.46	1.16	0.47	1.20	0.44	1.39	0.30	1.34	0.45	1.41	0.36	1.42	0.17	0.63
	L	-0.01	-0.02	0.06	0.18	0.05	0.16	-0.20	-0.68	-0.23	-0.85	-0.23	-0.85	-0.24	-0.91	0.09	0.23
	H-L	0.30	2.60	0.40	2.09	0.42	2.05	0.64	2.86	0.53	2.53	0.68	2.93	0.60	2.74	0.07	0.21
1%	H	0.30	1.05	0.84	1.43	0.84	1.43	0.74	1.51	0.44	1.28	0.76	1.52	0.58	1.45	0.14	0.43
	L	-0.01	-0.02	0.17	0.39	0.17	0.40	-0.39	-1.09	-0.41	-1.24	-0.42	-1.20	-0.41	-1.29	0.19	0.34
	H-L	0.30	2.60	0.67	1.62	0.67	1.61	1.14	2.79	0.85	2.37	1.18	2.73	0.99	2.54	-0.05	-0.11
Opt	H	0.30	1.05	0.36	1.06	0.37	1.12	0.35	1.32	0.26	1.28	0.36	1.33	0.31	1.35	0.17	0.71
	L	-0.01	-0.02	-0.05	-0.15	-0.06	-0.11	-0.12	-0.46	-0.15	-0.60	-0.15	-0.63	-0.16	-0.65	0.07	0.21
	H-L	0.27	0.60	0.42	0.88	0.45	0.91	0.59	1.18	0.55	1.15	0.66	1.50	0.62	1.37	0.01	0.01
I10	H	0.36	1.22	0.43	1.20	0.45	1.28	0.44	1.43	0.32	1.31	0.48	1.50	0.39	1.37	0.17	0.71
	L	-0.05	-0.15	-0.09	-0.28	-0.10	-0.29	-0.17	-0.57	-0.21	-0.77	-0.21	-0.79	-0.22	-0.84	0.07	0.21
	H-L	0.41	2.75	0.53	2.76	0.55	2.78	0.61	2.70	0.53	2.36	0.69	2.91	0.61	2.52	0.10	0.37
I50	H	0.44	1.34	0.50	1.29	0.54	1.38	0.55	1.28	0.36	0.85	0.55	1.20	0.48	1.04	0.17	0.71
	L	-0.08	-0.30	-0.13	-0.41	-0.14	-0.44	-0.22	-0.72	-0.30	-1.02	-0.27	-0.96	-0.29	-1.04	0.07	0.21
	H-L	0.52	2.57	0.64	2.65	0.68	2.67	0.77	2.14	0.65	1.51	0.82	2.09	0.77	1.83	0.10	0.37

Table 11: Variation of financial performance across random trials

This table reports the basic statistics of the annualized mean excess returns (MR) and Sharpe ratios (SR) across fifty random trials. The base model is MOM-SZ-NOM.

(a) Value-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
Average	H	0.21	0.67	0.23	0.64	0.24	0.72	0.25	0.95	0.18	0.94	0.23	0.94	0.20	0.95
	L	-0.01	-0.02	-0.08	-0.25	-0.10	-0.27	-0.13	-0.47	-0.13	-0.54	-0.12	-0.48	-0.12	-0.48
	H-L	0.22	1.39	0.32	1.47	0.34	1.51	0.38	1.77	0.31	1.61	0.35	1.66	0.31	1.61
Max	H	0.24	0.77	0.27	0.73	0.27	0.78	0.26	1.06	0.19	1.05	0.25	1.03	0.22	1.12
	L	0.01	0.05	-0.06	-0.22	-0.07	-0.17	-0.11	-0.39	-0.11	-0.46	-0.09	-0.41	-0.09	-0.37
	H-L	0.25	1.64	0.35	1.61	0.37	1.69	0.41	1.91	0.33	1.74	0.38	1.81	0.34	1.80
Min	H	0.19	0.61	0.21	0.58	0.22	0.66	0.22	0.85	0.16	0.86	0.21	0.86	0.18	0.87
	L	-0.03	-0.09	-0.10	-0.27	-0.11	-0.32	-0.15	-0.55	-0.15	-0.63	-0.14	-0.57	-0.13	-0.60
	H-L	0.20	1.22	0.28	1.33	0.30	1.34	0.34	1.51	0.28	1.47	0.32	1.46	0.28	1.42
Std	H	1.11	3.10	1.44	3.02	1.25	2.94	1.08	4.67	0.63	3.87	0.72	3.75	0.87	4.88
(×100)	L	0.72	2.35	0.89	2.71	0.80	2.46	1.06	3.84	1.02	4.31	1.02	4.36	1.08	4.75
	H-L	1.34	9.56	1.62	6.88	1.53	7.89	1.42	8.62	1.23	7.00	1.29	7.75	1.51	8.08

(b) Equal-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
Average	H	0.30	1.05	0.36	1.06	0.37	1.12	0.35	1.32	0.26	1.28	0.36	1.33	0.31	1.35
	L	-0.01	-0.02	-0.05	-0.15	-0.06	-0.11	-0.12	-0.46	-0.15	-0.60	-0.15	-0.63	-0.16	-0.65
	H-L	0.30	2.60	0.41	2.61	0.43	2.56	0.48	2.66	0.41	2.39	0.51	2.84	0.47	2.64
Max	H	0.30	1.09	0.36	1.07	0.38	1.15	0.37	1.36	0.28	1.35	0.38	1.38	0.33	1.42
	L	0.00	0.01	-0.04	-0.11	-0.05	-0.15	-0.11	-0.42	-0.13	-0.52	-0.13	-0.55	-0.14	-0.57
	H-L	0.32	2.75	0.43	2.85	0.45	2.76	0.50	2.86	0.44	2.67	0.54	3.01	0.49	2.98
Min	H	0.29	1.03	0.35	1.04	0.36	1.09	0.34	1.27	0.25	1.21	0.35	1.29	0.29	1.30
	L	-0.02	-0.06	-0.06	-0.28	-0.07	-0.23	-0.14	-0.48	-0.16	-0.65	-0.16	-0.68	-0.17	-0.73
	H-L	0.29	2.36	0.39	2.42	0.41	2.38	0.46	2.46	0.39	2.24	0.49	2.63	0.45	2.45
Std	H	0.34	1.01	0.36	0.93	0.42	1.07	0.56	2.27	0.62	2.79	0.57	2.25	0.70	2.49
(×100)	L	0.36	1.25	0.44	1.42	0.44	1.45	0.44	1.71	0.56	2.53	0.54	2.45	0.56	2.68
	H-L	0.54	9.18	0.76	8.56	0.80	8.92	0.85	9.81	0.93	9.39	0.92	9.94	0.98	10.06

Table 12: Robustness test

This table reports the results of the robustness check in Appendix B. ‘Rolling’ is the results from the rolling window method, and ‘Ex-Micro’, ‘Ex-NYSE10’, and ‘Ex-NYSE20’ are respectively the results excluding stocks below the 5% quantile, NYSE-size10% quantile, and NYSE-size 20% quantile. ‘GKX’ is the results from the sample period of Gu et al. (2020), 1987.01-2016.12, and ’From1947’ is from the sample period 1947.05-2017.01. The base model is MOM-SZ-NOM.

