1. 논문제목 (국문 혹은 영문)

(국문)

(영문)

Stock Return Predictability and Machine Learning: Incorporating Transaction Costs

2. 분야 : 재무관리

3. 저자명 및 소속(국문/영문) :

조용환 (한국주택금융공사, yhjo@hf.go.kr),
김용휘 (한국전자기술연구원, yonghwi@keti.re.kr),
신상욱 (울산과학기술원, simonshin@unist.ac.kr)

Stock Return Predictability and Machine Learning: Incorporating Transaction Costs

Yonghwan Jo* Yonghwi Kim† Sang-Ook (Simon) Shin‡

June 19, 2025

Abstract

We propose a machine learning model that incorporates transaction costs into return prediction using high-dimensional stock-level characteristics. Building on the conditional autoencoder of Gu, Kelly and Xiu (2021), our transaction cost-adjusted model penalizes costly-to-trade stocks during estimation. We show that this approach improves out-of-sample predictive power and portfolio performance, particularly when microcaps are excluded. In the nonmicrocap sample, the our model delivers a 40% higher Sharpe ratio and substantially larger alpha than the plain model, suggesting that accounting for transaction costs enhances signal quality and economic relevance.

Keywords: Stock Return, Transaction Cost, Machine Learning, Conditional Autoencoder, Neural Networks

***JEL* Classification**: G10, G11, G12, C45

---

*Research Fellow, Housing Finance Research Institute, Korea Housing-Finance Corporation. Email: yhjo@hf.go.kr

†Senior Researcher, IT Convergence System Research Center, Korea Electronics Technology Institute. Email: yonghwi@keti.re.kr

‡Assistant Professor of Finance, School of Business Administration, Ulsan National Institute of Science and Technology. Email: simonshin@unist.ac.kr

1 Introduction

Empirical studies in asset pricing have identified numerous characteristic-based anomalies that explain cross-sectional variations in stock returns. Collectively, these characteristics exhibit predictive power for stock returns, as demonstrated by Green, Hand and Zhang (2017) and Kelly, Pruitt and Su (2019). This predictability is further enhanced with the application of machine learning models. Specifically, portfolios constructed using machine learning signals, such as those developed by Gu, Kelly and Xiu (2020, 2021), tend to achieve higher Sharpe ratios than portfolios based on linear models.

However, these portfolios are often criticized for including many costly-to-trade stocks and having high turnover, both of which lead to substantial transaction costs. While they may achieve high ex post Sharpe ratios, their profitability can be eroded as trading costs outweigh the incremental gains in returns. More importantly, high transaction costs involved in these models raise the concern that their enhanced predictability may stem from capturing trading frictions rather than genuine relationships between characteristics and expected returns.[1]

Indeed, the key characteristics identified by Gu, Kelly and Xiu (2020) or Gu, Kelly and Xiu (2021) as most predictive of stock returns—such as 1-month momentum, size, and idiosyncratic volatility—are also closely linked to trading frictions, as classified by Hou, Xue and Zhang (2020), and often entail high transaction costs. Furthermore, Avramov, Cheng and Metzker (2023) show that the predictive power of machine learning models is largely concentrated in small stocks, with portfolio profitability declining sharply once those are excluded.

This concentration on small stocks is problematic. As of 2021, microcaps—defined as stocks below the 20^{th} percentile of NYSE market capitalization—accounted for less than 1% of market capitalization but nearly half of all stocks. Their prevalence biases machine learning models that use standard loss functions treating all stocks equally, regardless of economic

---

[1] Consistent with this argument, Hou, Xue and Zhang (2020) document that 96% of the anomalies in the trading friction category (including liquidity, market microstructure, etc.) fail to pass 95% significance test after controlling for microcaps, which are associated with higher transaction costs.

significance. As a result, the long-lived return predictability documented in these models tends to reflect microcap characteristics—stocks that are often illiquid, hard to short, and less relevant for practical investing—rather than the larger, more economically meaningful securities that dominate institutional portfolios. Thus, ignoring transaction costs when estimating these models not only undermines signal reliability but also limits their real-world applicability and their ability to identify meaningful risk–return tradeoffs.

To examine in detail how transaction costs influence estimation, consider the simple return-beta relationship:

$$r_{i,t} = \beta'_{i,t-1} f_t + \varepsilon_{i,t}, \tag{1}$$

where $r_{i,t}$ represents the return on stock i, and $\beta_{i,t-1}$ captures its exposure to factor f_t. Stocks with high $\beta_{i,t-1}$ are predicted to have higher returns, signaling a potential buy. However, if transaction costs are prohibitive, investors are less likely to purchase these stocks, leading to underpricing and subsequently higher returns than would occur in a low transaction cost environment. Conversely, stocks with low $\beta_{i,t-1}$ indicate a sell signal. In the presence of high transaction costs, investors are less inclined to short sell, resulting in overpricing and generally lower returns. If a machine learning model predominantly learns the return-beta relationship by minimizing mean squared errors from stocks with high transaction costs in-sample, to the extent this mispricing persists, the model's predictive power out-of-sample may be driven by such frictions.

To address this concern, we propose a novel machine learning model that incorporates transaction costs as a central component. Our approach involves developing an objective function that systematically accounts for transaction costs by penalizing stocks that are costly to trade. The model thus places greater emphasis on stocks with lower transaction costs and reduces focus on those with higher costs. Unlike prior studies that account for transaction costs only after portfolio formation (e.g., Novy-Marx and Velikov, 2016; Detzel, Novy-Marx and Velikov, 2023), our approach incorporates these frictions from the outset by embedding them directly into the model's optimization. This framework enables a more robust assessment of the predictive power of machine learning signals by incorporating transaction costs, aiming to enhance the model's effectiveness in linking characteristics with returns.

We employ the conditional autoencoder (CA) model proposed by Gu, Kelly and Xiu (2021) to predict expected returns for individual U.S. stocks. This model is highly flexible, effectively capturing the nonlinear relationship between stock characteristics and factor exposures (i.e., $\beta_{i,t}$), while constructing factors linearly from individual returns. It has been shown to outperform both linear models, such as Instrumented Principal Component Analysis (IPCA) by Kelly, Pruitt and Su (2019), and nonlinear models, such as the deep neural network (DNN) approach in Gu, Kelly and Xiu (2020). Additionally, Avramov, Cheng and Metzker (2023) find that the CA model exhibits low time-series variation and performs particularly well in low market states.