(a) Value-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
Rolling	H	0.21	0.67	0.21	0.59	0.22	0.67	0.23	0.92	0.17	0.93	0.22	0.93	0.19	0.97
	L	-0.01	-0.02	-0.09	-0.24	-0.10	-0.27	-0.12	-0.40	-0.12	-0.50	-0.10	-0.42	-0.10	-0.40
	H-L	0.22	1.32	0.29	1.40	0.32	1.41	0.35	1.62	0.30	1.52	0.32	1.55	0.29	1.52
Ex-Micro	H	0.20	0.64	0.20	0.58	0.21	0.66	0.23	0.91	0.17	0.90	0.21	0.92	0.18	0.92
	L	-0.02	-0.06	-0.08	-0.20	-0.09	-0.24	-0.10	-0.35	-0.10	-0.42	-0.08	-0.35	-0.08	-0.30
	H-L	0.22	1.27	0.27	1.37	0.29	1.35	0.33	1.53	0.27	1.40	0.30	1.44	0.26	1.37
Ex-NYSE10	H	0.19	0.57	0.19	0.61	0.20	0.69	0.21	0.81	0.17	0.91	0.19	0.80	0.18	0.89
	L	-0.02	-0.06	0.01	0.05	-0.04	-0.14	-0.07	-0.20	-0.06	-0.28	-0.05	-0.14	-0.05	-0.15
	H-L	0.20	1.07	0.17	1.05	0.24	1.25	0.28	1.42	0.24	1.26	0.24	1.20	0.23	1.23
Ex-NYSE20	H	0.18	0.55	0.17	0.56	0.18	0.66	0.20	0.84	0.18	0.95	0.19	0.86	0.18	0.93
	L	0.01	0.04	0.04	0.14	-0.04	-0.14	-0.02	-0.06	-0.02	-0.08	-0.02	-0.08	-0.01	-0.03
	H-L	0.18	0.84	0.13	0.82	0.22	1.14	0.23	1.22	0.19	1.01	0.20	1.08	0.19	1.03
GKX	H	0.20	0.60	0.23	0.59	0.24	0.66	0.23	0.86	0.14	0.77	0.21	0.82	0.16	0.80
	L	-0.02	-0.04	-0.09	-0.22	-0.10	-0.27	-0.16	-0.50	-0.15	-0.54	-0.13	-0.50	-0.13	-0.52
	H-L	0.22	1.28	0.31	1.32	0.34	1.35	0.39	1.69	0.29	1.41	0.34	1.49	0.29	1.43
From1947	H	0.18	0.66	0.21	0.65	0.22	0.71	0.22	0.90	0.16	0.91	0.21	0.88	0.18	0.93
	L	-0.04	-0.11	-0.09	-0.30	-0.11	-0.37	-0.16	-0.64	-0.16	-0.62	-0.15	-0.64	-0.15	-0.63
	H-L	0.22	1.60	0.30	1.61	0.33	1.69	0.38	1.94	0.32	1.68	0.36	1.77	0.33	1.80

(b) Equal-weighted

		Org		Rank		Prob		PrDf1		PrDf5		Return		Sharpe	
		MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR	MR	SR
Rolling	H	0.29	1.02	0.33	1.02	0.34	1.08	0.33	1.32	0.25	1.30	0.33	1.32	0.28	1.37
	L	0.01	0.04	-0.04	-0.14	-0.04	-0.07	-0.12	-0.40	-0.14	-0.57	-0.15	-0.57	-0.15	-0.63
	H-L	0.28	2.62	0.37	2.74	0.39	2.65	0.44	2.66	0.39	2.32	0.48	2.92	0.44	2.65
Ex-Micro	H	0.23	0.84	0.24	0.77	0.25	0.86	0.26	1.12	0.21	1.14	0.25	1.11	0.23	1.19
	L	-0.03	-0.09	-0.08	-0.28	-0.09	-0.28	-0.14	-0.48	-0.16	-0.62	-0.15	-0.61	-0.16	-0.64
	H-L	0.26	2.43	0.32	2.63	0.34	2.55	0.40	2.51	0.37	2.25	0.41	2.66	0.38	2.43
Ex-NYSE10	H	0.20	0.65	0.20	0.69	0.21	0.72	0.22	0.88	0.19	1.01	0.21	0.91	0.20	1.00
	L	-0.02	-0.05	0.01	0.05	-0.03	-0.09	-0.06	-0.28	-0.07	-0.33	-0.07	-0.26	-0.07	-0.33
	H-L	0.23	1.62	0.19	1.46	0.24	1.53	0.28	1.62	0.27	1.52	0.28	1.63	0.27	1.58
Ex-NYSE20	H	0.19	0.64	0.17	0.63	0.20	0.73	0.20	0.86	0.18	1.01	0.20	0.92	0.18	0.98
	L	0.00	0.01	0.04	0.14	-0.02	-0.05	0.01	0.03	0.05	0.08	0.02	0.05	-0.03	-0.13
	H-L	0.19	1.27	0.14	1.08	0.22	1.42	0.24	1.39	0.21	1.17	0.23	1.35	0.21	1.21
GKX	H	0.30	0.99	0.37	1.00	0.38	1.07	0.36	1.23	0.24	1.16	0.36	1.25	0.30	1.25
	L	-0.01	-0.05	-0.04	-0.17	-0.05	-0.16	-0.12	-0.43	-0.15	-0.59	-0.16	-0.60	-0.16	-0.65
	H-L	0.31	2.31	0.41	2.31	0.43	2.29	0.48	2.36	0.39	2.05	0.52	2.57	0.46	2.33
From1947	H	0.27	1.02	0.32	1.00	0.34	1.08	0.33	1.25	0.25	1.21	0.33	1.23	0.29	1.26
	L	-0.02	-0.06	-0.06	-0.21	-0.07	-0.28	-0.13	-0.52	-0.15	-0.61	-0.16	-0.67	-0.15	-0.68
	H-L	0.29	2.81	0.38	2.66	0.41	2.66	0.46	2.79	0.40	2.45	0.49	2.82	0.44	2.72

Table 13: Effects of number of classes

This table examines the impact of the number of classes (K) by changing its value from 5 to 100. In each panel, the upper part reports the performance of the portfolios constructed from the top and bottom quantiles: *e.g.*, when $K = 100$, H portfolio is made of the top 1% stocks. The lower part reports the performance of the portfolios constructed from the top and bottom deciles, *i.e.*, the portfolios are made of top/bottom 10% regardless of the number of classes. The DM model used is MOM-SZ-NOM with Return reclassification.

(a) Value-weighted

	$K = 5$			$K = 10$			$K = 20$			$K = 50$			$K = 100$		
	H(20)	L(20)	H-L	H(10)	L(10)	H-L	H(5)	L(5)	H-L	H(2)	L(2)	H-L	H(1)	L(1)	H-L
Mean	0.169	0.003	0.166	0.235	-0.116	0.350	0.268	-0.174	0.442	0.469	-0.254	0.723	0.588	-0.398	0.986
Stddev	0.199	0.206	0.163	0.249	0.240	0.212	0.276	0.265	0.247	0.451	0.322	0.388	0.607	0.380	0.522
Sharpe	0.849	0.015	1.022	0.944	-0.483	1.656	0.971	-0.657	1.789	1.040	-0.789	1.864	0.969	-1.047	1.889
	H(10)	L(10)	H-L	H(10)	L(10)	H-L	H(10)	L(10)	H-L	H(10)	L(10)	H-L	H(10)	L(10)	H-L
Mean	0.216	-0.054	0.270	0.235	-0.116	0.350	0.251	-0.080	0.331	0.255	-0.069	0.324	0.275	-0.110	0.386
Stddev	0.219	0.227	0.191	0.249	0.240	0.212	0.248	0.241	0.216	0.261	0.230	0.207	0.283	0.238	0.232
Sharpe	0.986	-0.238	1.418	0.944	-0.483	1.656	1.012	-0.332	1.528	0.977	-0.300	1.566	0.972	-0.462	1.662

(b) Equal-weighted

	$K = 5$			$K = 10$			$K = 20$			$K = 50$			$K = 100$		
	H(20)	L(20)	H-L	H(10)	L(10)	H-L	H(5)	L(5)	H-L	H(2)	L(2)	H-L	H(1)	L(1)	H-L
Mean	0.274	-0.066	0.342	0.365	-0.150	0.515	0.484	-0.247	0.731	0.689	-0.331	1.020	0.960	-0.462	1.422
Stddev	0.229	0.221	0.138	0.274	0.244	0.181	0.339	0.266	0.238	0.452	0.318	0.359	0.600	0.332	0.496
Sharpe	1.197	-0.299	2.471	1.332	-0.615	2.842	1.428	-0.929	3.068	1.524	-1.041	2.840	1.600	-1.392	2.866
	H(10)	L(10)	H-L	H(10)	L(10)	H-L	H(10)	L(10)	H-L	H(10)	L(10)	H-L	H(10)	L(10)	H-L
Mean	0.323	-0.144	0.467	0.365	-0.150	0.515	0.393	-0.165	0.557	0.398	-0.161	0.559	0.409	-0.171	0.580
Stddev	0.258	0.237	0.175	0.274	0.244	0.181	0.285	0.241	0.186	0.295	0.246	0.192	0.306	0.242	0.195
Sharpe	1.252	-0.608	2.662	1.332	-0.615	2.842	1.379	-0.685	2.996	1.349	-0.654	2.906	1.337	-0.707	2.969

Table 14: Correlations between model variations

This table reports the correlations between the eight variations of the DM model: MOM-NOM, MOM-ORD, RET-NOM, RET-ORD, MOM-SZ-NOM, MOM-SZ-ORD, RET-SZ-NOM, RET-SZ-ORD. The correlations are calculated using the portfolio returns.

(a) Long

	MN	MO	RN	RO	MSN	MSO	RSN	RSO
MOM-NOM (MN)	1.000							
MOM-ORD (MO)	0.967	1.000						
RET-NOM (RN)	0.905	0.906	1.000					
RET-ORD (RO)	0.927	0.945	0.944	1.000				
MOM-SZ-NOM (MSN)	0.916	0.921	0.889	0.909	1.000			
MOM-SZ-ORD (MSO)	0.918	0.935	0.890	0.923	0.972	1.000		
RET-SZ-NOM (RSN)	0.890	0.890	0.902	0.918	0.926	0.923	1.000	
RET-SZ-ORD (RSO)	0.906	0.921	0.916	0.956	0.940	0.961	0.946	1.000

(b) Short

	MN	MO	RN	RO	MSN	MSO	RSN	RSO
MOM-NOM (MN)	1.000							
MOM-ORD (MO)	0.967	1.000						
RET-NOM (RN)	0.905	0.906	1.000					
RET-ORD (RO)	0.927	0.945	0.944	1.000				
MOM-SZ-NOM (MSN)	0.916	0.921	0.889	0.909	1.000			
MOM-SZ-ORD (MSO)	0.918	0.935	0.890	0.923	0.972	1.000		
RET-SZ-NOM (RSN)	0.890	0.890	0.902	0.918	0.926	0.923	1.000	
RET-SZ-ORD (RSO)	0.906	0.921	0.916	0.956	0.940	0.961	0.946	1.000

(c) Long-Short

	MN	MO	RN	RO	MSN	MSO	RSN	RSO
MOM-NOM (MN)	1.000							
MOM-ORD (MO)	0.967	1.000						
RET-NOM (RN)	0.905	0.906	1.000					
RET-ORD (RO)	0.927	0.945	0.944	1.000				
MOM-SZ-NOM (MSN)	0.916	0.921	0.889	0.909	1.000			
MOM-SZ-ORD (MSO)	0.918	0.935	0.890	0.923	0.972	1.000		
RET-SZ-NOM (RSN)	0.890	0.890	0.902	0.918	0.926	0.923	1.000	
RET-SZ-ORD (RSO)	0.906	0.921	0.916	0.956	0.940	0.961	0.946	1.000

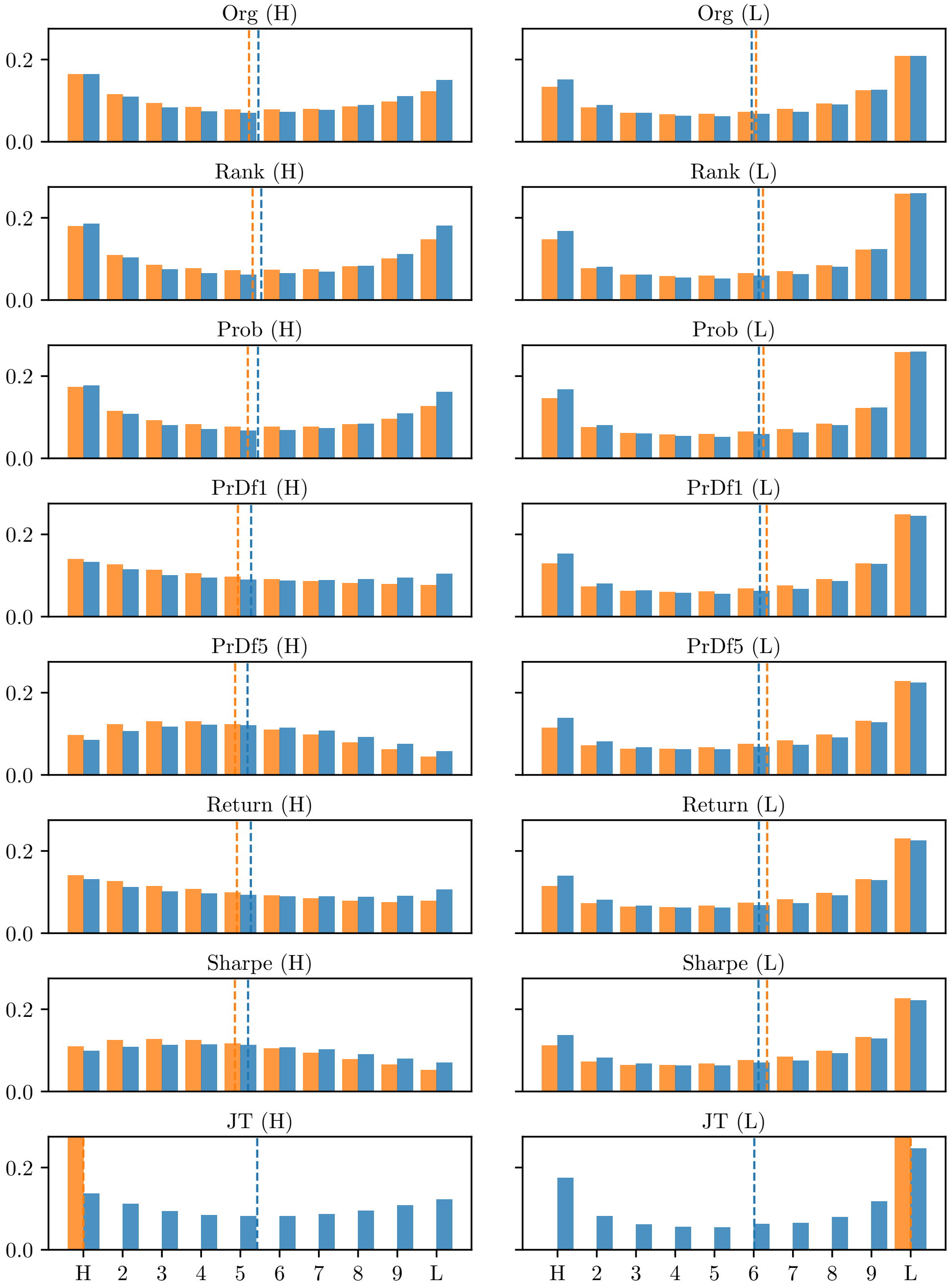


Figure 9: Cross-sectional relative return distribution after reclassification: MOM-NOM

This figure presents the cross-sectional relative return distributions in the high- and low-return deciles predicted by the momentum-based nominal classifier (MOM-NOM). The title of each graph denotes the reclassification method, and the horizontal axis denotes the one-month ahead return class. The orange bars represent the predicted probabilities and the blue bars represent the actual frequencies. The dotted lines represent the average return class. The sample period is from 1975.01 to 2017.01.