Building on this framework, we introduce a penalty term that forces the estimation of $\beta_{i,t}$ to approach zero for stocks with high transaction costs. If the penalty is overly stringent, it effectively excludes high-transaction-cost stocks from the learning process, forfeiting the opportunity to capture the pricing relationship between returns and betas for these stocks. Conversely, if the penalty is too lenient, the estimation closely resembles that of Gu, Kelly and Xiu (2021). In this way, our approach generalizes machine learning models by explicitly incorporating transaction costs, offering a flexible framework that balances penalization with the retention of valuable information from high-cost stocks.

Using a sample of U.S stocks listed on the NYSE, AMEX, and NASDAQ, we examine the performance of a transaction cost-adjusted conditional autoencoder (TC model), which modifies the original CA model of Gu, Kelly and Xiu (2021) by incorporating transaction costs into the optimization. We evaluate how well 94 stock-level characteristics—originally used in Gu, Kelly and Xiu (2020)—predict stock returns under this modified framework, and compare its performance to the original CA model (hereafter, plain model). Our empirical tests show that the TC model consistently delivers higher out-of-sample predictability.

Specifically, the out-of-sample predictive R^2 is 0.44% in the TC model, compared to 0.27% in the plain model. The improvement becomes more pronounced when microcap stocks are excluded: the TC model achieves an R^2 of 0.27%, while the plain model yields only 0.03%.

This enhanced return–beta relationship translates into improved ex post portfolio performance. We sort stocks based on expected returns and construct zero-cost portfolios by taking long positions in the top decile and short positions in the bottom decile. In the full sample, the TC model delivers a modest improvement in the Sharpe ratio relative to the plain model (1.512 vs. 1.442). Consistent with the R2 results, the performance gap widens significantly in the nonmicrocap sample, where the TC model achieves a Sharpe ratio of 1.071 compared to 0.758 for the plain model—representing a more than 40% improvement. To assess whether the superior performance of the TC model is due to greater exposure to systematic risk, we compare abnormal returns relative to the Fama-French six-factor model. In the full sample, the TC model yields a slightly higher monthly alpha (2.86%) than the plain model (2.83%). However, in the nonmicrocap sample, the difference becomes substantial: the TC model produces an alpha of 1.10%, compared to 0.66% for the plain model. These findings suggest that the improved performance is driven by better return predictability, not higher risk exposure.

While overall portfolio performance is higher in the full sample—even after accounting for transaction costs—the TC model performs particularly well in the nonmicrocap sample. Microcap stocks are notoriously difficult to arbitrage due to high trading frictions. As shown in Avramov, Cheng and Metzker (2023), the predictive power of machine learning models diminishes substantially when microcaps are excluded. However, by incorporating "economic restrictions" such as transaction costs into the model, we effectively restore a significant portion of the performance drop.

Next, we examine whether the observed performance gains stem from improvements at the characteristic level. Specifically, we seek to identify which characteristics are most important in explaining the cross-section of stock returns. To this end, we employ two complementary approaches: absolute partial derivatives and the reduction in out-of-sample performance (i.e., risk-adjusted returns). Compared to the plain model, the predictability of the TC model is less reliant on a small subset of characteristics, most of which are related to trading frictions. Specifically, the average importance of the top five characteristics decreases from 0.75 in the plain model to 0.50 in the TC model using the full sample. Focusing on out-of-sample performance measured by FF6 alpha, the average percentage reduction of the top five characteristics declines from 24% in the plain model to 20% in the TC model. A similar pattern is observed in the nonmicrocap sample.

Unlike Gu, Kelly and Xiu (2020), which focus on reductions in in-sample R^2, we document that a substantial number of characteristics exhibit negative contributions to out-of-sample performance. However, when transaction cost constraints are incorporated into the optimization process, the number of such negatively contributing characteristics declines significantly—from 36 to 28 in the full sample, and from 37 to 9 in the non-microcap sample. Negative contributions may arise for several reasons, including spurious relationships, high covariance among characteristics, and overfitting to noise. Our attempt to address the issue of high transaction costs appears effective in mitigating these problems. As a result, the TC model enhances the quality of signal processing, leading to improved out-of-sample performance.

Related Literature

Traditionally, empirical asset pricing studies that aim to explain cross-sectional differences in asset returns begin with a small set of characteristics motivated by theory or economic reasoning. The significance of these characteristics is typically tested using characteristic-sorted portfolios or cross-sectional regressions (e.g., the Fama–MacBeth method).

Recently, a growing literature applies machine learning to empirical asset pricing. Gu, Kelly and Xiu (2020) implement various machine learning methods on a comprehensive dataset that includes all traded stocks. Building on this, Gu, Kelly and Xiu (2021) propose a conditional autoencoder model, while Chen, Pelger and Zhu (2024) incorporate no-arbitrage constraints into neural network architectures such as feedforward networks, recurrent long short-term memory (LSTM) networks, and generative adversarial networks (GANs). Cong, Tang, Wang and Zhang (2021) adopt reinforcement learning algorithms to directly optimize portfolios. Jiang, Kelly and Xiu (2023) introduce a novel approach that uses image-based features extracted from stock price charts to identify predictive patterns.

These machine learning methods adopt a data-driven approach to uncover nonlinear relationships directly from the data. However, this strength also makes them sensitive to the composition of the input sample. Despite their strong performance in return prediction, machine learning models often overfit to stocks that dominate the sample size but are difficult to trade (i.e., microcaps) undermining the implementability of the resulting strategies.

One potential solution is to exclude microcaps from the training sample. However, this introduces both theoretical and practical concerns. Theoretically, if nonmicrocaps do not span the payoff space required to construct the pricing kernel, training the model on this subsample alone can lead to biased estimates of expected returns. Empirically, given that microcaps constitute a large portion of the cross-section by count, excluding them substantially reduces the number of observations relative to the number of parameters. As a result, the characteristic space becomes sparse and training efficiency deteriorates.[2] In contrast, we incorporate transaction costs directly into the machine learning objective function while retaining the entire universe of stocks during training. This approach reweights the optimization process, emphasizing more economically significant and implementable securities.[3]

Historically, empirical asset pricing studies have largely ignored transaction costs, assuming frictionless markets. This approach emphasizes statistical significance over economic relevance. However, transaction costs can substantially reduce portfolio profitability and

---

[2] In Table 4, the Sharpe ratio of portfolios trained only on nonmicrocaps using the plain model performs no better than portfolios of nonmicrocaps from our transaction cost-adjusted model.

[3] Avramov, Cheng and Metzker (2023) employ a value-weighted loss function to address the small-stock bias, yet report similar performance to equal-weighted alternatives.

challenge the interpretation of return predictability.[4] Some studies attempt to mitigate this issue by forming value-weighted portfolios and/or sorting by NYSE breakpoints. Yet, as shown by Novy-Marx and Velikov (2016), many anomalies—particularly those involving high turnover or difficult-to-trade stocks—become statistically insignificant after adjusting for transaction costs.

While most of these studies adjust for costs after constructing signals to sort assets, others explicitly incorporate frictions during optimization. For instance, Gârleanu and Pedersen (2013) and Jensen, Kelly, Malamud and Pedersen (2024) account for transaction costs when constructing optimal portfolios. In a similar spirit, we optimize the model by embedding transaction costs into the loss function. However, our objective is not just to improve portfolio choice, but to correct the biases in return predictability caused by these frictions. We directly penalize costly-to-trade stocks during training, so that return predictability is based on net, rather than gross, returns.

Beyond improving overall predictability, it is equally important to understand which characteristics drive the performance. Identifying key predictors helps distinguish between anomalies that reflect genuine risk–return relationships and those that merely capture short-lived mispricings that persist because they are too costly to be arbitraged away. Green, Hand and Zhang (2017) consider 94 characteristics and find that only a small subset offers independent explanatory power. Gu, Kelly and Xiu (2020, 2021) rank the importance of characteristics by assessing the reduction in total in-sample R^2 when each characteristic is excluded.

We extend their approach by measuring characteristic importance based on the out-of-sample reduction in portfolio performance (e.g., FF6 alpha, Sharpe ratio, or predictive R^2). This method enables us to quantify the magnitude of each characteristic's impact on out-of-sample performance, including those with negative effects. To the extent that

---

[4] For example, Detzel, Novy-Marx and Velikov (2023) demonstrate that asset pricing models that ignore transaction costs fail to span the achievable mean-variance efficient frontier. They provide empirical evidence that high transaction costs create a substantial gap between the gross and net maximum squared Sharpe ratios, particularly for factor models with high turnover.

in-sample importance does not translate to out-of-sample settings, our results may diverge from theirs. While we confirm that characteristics identified as important by Gu, Kelly and Xiu (2020, 2021) remain influential in our model, their dominance is less pronounced, and fewer characteristics exhibit negative predictive power. By incorporating transaction costs into model estimation, we reduce biases in return prediction and improve both out-of-sample performance and the economic relevance of machine learning signals.

The rest of the paper is organized as follows. Section 2 outlines the methodology for estimating the transaction cost-adjusted model. Section 3 presents the empirical results. Section 4 concludes.

2 Methodology

In this section, we first outline the approach used to measure transaction costs. Next, we derive the objective function that incorporates transaction costs. Lastly, we provide an overview of the conditional autoencoder model employed in our study and then the estimation procedure.

2.1 Measuring transaction costs

To measure transaction costs, we follow an approach similar to Brandt, Santa-Clara and Valkanov (2009), Hand and Green (2011), and DeMiguel, Martín-Utrera, Nogales and Uppal (2020). Specifically, to capture cross-sectional variation, transaction costs $c_{i,t}$ are modeled as inversely proportional to stock value, while their time-series variation is captured by a gradual decline over time:

$$c_{i,t} = y_t \max\left(\zeta_{i,t-1}, 0.04\%\right), \tag{2}$$

where y_t captures the time-series variation, decreasing linearly from 5.82 in 1957 to 1 in 2002 and remaining constant thereafter.[5] The cross-sectional component is given by $\zeta_{i,t-1} = 0.2253\% - 0.2712\% \times v_{i,t-1}$, where $v_{i,t-1}$ is the normalized log market value of stocks, scaled such that its minimum is zero and its maximum is one.[6] The coefficients for $\zeta_{i,t-1}$ are chosen to yield a median transaction cost of approximately 10 basis points (bps),[7] with the 25^{th} and 75^{th} percentiles at 6 bps and 14 bps, respectively. Since this formulation allows transaction costs to be too low or even negative, a minimum transaction cost of 4 bps is imposed. As a result, in 2002 or later, the smallest stocks have transaction costs of 23 bps, while the largest stocks have costs of 4 bps.

Figure **??** illustrates the distribution of monthly transaction costs from 1957 to 2021. The solid line represents the median, with other percentiles also displayed. Notably, from 2002 onward, the 5^{th} percentile is bounded at a minimum of 4 bps. Our specification effectively captures the downward trend in transaction costs over time while maintaining sufficient cross-sectional variation.

2.2 Objective function adjusted for transaction costs

Following Kelly, Pruitt and Su (2019) and Gu, Kelly and Xiu (2021), the excess returns on stocks can be represented using a K-factor structure:

$$r_{i,t} = \beta'_{i,t-1} f_t + \varepsilon_{i,t}, \tag{3}$$

where f_t is the $K \times 1$ vector of factors, and $\beta_{i,t-1}$ is the $K \times 1$ vector of factor loadings. When transaction costs are incorporated, the net return realized by investors is adjusted

---

[5] This ensures that transaction costs in 1974 are four times higher than in 2002, consistent with Brandt, Santa-Clara and Valkanov (2009).

[6] Brandt, Santa-Clara and Valkanov (2009) assume $\zeta_{i,t-1} = 0.60\% - 0.25\% \times v_{i,t-1}$. We modify their model to better align the distribution with Frazzini, Israel and Moskowitz (2018), who estimate transaction costs based on live trade execution data.