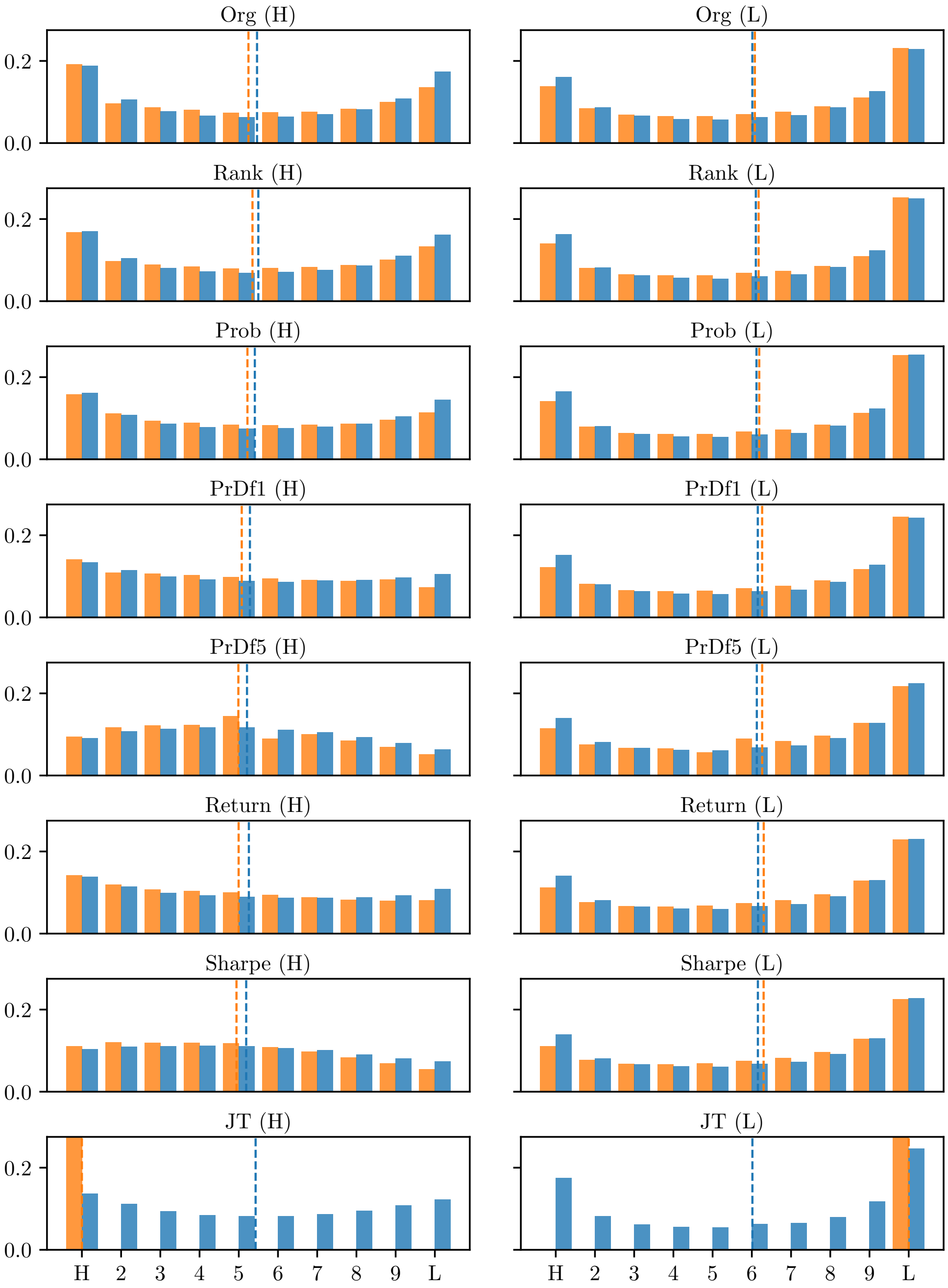
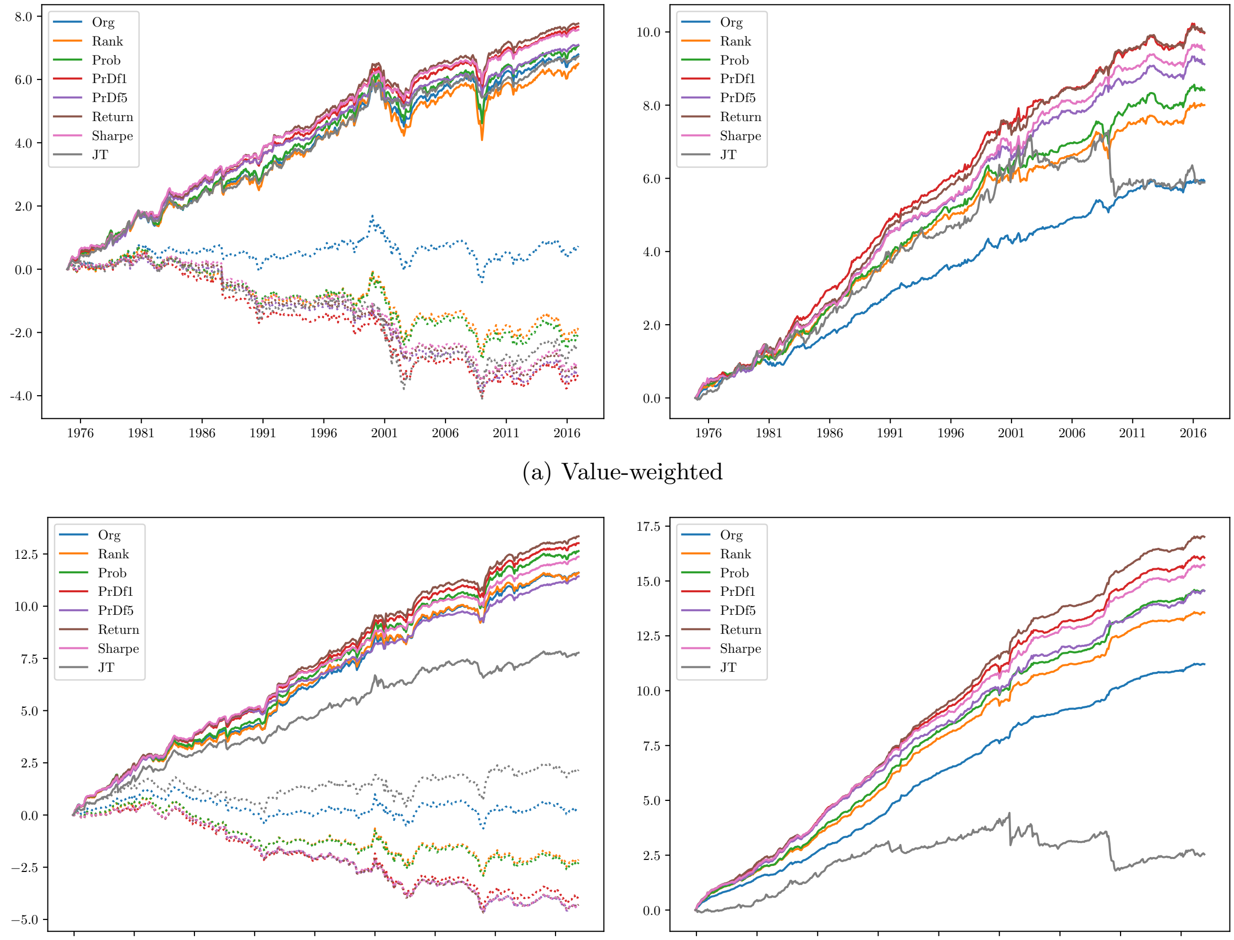


Figure 10: Cross-sectional relative return distribution after reclassification: MOM-ORD

This figure presents the cross-sectional relative return distributions in the high- and low-return deciles predicted by the momentum-based ordinal classifier (MOM-ORD). The title of each graph denotes the reclassification method, and the horizontal axis denotes the one-month ahead return class. The orange bars represent the predicted probabilities and the blue bars represent the actual frequencies. The dotted lines represent the average return class. The sample period is from 1975.01 to 2017.01.