[7] According to Frazzini, Israel and Moskowitz (2018), the average one-way transaction cost for U.S. stocks is approximately 10 bps.

accordingly: it decreases by $c_{i,t}$ for long positions and increases by $c_{i,t}$ for short positions. Thus, the net return is given by $r_{i,t} - \tilde{c}_{i,t}$, where $\tilde{c}_{i,t}$ represents the signed transaction cost, defined as:

$$\tilde{c}_{i,t} = \begin{cases} c_{i,t}, & \text{if stock } i \text{ is held in a long position,} \\ -c_{i,t}, & \text{if stock } i \text{ is held in a short position.} \end{cases} \tag{4}$$

Transaction costs systematically introduce a bias in the returns described in Equation (3), causing them to deviate from the components attributed to stock characteristics (i.e., $\beta'_{i,t-1} f_t$). In particular, even when stocks exhibit higher expected returns, investors may postpone purchasing until the price declines sufficiently to compensate these costs in the presence of significant transaction costs. Similarly, high transaction costs deter investors from short selling stocks with lower expected returns. As a result, stocks with $\beta'_{i,t-1} f_t > 0$ (< 0) tend to be underpriced (overpriced), leading to higher (lower) realized returns.

To address this issue, we extend Equation (3) to account for transaction costs as follows:

$$r_{i,t} = \beta'_{i,t-1} f_t + \tilde{c}_{i,t} + \varepsilon_{i,t}. \tag{5}$$

By using net returns, which incorporate transaction costs, this formulation accurately reflects investors' real-world economic outcomes and thus provides a more precise representation of the risk-reward trade-off. Accordingly, our goal is to adjust the estimation framework to properly incorporate these effects into the optimization process. At time t, in a reference case where $\tilde{c}_{i,t} = 0$ for all stock i, a factor-mimicking portfolio f_t and its exposure $\beta_{i,t-1}$ can be estimated by minimizing the mean squared errors (MSEs), following Kelly, Pruitt and Su (2019) and Gu, Kelly and Xiu (2021):

$$\frac{1}{N} \sum_{n=1}^{N} \left(r_{i,t} - \tilde{r}_{i,t} \right)^2, \tag{6}$$

where $\tilde{r}_{i,t}$ denotes the expected return, given by $\tilde{r}_{i,t} \equiv \beta'_{i,t-1} f_t$. When transaction costs are present, the objective function is modified as follows:

$$\frac{1}{N}\sum_{n=1}^{N}\left(r_{i,t}-\tilde{r}_{i,t}-\tilde{c}_{i,t}\right)^2. \tag{7}$$

Equation (7) can be rewritten as $\frac{1}{N}\sum_{n=1}^{N}\left(r_{i,t}-\tilde{r}_{i,t}\right)^2-\frac{2}{N}\sum_{n=1}^{N}\tilde{c}_{i,t}\left(r_{i,t}-\tilde{r}_{i,t}\right)+\frac{1}{N}\sum_{n=1}^{N}\tilde{c}_{i,t}^2$, using Equation (6). By suppressing terms unrelated to cross-sectional differences, the objective function simplifies to:

$$\frac{1}{N}\sum_{n=1}^{N}\left(r_{i,t}-\tilde{r}_{i,t}\right)^2+\frac{2}{N}\sum_{n=1}^{N}\tilde{c}_{i,t}\left(\beta_{i,t-1}'f_t\right), \tag{8}$$

where the second term is positive, reflecting the positive covariance between signed transaction costs ($\tilde{c}_{i,t}$) and expected returns ($\beta_{i,t-1}'f_t$), as discussed earlier.

When transaction costs are high, the absolute value of $c_{i,t}$ is large, but its sign remains uncertain since stocks have not yet been classified as long or short during estimation. To address this, as outlined in Section 2.1, we model transaction costs as inversely proportional to stock value: $c_{i,t}=g(v_{i,t-1})$, where $g(\cdot)$ is a monotonically decreasing function that ensures positivity, as defined in Equation (2). To maintain the positivity of the second term in Equation (8), we replace $\tilde{c}_{i,t}$ with $g(v_{i,t-1})$ and take the absolute values of the elements in $\beta_{i,t-1}$. Additionally, we simplify the vector of factors f_t to $\mathbf{1}_{K\times 1}$, where $\mathbf{1}_{K\times 1}$ represents a $K\times 1$ vector of ones.[8] Consequently, in our empirical analysis, we modify the second term in Equation (8) as $\frac{2}{N}\sum_{i=1}^{N}g(v_{i,t-1})\left|\beta_{i,t-1}'\right|\mathbf{1}_{K\times 1}$.

We then aggregate the objective function at time t over T periods, resulting in the following objective function:

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N}\left(r_{i,t}-\tilde{r}_{i,t}\right)^2+\frac{2}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N}g(v_{i,t-1})\left|\beta_{i,t-1}'\right|\mathbf{1}_{K\times 1}, \tag{9}$$

---

[8] This implies that each factor imposes an equal penalty on beta estimation. Another key reason for this simplification is that, while all elements of f_t should theoretically be positive—indicating a positive risk premium—they may empirically take negative values. In this case, the cost function decreases, potentially exacerbating the bias introduced by transaction costs.

The first term in Equation (9) corresponds to the standard objective function for minimizing MSEs. In contrast, the second term serves as a penalty on the estimation of $\beta_{i,t-1}$, driving it closer to zero for stocks with higher transaction costs. This adjustment ensures that the model places greater emphasis on stocks with lower transaction costs while giving less weight to those with higher transaction costs. Ignoring the second term would lead to a biased estimation of $\tilde{r}_{i,t}$, particularly in the presence of substantial cross-sectional variation in transaction costs, as proxied by $g(v_{i,t-1})$.

2.3 Conditional autoencoder model

We adopt the conditional beta pricing model proposed by Gu, Kelly and Xiu (2021) as our benchmark model. This model extends dimension reduction techniques such as IPCA by recursively identifying a latent factor structure from stock returns. The beta pricing model is formulated as a K-factor structure:

$$r_{i,t} = \beta\left(z_{i,t-1}\right)' f_t + \varepsilon_{i,t}, \tag{10}$$

where $\beta\left(z_{i,t-1}\right)$ represents the conditional factor exposure based on firm characteristics $z_{i,t-1}$, and f_t is the K-dimensional latent factors.

Building on this K-factor structure, we implement a conditional autoencoder following Gu, Kelly and Xiu (2021). This approach utilizes two feed-forward neural networks, where the output layer reconstructs stock returns $r_{i,t}$ as the dot product of $\beta\left(z_{i,t-1}\right)$ and f_t. Within the autoencoder framework, the two neural networks serve as non-linear estimators: one for the conditional beta exposures and the other for the K-dimensional latent factors.

For implementation, we use a 5-factor model with two hidden layers for the $\beta\left(z_{i,t-1}\right)$ estimator. To ensure that our model replicates not only the structure but also the performance of Gu, Kelly and Xiu (2021), we match the hypothesis space by aligning hyperparameters, activation functions, training scheme, regularization techniques, and optimization algorithms. Specifically, we employ the ReLU activation function with batch normalization for the $\beta\left(z_{i,t-1}\right)$ estimator. During training, we incorporate three regularization techniques: early stopping, a penalty function, and an ensemble approach.

We incorporate a regularization technique to mitigate overfitting and introduce a parameter to control the penalization of transaction costs. This leads to the following regularized objective function:

$$\mathcal{L}\left(\theta;\cdot\right)=\frac{1}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N}\left(r_{i,t}-\tilde{r}_{i,t}\right)^{2}+\frac{\lambda}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N}g(v_{i,t-1})\left|\beta_{i,t-1}'\right|\mathbf{1}_{K\times 1}+\phi\left(\theta;\cdot\right), \tag{11}$$

where $\lambda > 0$ is a tuning parameter, θ denotes the weight parameters of the neural networks, and $\phi\left(\theta;\cdot\right)$ corresponds the LASSO penalization following Gu, Kelly and Xiu (2021). Since transaction costs are estimated rather than directly observed, the degree to which beta estimation is penalized for stocks with high transaction costs remains uncertain. Therefore, λ functions as a hyperparameter that regulates the overall magnitude of the penalty term for transaction costs. We examine various values for this parameter, and our empirical tests indicate that setting $\lambda = 1e-1$ (i.e., with the penalty term contributing 10% to the cost function) yields the best performance.[9]

Our study examines 65 years of monthly individual stock returns, spanning from January 1957 to December 2021, combined with time-lagged firm characteristics. Monthly stock return data is collected from the Center for Research in Securities Prices (CRSP) for all firms listed on the NYSE, AMEX, and NASDAQ. The time-lagged firm characteristics are constructed from 94 stock return predictors documented in the appendix of Gu, Kelly and Xiu (2020).[10] Since these characteristics are known to be highly skewed and leptokurtic, we apply rank normalization to transform them into the interval (-1, 1) for each month. For our analysis, the 65-year dataset is divided into three periods: an 18-year training period (1957–1974), a 12-year validation period (1975–1986), and a 35-year out-of-sample

---

[9] Other choices of λ include $1e-0$, $1e-2$, $1e-3$, and $1e-4$. Results based on these values exhibit a similar pattern.

[10] We download the firm characteristics from Dacheng Xiu's website.

test period (1987–2021). During the out-of-sample test, we implement a recursive re-training approach, where each model is re-trained annually with an expanding training sample while maintaining a fixed validation sample size.[11]

3 Empirical Results

This section presents our main empirical findings, comparing the performance of our transaction cost-adjusted machine learning model (TC model) with the standard conditional autoencoder model (plain model). We first examine the predictive power of the two models, followed by an evaluation of portfolio performance metrics. Finally, we investigate which stock characteristics drive these results.

3.1 Predictive Power of Alternative CA Models

We evaluate the predictive performance of the conditional autoencoder model under both the plain model and the transaction cost-adjusted model. The objective is to assess whether incorporating transaction costs directly into the estimation process enhances predictive accuracy. While machine learning models are often criticized for deriving predictive power primarily from microcap stocks, integrating transaction costs enables the model to more effectively filter out transient and high-cost signals, thereby enhancing robustness and yielding economically meaningful predictions. For assessing predictive performance, we use the out-of-sample predictive R^2 methodology following Campbell and Thompson (2008):

$$R^2_{\mathrm{OOS}} = 1 - \frac{\sum_{(i,t)\in t_{test}} (r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in t_{test}} (r_{i,t+1} - \tilde{r}_{i,t+1})^2} \tag{12}$$

where t_{test} represents the test dataset spanning from January 1987 to December 2021, $r_{i,t+1}$ denotes the realized excess return for stock i at month $t+1$, $\hat{r}_{i,t+1}$ is the predicted excess return

---

[11] That is, instead of extending the validation period, it is shifted forward by one year for each successive out-of-sample test.

from a machine learning model, and $\tilde{r}_{i,t+1}$ is the predicted excess return from a benchmark model. Following Gu, Kelly and Xiu (2020), we use a benchmark of 0 (i.e., $\tilde{r}_{i,t+1} = 0$), which implies the model is evaluated against a prediction of zero excess returns. A positive R^2_{OOS} indicates that the predictive model outperforms the benchmark in terms of out-of-sample predictive accuracy (lower average mean-squared prediction error), while a negative R^2_{OOS} implies the benchmark performs better.

Table 1 reports the predictive R^2 statistics, showing that the TC model consistently outperforms the plain model across all sample specifications. In the full sample, the overall predictive R^2 improves from 0.27% to 0.44%. Although the predictive R^2 is generally higher in the full sample, the relative improvement is particularly striking for the nonmicrocap stocks, where the predictive R^2 increases ninefold, from 0.03% to 0.27%. These results indicate that the plain model is primarily optimized to identify predictive patterns driven by microcap stocks, whereas incorporating transaction costs penalizes high-cost stocks—mostly microcaps—leading the model to shift its focus towards nonmicrocap stocks, where signals are more stable and implementable.