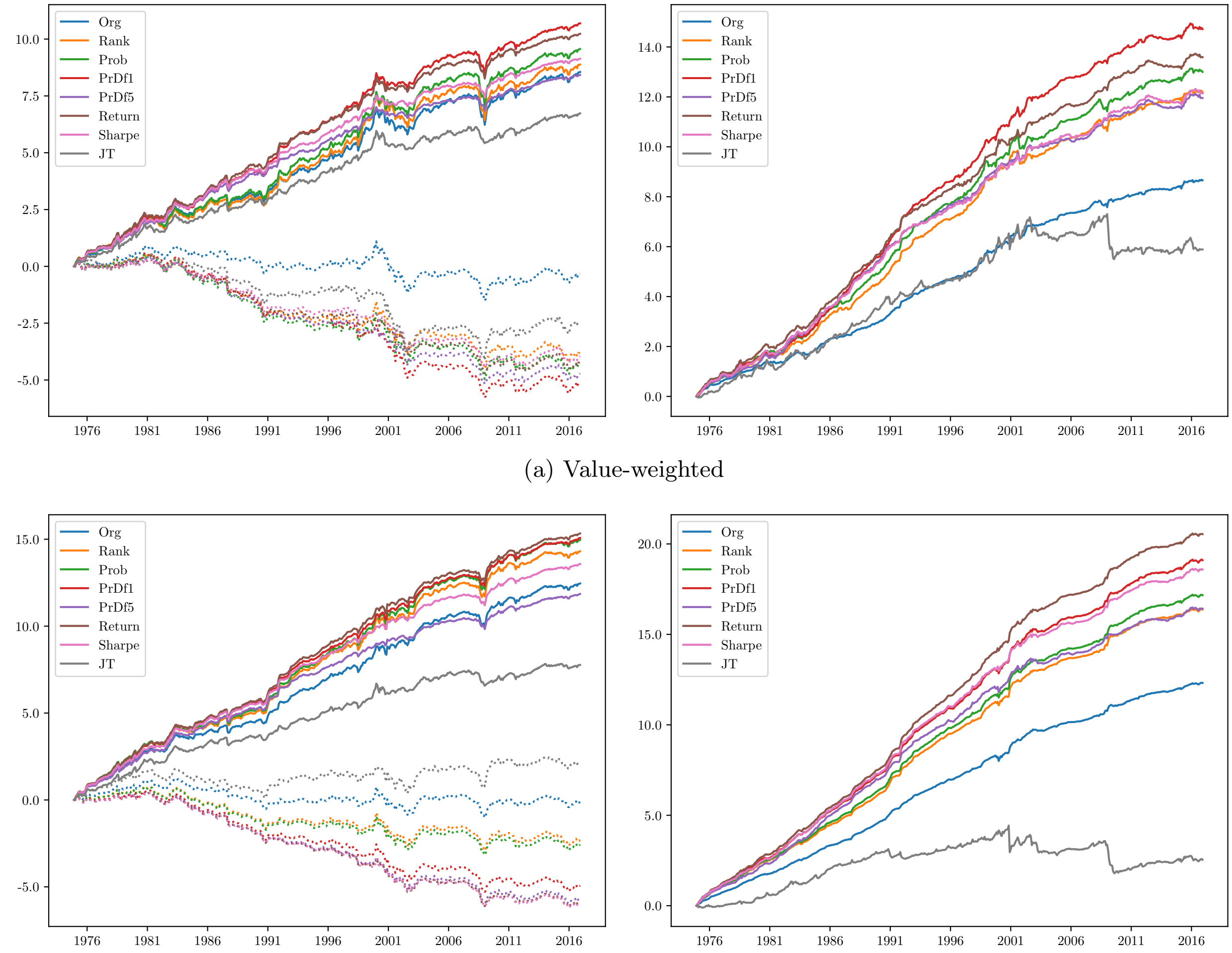


(a) Value-weighted

(b) Equal-weighted

Figure 11: Cumulative returns: MOM-NOM

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-NOM. The left panels present long (solid line) and short (dashed line) portfolios separately, and the right panels present the long-short portfolios.

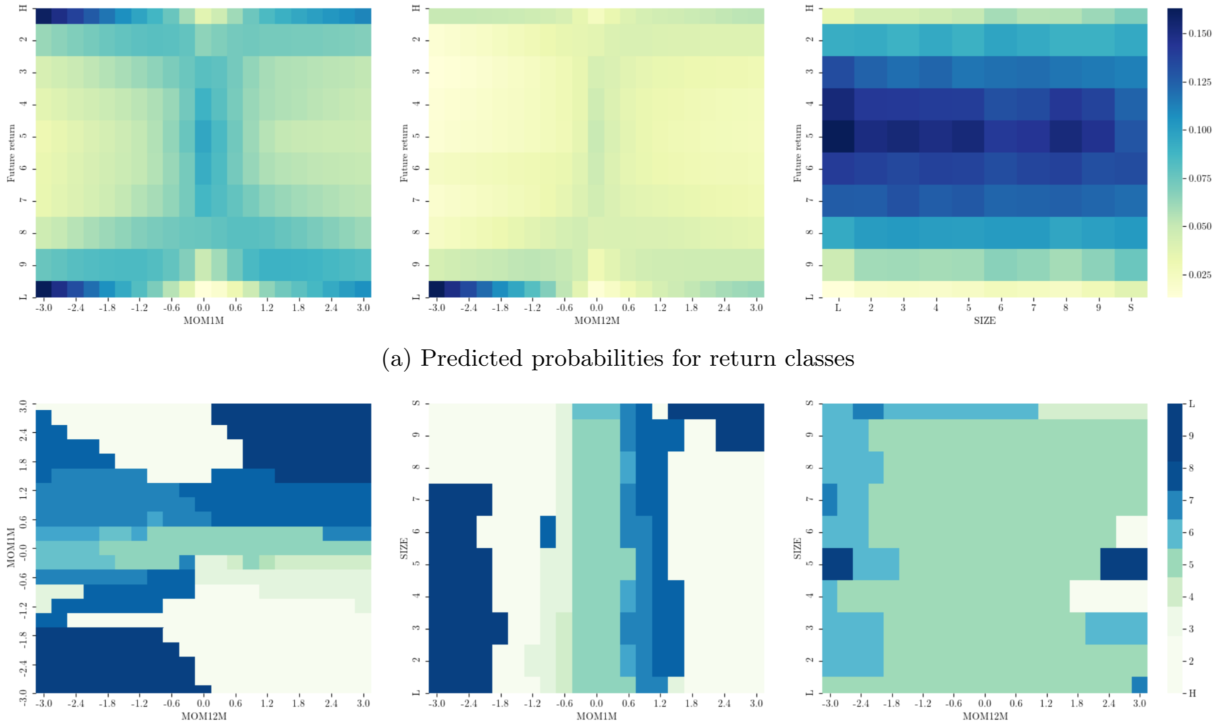


(a) Value-weighted

(b) Equal-weighted

Figure 12: Cumulative return: MOM-SZ-NOM

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-SZ-NOM. The left panels present long (solid line) and short (dashed line) portfolios separately, and the right panels present the long-short portfolios.