We further analyze the results by trading direction, based on the sign of predicted returns. The evidence indicates that predictive power is considerably higher for stocks with positive predicted returns (long positions) than for those with negative predicted returns (short positions), possibly reflecting the impact of short-sale constraints. More importantly, the TC model significantly enhances prediction accuracy across both long and short positions. This improvement is particularly pronounced for nonmicrocap stocks with positive predicted returns, where the difference in predictive R^2 reaches 0.43%, suggesting that incorporating transaction costs is especially effective for stocks that are more liquid and less constrained by short-selling limitations.

Figure 2 reports the predictive R^2 across transaction cost quintiles. Stocks are sorted by transaction costs, and the average R^2 is computed within each quintile. As expected, predictive performance is generally higher for stocks with greater transaction costs, consistent with the pattern in Table 1 in that the full sample exhibits higher R^2 than the nonmicrocap sample.

More importantly, the TC model exhibits consistently higher predictive R^2 across all transaction cost quintiles, with the most pronounced improvement in the lowest transaction cost quintile (from 0.06% to 0.36%). By adjusting the objective function to place greater weight on stocks with lower transaction costs, our model has successfully improved its ability to capture predictive patterns—whether risk-based or behavioral—in these more economically significant stocks. This aligns with the evidence from Table 1 that the nonmicrocap sample experiences more substantial improvements than the full sample. Notably, although modest, the improvement in the highest transaction cost quintile (from 0.60% to 0.63%) confirms that the model benefits all stocks—not just those with low transaction costs. This suggests that the model does not enhance predictive power for low-cost stocks at the expense of high-cost ones. Rather, by accounting for economically motivated constraints, the TC model improves predictive accuracy across the entire cross-section.

3.2 Portfolio Performance under Alternative CA Models

Table 2 reports the performance of portfolios sorted in decile by predicted returns estimated from the conditional autoencoder models. Specifically, we employ a CA model with five factors and two hidden layers (CAE5(2)). In the plain model, portfolio returns are highly profitable, consistent with the high-minus-low (H−L) portfolio results reported by Gu, Kelly and Xiu (2021), generating a monthly return of 3.1% in the full sample (Panel A). When the sample is restricted to nonmicrocap stocks, the return decreases to 1.3% (Panel B).

We also report portfolio returns net of transaction costs. Following DeMiguel, Garlappi and Uppal (2009) and Detzel, Novy-Marx and Velikov (2023), we define the net return for high-decile portfolios as:

$$r_{p,t}^{\text{net}} = r_{p,t} - (1 + r_{p,t}) \sum_{i=1}^{N} \left|\omega_{i,t} - \omega_{i,t-}\right| c_{i,t-1}, \tag{13}$$

where $r_{p,t}$ represents the gross return of the portfolio, $\omega_{i,t}$ denotes the weight of stock i in month t after rebalancing, and $\omega_{i,t-} = \frac{\omega_{i,t-1}(1+r_{p,t})}{\sum_{j=1}^{N}\omega_{j,t-1}(1+r_{p,t})}$ is the corresponding weight before rebalancing. The transaction cost for stock i at month $t-1$ is denoted by $c_{i,t-1}$, as estimated in Section 2. For low-decile portfolios, the net return is given by:

$$r_{p,t}^{\text{net}} = r_{p,t} + (1 + r_{p,t}) \sum_{i=1}^{N} \left|\omega_{i,t} - \omega_{i,t-}\right| c_{i,t-1}. \tag{14}$$

In this way, the net return of the hedge portfolio is computed as the difference between the net returns of the high- and low-decile portfolios. Considering transaction costs reduces portfolio performance similarly across models: net returns are lower than gross returns by 0.33% in the full sample and by 0.16% in the nonmicrocap sample. Nevertheless, the net returns remain economically and statistically significant.

To assess whether the improvement in predictive power of our transaction cost-adjusted model, as shown in Table 1, translates into better portfolio performance, we compare the returns of portfolios constructed using the plain model and the TC model. Imposing economic restrictions—such as transaction costs—can deteriorate in-sample performance by limiting the model's flexibility to fit the data. However, when such restrictions reflect economically meaningful signals and help prevent overfitting, they may lead to improved performance out-of-sample.

In Panel A, we compare the performance of the TC model to the plain model using the full sample. The hedge portfolio delivers approximately 5% higher monthly returns (3.25% vs. 3.10%) and a higher Sharpe ratio (1.51 vs. 1.44) under the TC model. The advantage of the TC model becomes significantly more pronounced in Panel B, which focuses on nonmicrocap stocks. While the performance of the plain model deteriorates notably once microcaps are excluded, the TC model continues to perform robustly, with the H–L portfolio generating substantially higher gross monthly returns (1.75% vs. 1.29%). The Sharpe ratio also improves markedly—by more than 40%—rising from 0.76 to 1.07. These findings are consistent with the evidence in Table 1, where the predictive power of the TC model improves considerably in the nonmicrocap sample, translating into a substantial enhancement in portfolio performance.

The outperformance of the TC model relative to the plain model can be decomposed into two components: the return attributable to systematic risk (risk premium) and the abnormal return. In Panel B, both the low and high decile portfolios in the TC model exhibit lower exposure to systematic risk compared to their counterparts in the plain model, suggesting that some of the Fama-French six factors tend to load on stocks with higher transaction costs. The reduction in systematic risk is comparable across the two portfolios, resulting in only a slight difference in the risk premium of the hedge portfolio: 0.65% $(= 1.75 - 1.10)$ in the TC model versus 0.63% $(= 1.29 - 0.66)$ in the plain model. Accordingly, approximately 95% of the improvement in portfolio performance in the nonmicrocap sample is attributable to abnormal returns.[12]

Moreover, although the increase in abnormal return arises from both the long and short positions, it is primarily driven by the long leg. Specifically, 74% of the abnormal return improvement comes from the long position.[13] This finding aligns with Table 1, which shows that the TC model significantly improves predictability for stocks with positive expected returns. Overall, these results indicate that the superior performance of the TC model is not simply due to increased exposure to common risk factors (e.g., Fama-French factors), nor is it predominantly driven by short positions, which are often constrained by short-sale limitations.[14]

---

[12] That is, $\frac{1.10-0.661}{1.751-1.288} = 95\%$.