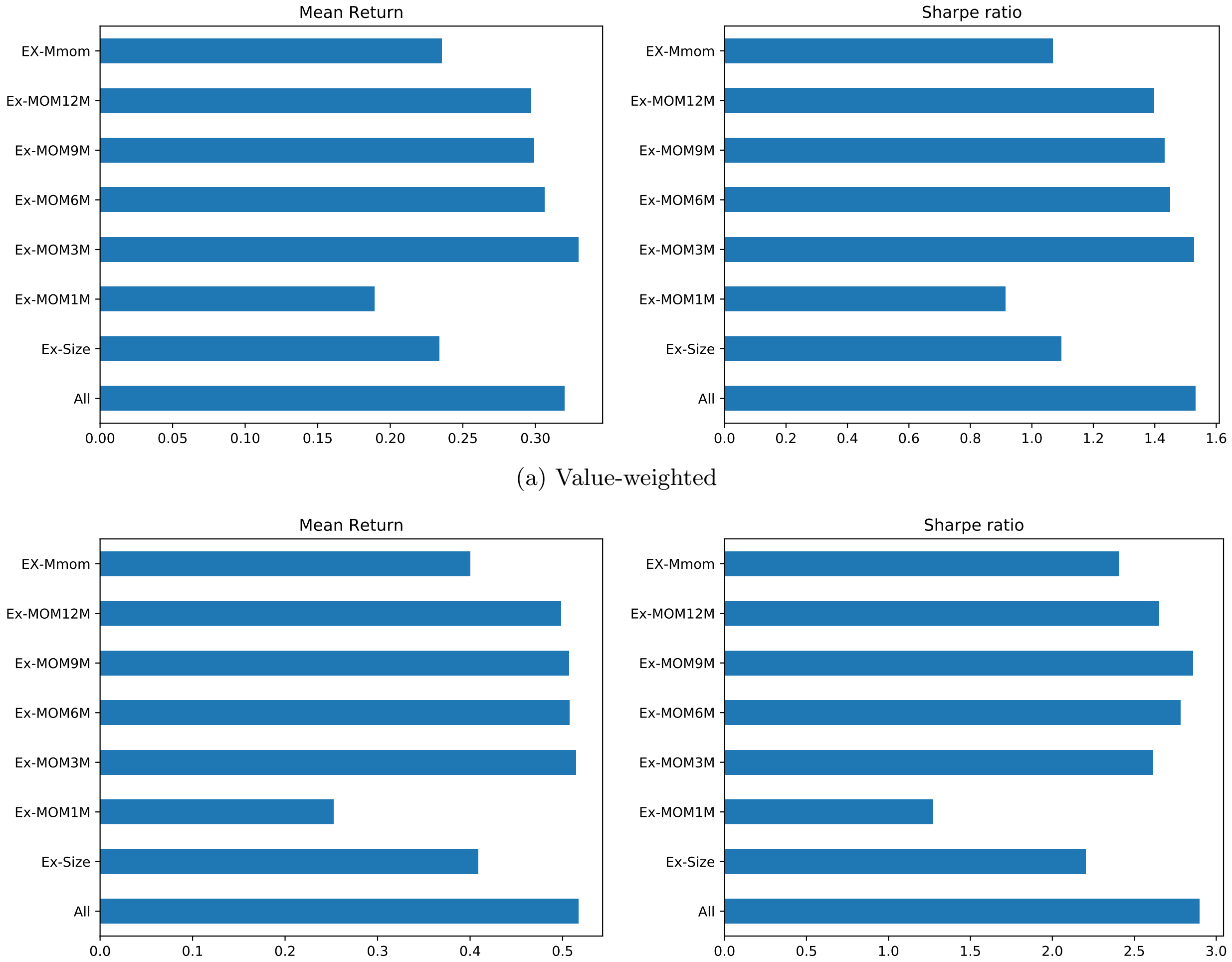


(a) Predicted probabilities for return classes

(b) Predicted return classes

Figure 13: Feature values vs. predicted return classes

This figure shows the sensitivity of the model with respect to the one-month momentum (MOM1M), one-year momentum (MOM12M), and size (SIZE) features. The top panel demonstrates the predicted probability for each return class given the value of a feature, and the bottom panel demonstrates the predicted class given the values of a pair of features. All features are standardized except for the size dummies, and the values of the other features are assumed to be 0.

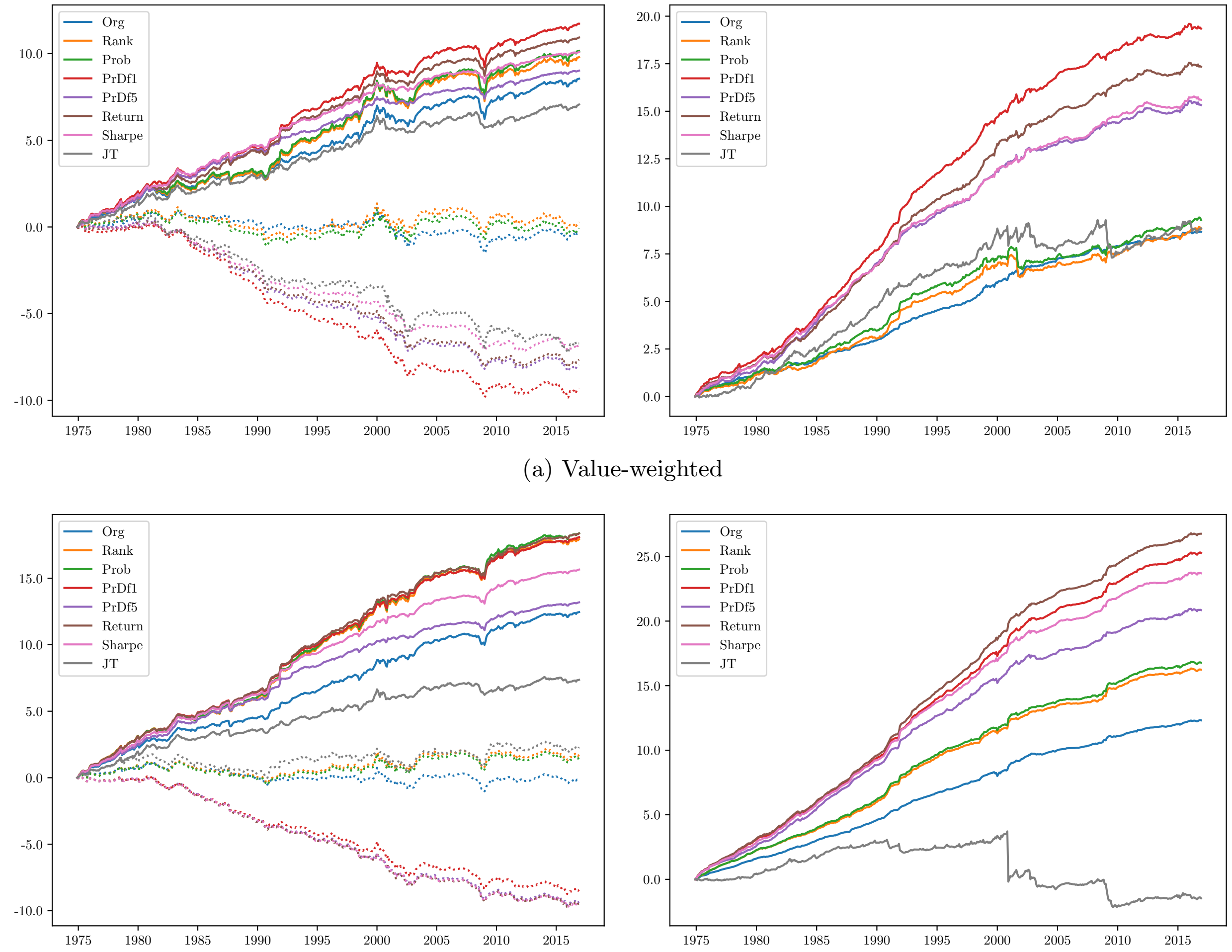


(a) Value-weighted

(b) Equal-weighted

Figure 14: Variable importance

This figure compares the performances of MOM-SZ-NOM when some of the input features are omitted. 'All' includes all the input features, 'Ex-Size' excludes the size dummies, 'Ex-MOMmM' excludes m-month momentum related features, $nMOM_m$ and M_{MOM_m}, and 'Ex-Mmom' excludes all cross-sectional means, M_{MOM_m}. A short bar implies that the associated variable is important. The model is retrained multiple times for each set of input features and the performance is averaged to minimize the impact of the randomness in learning.

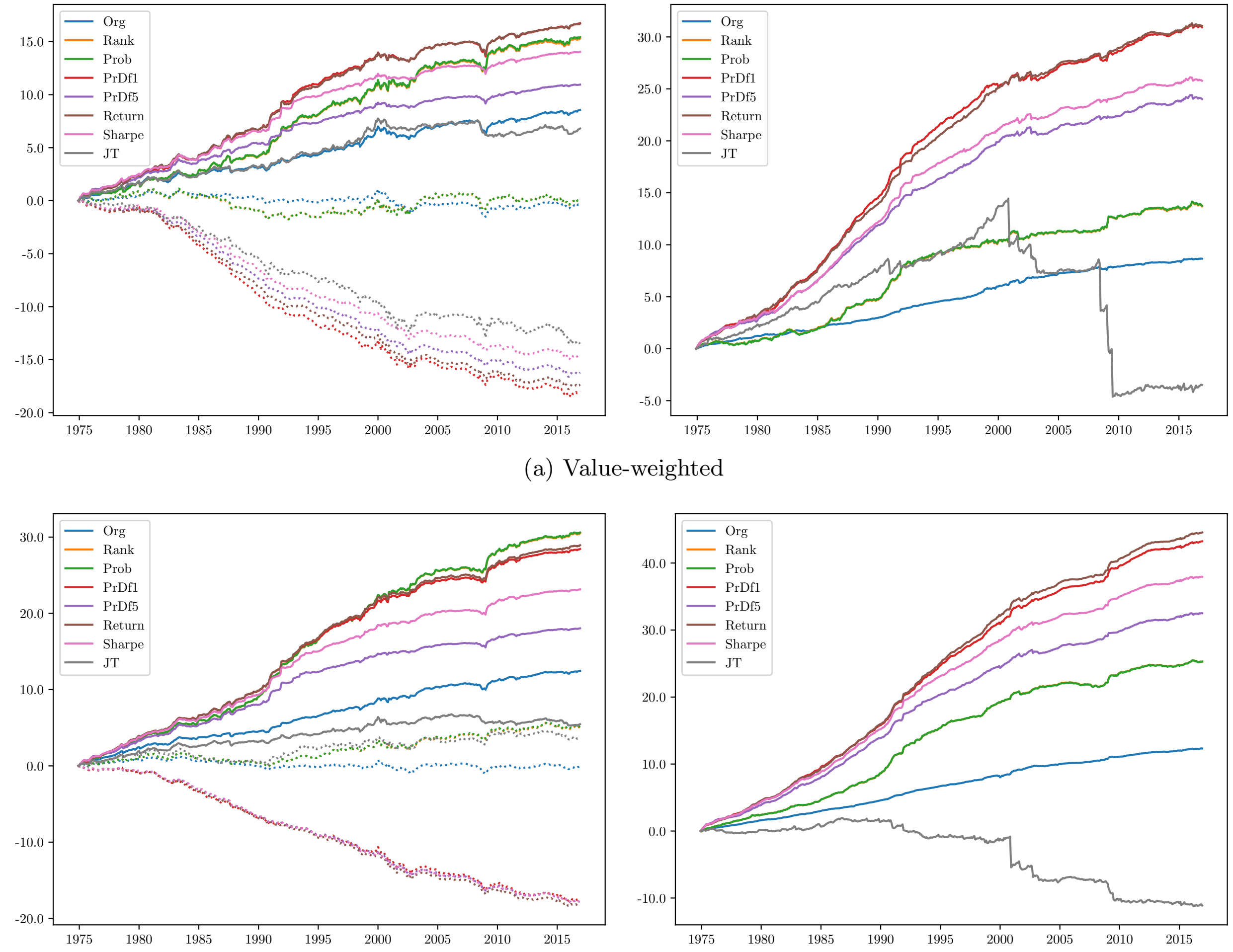


(a) Value-weighted

(b) Equal-weighted

Figure 15: Cumulative return: MOM-SZ-NOM (5%)

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-SZ-NOM. The portfolios are constructed using the top and bottom 5% of the stock universe.

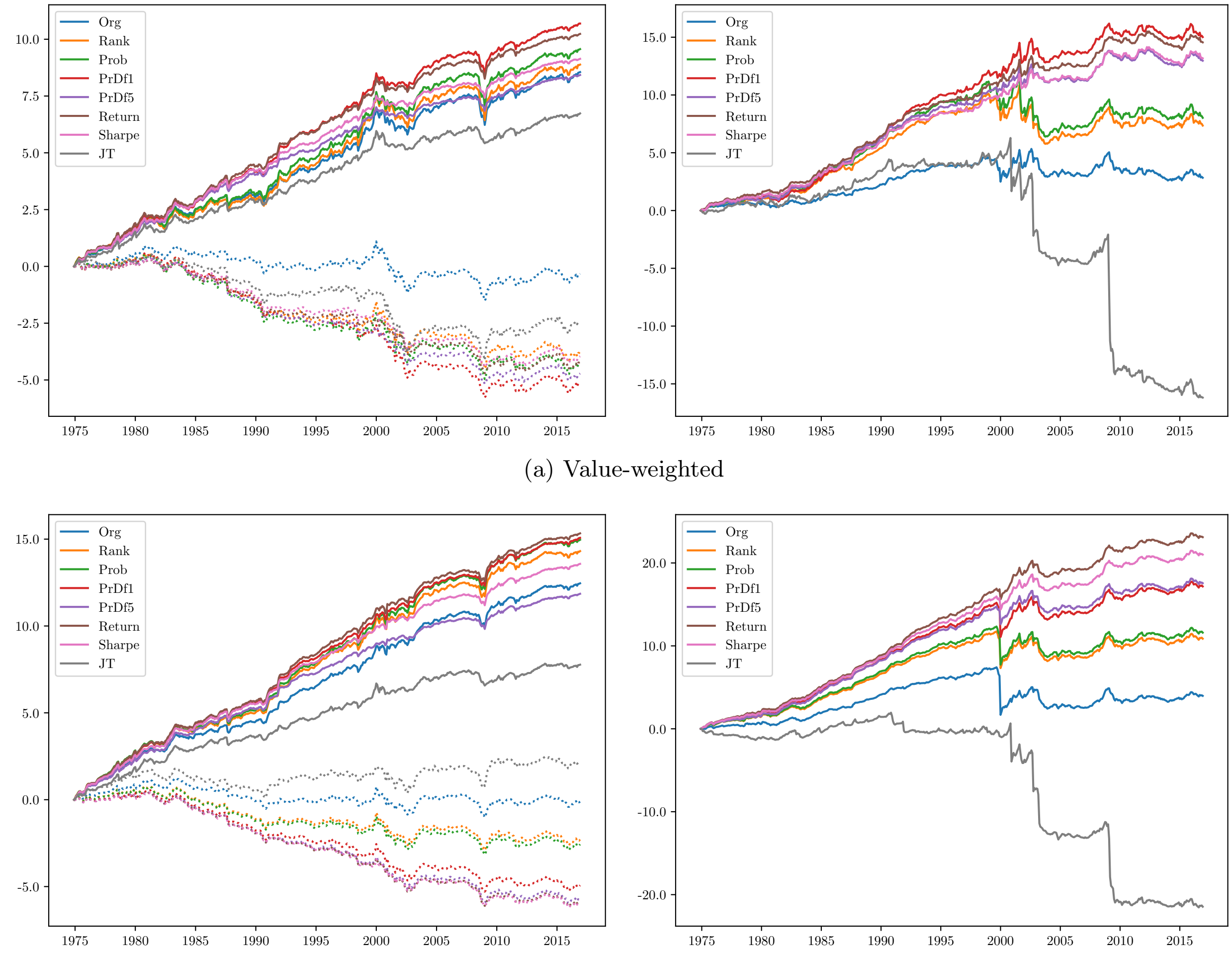


(a) Value-weighted

(b) Equal-weighted

Figure 16: Cumulative return: MOM-SZ-NOM (1%)

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-SZ-NOM. The portfolios are constructed using the top and bottom 1% of the stock universe.