[13] This corresponds to $\frac{0.683-0.359}{1.10-0.661} = 74\%$.

[14] Our approach does not directly accommodate short-sale constraints. However, this evidence suggests that the TC model does not improve predictability primarily by capturing prevalent overpricing in stock markets, as documented by Stambaugh, Yu and Yuan (2012). Related studies, such as Drechsler and Drechsler (2016), Engelberg, Reed and Ringgenberg (2018), and Patton and Weller (2020), emphasize the impact of short-sale constraints on limiting arbitrage.

3.3 Characteristic Importance Analysis

Our TC model enhances portfolio performance by systematically recalibrating the importance of trading friction characteristics while preserving the influence of fundamental signals. This section examines these shifts through two complementary approaches: partial derivatives to understand the model's internal mechanics and feature removal to measure economic impact on risk-adjusted returns. These methods reveal that the TC model reduces reliance on short-term price trends and liquidity measures predominantly in microcap stocks, while maintaining consistent characteristic relationships in economically significant securities.

3.3.1 Importance Based on Partial Derivatives

While Gu, Kelly and Xiu (2020) measure characteristic importance using in-sample predictive R^2, we extend their approach to the out-of-sample context, providing a more robust assessment of how each characteristic influences model predictions in real-world applications. This approach captures the model's internal decision-making process and helps identify whether the TC model's superior performance stems from systematically different valuations of specific characteristics.

We measure characteristic influence using partial derivatives of predicted returns with respect to each input variable, following Dimopoulos, Bourret and Lek (1995):

$$D_{OOS} = \sum_{(i,t)\in T_{training}} \left| \frac{\partial r_{i,t}}{\partial z_{i,t-1}} \Big|_{z'_{i,t-1}} \right| \tag{15}$$

Panel A of Figure 3 presents results for the full sample. The plain model exhibits strong concentration in a few characteristics, with 1-month momentum (mom1m) dominating at an importance of 0.97, followed by log(market size) (mvel1, 0.77), return volatility (retvol, 0.71), and illiquidity (ill, 0.65). These align closely with Gu, Kelly and Xiu (2021), with 17 of the top 20 characteristics matching their findings. The dominant characteristics form distinct groups: (1) Price trends (momentum variables and maximum daily return), (2) Liquidity measures (market size, illiquidity, bid-ask spread, trading volume metrics), and (3) Risk measures (volatility and beta variables).

The TC model systematically reduces the importance of these trading-friction characteristics as classified by Hou, Xue and Zhang (2020). Quantitatively, 1-month momentum falls from 0.97 to 0.54 (44.48% decrease), market equity drops from 0.78 to 0.58 (25.4% decrease), and illiquidity declines from 0.65 to 0.46 (29.3% decrease). Similarly, bid-ask spread decreases by 37.0%, dollar trading volume falls by 33.7%, and turnover drops by 33.3%. At the category level, the TC model reduces the average importance of Price trend characteristics by 26.12%, Liquidity characteristics by 30.09%, and Risk measures by 20.10%.

Most remarkably, when analyzing only nonmicrocap stocks (Panel B of Figure 3), the differences in characteristic importance between Plain and TC models almost completely disappear. Most characteristics show differences of less than 1% in importance between the models, with some variables like 36-month momentum, book-to-market, and firm age actually becoming slightly more important in the TC model. This pattern confirms that our TC model primarily adjusts the influence of characteristics in microcap stocks—those with high trading costs—while preserving variable relationships for economically significant stocks.

3.3.2 Importance Based on Risk-Adjusted Returns

Our second approach quantifies the contribution of each characteristic to portfolio performance, following methods proposed by Gu, Kelly and Xiu (2020) and extended in Jo and Kim (2025). This approach directly measures economic impact, revealing which characteristics genuinely drive risk-adjusted returns beyond standard factors. Jo and Kim (2025) demonstrate that conventional importance measures based on in-sample predictive R^2 can be misleading due to microcap bias and disconnection from economic significance.

For a trained model f^* with input characteristics matrix Z, we measure each characteristic's importance by calculating how portfolio performance changes when that characteristic is neutralized:

$$\Delta_j = g(f^*(Z)) - g(f^*(Z|z_j = 0)) \tag{16}$$

where $g(\cdot)$ represents Fama-French six-factor (FF6) adjusted returns and $Z|z_j = 0$ indicates the input matrix with the jth characteristic set to zero. We use Fama-French six-factor adjusted returns because it isolates returns not explained by common risk factors. We then normalize positive and negative impacts separately:

$$VI_j^{positive} = \frac{\Delta_j}{\sum_{k \in K^+}(\Delta_k)}, \quad \text{where } \Delta_j \geq 0 \tag{17}$$

$$VI_j^{negative} = \frac{\Delta_j}{\sum_{k \in K^-}(\Delta_k)}, \quad \text{where } \Delta_j < 0 \tag{18}$$

This approach identifies characteristics that generate genuine risk-adjusted returns beyond the Fama-French six-factor model, distinguishing between those that enhance performance and those that detract from it.

For the full sample (Panel A of Figure 4), the plain model's portfolio performance depends heavily on a few dominant characteristics: illiquidity (14.36%), 1-month momentum (12.93%), and 6-month momentum (9.62%), followed by return volatility (7.08%), dollar trading volume (6.45%), and bid-ask spread (5.81%). These are predominantly trading friction variables identified by Hou, Xue and Zhang (2020). The model also exhibits significant negative sensitivity to several characteristics, with sales-to-inventory (7.87%), sales-to-price (6.07%), and secured debt indicators (5.58%) reducing performance when included. Of the 94 characteristics, 34 exhibit negative importance—their removal actually improves performance.

The TC model shows a markedly different pattern. Positive influences become more evenly distributed, with return volatility (5.03%), 1-month momentum (4.92%), and both 6-month momentum and illiquidity (3.77%) as top contributors, but with substantially reduced magnitudes. Most striking are the shifts in negative contributors—market capitalization undergoes a dramatic reversal from positive (2.50%) to strongly negative (-53.53%), followed by beta (-16.14%) and book-to-market (-9.20%). Despite these extreme negative values, the total number of variables with negative importance falls to 27.