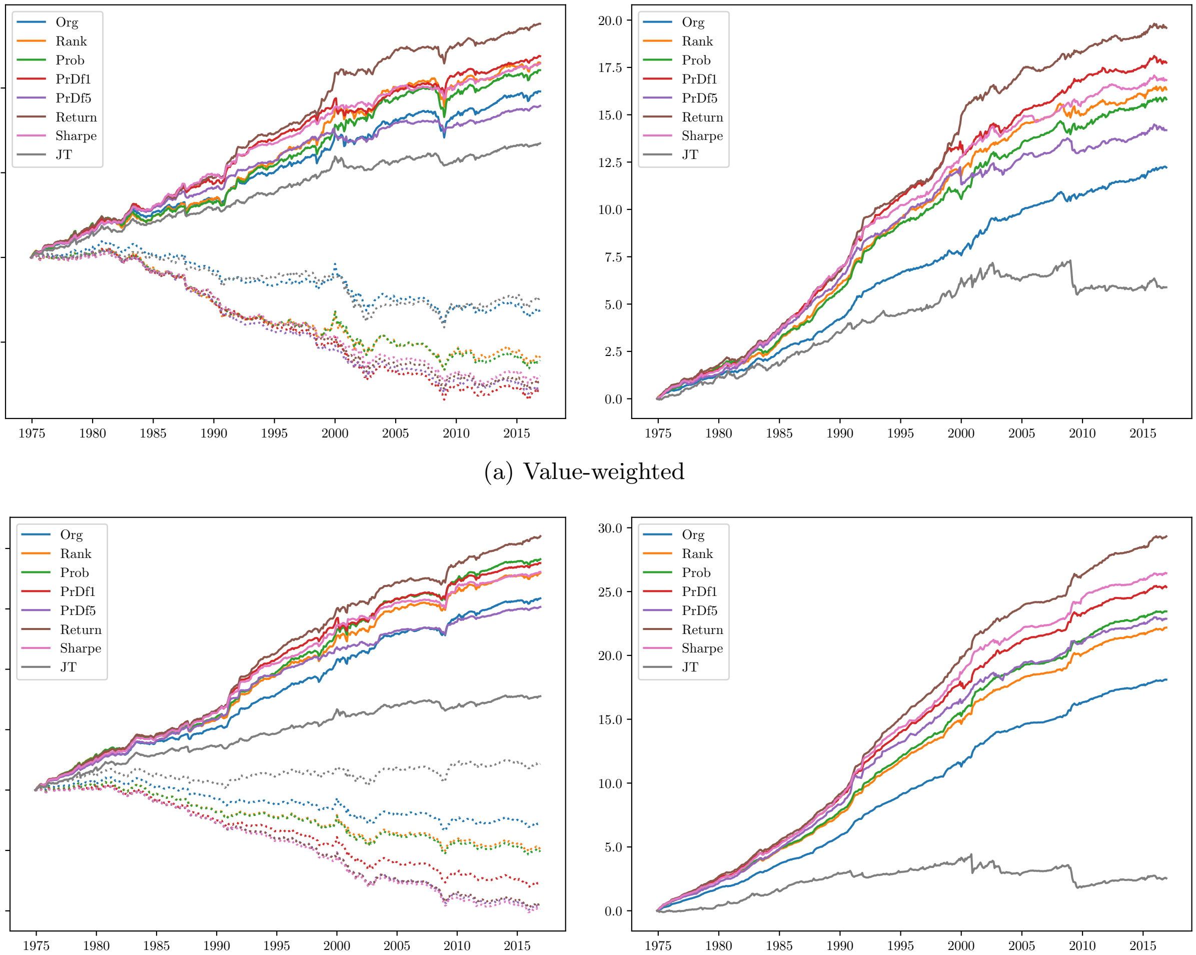


(a) Value-weighted

(b) Equal-weighted

Figure 17: Cumulative return: MOM-SZ-NOM (optimal long-short strategy)

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-SZ-NOM. The portfolios are constructed via the mean-variance optimal strategy described in Section A.2.

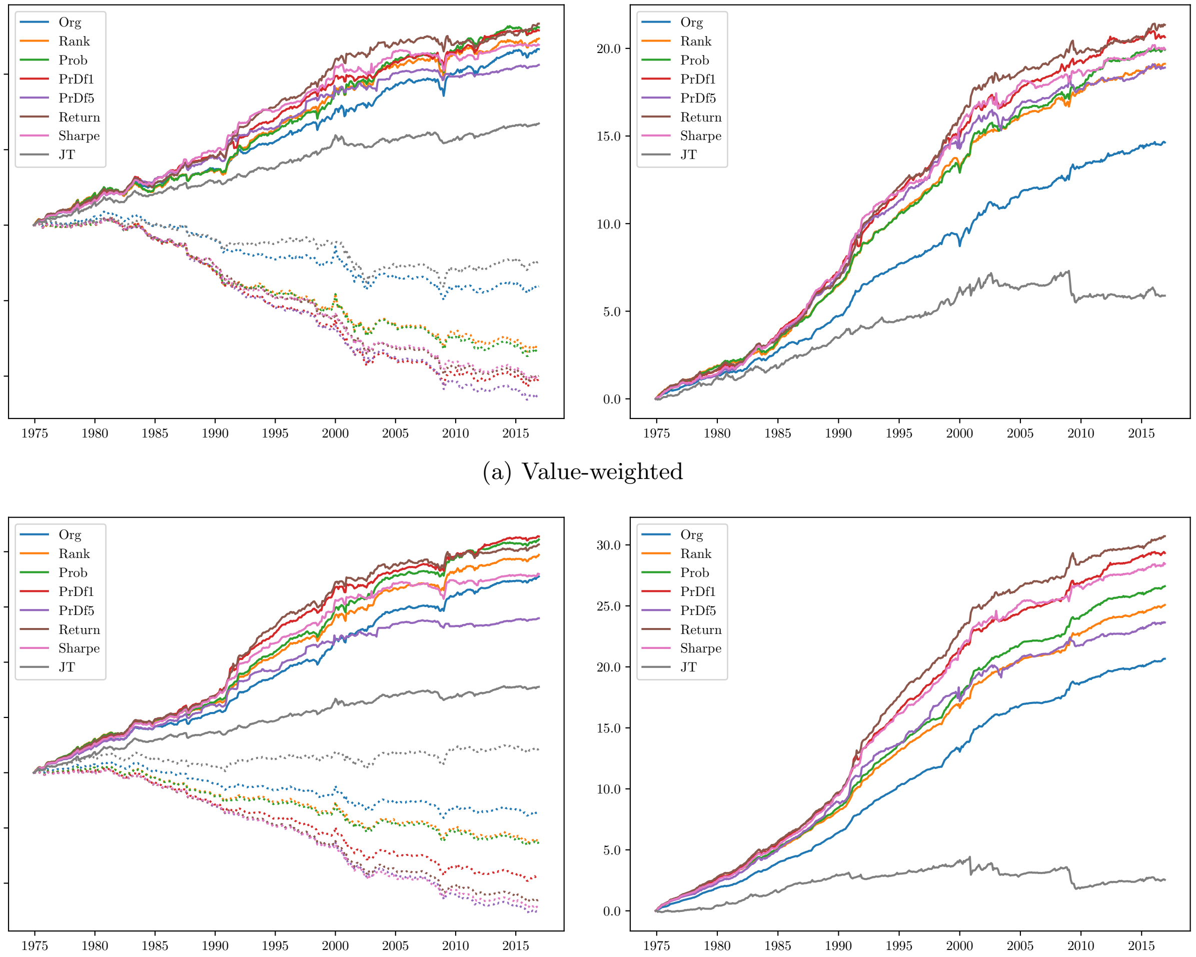


(a) Value-weighted

(b) Equal-weighted

Figure 18: Cumulative return: MOM-SZ-NOM (I10)

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-SZ-NOM. The portfolios are constructed using the stocks commonly included across ten random trials.



(a) Value-weighted

(b) Equal-weighted

Figure 19: Cumulative return: MOM-SZ-NOM (I50)

This figure compares the cumulative returns (logarithmic scale) of the long-short portfolios obtained from the reclassification criteria in Section 3.3 with the base model MOM-SZ-NOM. The portfolios are constructed using the stocks commonly included across fifty random trials.