The TC model transforms this pattern, exhibiting a more uniform distribution of importance across characteristics with only nine showing negative values. This radical redistribution represents a fundamental shift in how the model evaluates characteristics, moving from extreme dependence on a few liquidity variables to a more balanced approach that extracts value from a broader set of firm attributes.

3.3.3 Economic Implications

Both analytical approaches reveal consistent patterns that explain the TC model's superior performance. By incorporating transaction costs directly into the estimation process, our model systematically reduces dependence on characteristics associated with trading frictions while preserving characteristic relationships in economically significant securities. The TC model shifts from extreme concentration on a few liquidity variables to a more balanced utilization of diverse firm characteristics, and dramatically changes which characteristics negatively impact performance.

The category-level patterns are particularly revealing: price trend measures show an average 26% importance reduction in the full sample but minimal change in nonmicrocaps, while liquidity measures show a 30% reduction in the full sample but maintain consistent importance in economically significant stocks.

This selective recalibration allows the TC model to maintain strong predictive power where it matters most for institutional investors, who typically focus on larger, more liquid securities. By preventing the model from extracting misleading signals from difficult-to-arbitrage securities, our approach identifies characteristics that capture genuine risk-return relationships rather than transitory mispricings that persist due to arbitrage constraints. Our approach complements recent work by Jo and Kim (2025), who demonstrate that standard importance measures in machine learning models are heavily influenced by microcap stocks.

These findings align with Hou, Xue and Zhang (2020), who document that many anomalies in the trading friction category fail statistical significance tests once microcaps are controlled for. Our TC model effectively addresses this issue by automatically downweighting signals from high-friction securities during the estimation process, resulting in more economically meaningful and implementable investment strategies.

3.4 Performance of Linear Autoencoder Models

We investigate whether the observed performance improvements in our transaction cost model derive predominantly from linear or nonlinear relationships in stock characteristics. To test this hypothesis, we implement a six-factor conditional autoencoder model with no hidden layers (CAE6(0)), which effectively functions as a linear autoencoder model comparable to the Instrumented Principal Component Analysis (IPCA) of Kelly, Pruitt and Su (2019).

Our model specifications follow Gu, Kelly and Xiu (2021), who find that the five-factor model with two hidden layers (CAE5(2)) delivers optimal performance among nonlinear configurations, while the six-factor model with no hidden layers (CAE6(0)) performs best among linear specifications. These specific architectures have been established as standard benchmarks in the literature, with Avramov, Cheng and Metzker (2023) also adopting these same CAE5(2) and CAE6(0) configurations in their evaluation of machine learning models in asset pricing. By adopting these specific architectures, we ensure that each model represents the strongest possible implementation within its class (linear or nonlinear). The CAE5(2) model incorporates nonlinear transformations through its hidden layers with ReLU activation functions, while the CAE6(0) model represents a linear specification with direct connections between inputs and outputs. This deliberate selection of optimal architectures for both linear and nonlinear models allows us to more precisely isolate the contribution of nonlinearity to our results while controlling for other aspects of model design.

The theoretical foundation for investigating both linear and nonlinear specifications is supported by recent literature on asset pricing models. Chen, Pelger and Zhu (2024) argue that nonlinear asset pricing models can uncover risk factors that linear models fail to detect, particularly in complex market environments with multiple market frictions. Similarly, Kozak, Nagel and Santosh (2020) demonstrate that while simplified linear models often perform adequately in many asset pricing applications, specific market characteristics necessitate nonlinear approaches to capture more nuanced interactions.

Table 4 presents the performance metrics for the CAE6(0) models. Several observations emerge from this analysis. First, both plain and transaction cost versions of the linear model generate significant returns, but with lower magnitudes compared to their nonlinear counterparts. In the full sample, the TC model's HML portfolio generates a gross return of 1.645% (versus 3.249% for CAE5(2)) and a net return of 1.447% (versus 2.914% for CAE5(2)). This substantial difference highlights the importance of capturing nonlinear relationships in predicting stock returns.

Despite the overall lower performance, the TC model still consistently outperforms the plain model in the linear specification. For the full sample, the TC model generates higher gross returns (1.645% versus 1.295%), higher net returns (1.447% versus 1.092%), and improved Sharpe ratios (0.937 versus 0.796) compared to the plain model. The FF6 alpha is also higher for the TC model (0.823% versus 0.701%).

When examining nonmicrocap stocks, we observe that both models experience a performance decline, but the TC model maintains better performance. The TC model's HML portfolio delivers a gross monthly return of 1.206% (versus 1.071% for the plain model) and a net return of 1.035% (versus 0.891%).

Importantly, when analyzing the sources of risk-adjusted returns in the nonmicrocap sample, we find a notable difference in how performance is distributed between the long and short sides. In the plain model (Panel C), the FF6 risk-adjusted returns are concentrated in the long side, with the Low portfolio generating insignificant risk-adjusted returns of

-0.196% (t-statistic $= -1.48$) while the High portfolio produces significant risk-adjusted returns of 0.331%** (t-statistic $= 2.37$). In the TC model (Panel D), we observe that the performance is even more strongly concentrated in the long side, with insignificant risk-adjusted returns in the Low portfolio (-0.055% with t-statistic $= -0.42$) and larger, highly significant risk-adjusted returns in the High portfolio (0.457%*** with t-statistic $= 3.58$).

This shift in risk-adjusted return distribution indicates that even in a linear specification, our TC model reduces reliance on the short side for performance and enhances the contribution from the long side. While the High portfolio's risk-adjusted returns in the linear TC model do not reach the level of significance observed in the nonlinear specification, the rebalancing of return sources away from the short side represents an economically meaningful improvement given the implementation challenges associated with short positions.

The performance gap between the plain and TC models appears smaller in the linear specification than in the nonlinear models. For nonmicrocap stocks, the difference in gross HML returns is 0.135 percentage points (1.206% - 1.071%) for CAE6(0), compared to 0.463 percentage points (1.751% - 1.288%) for CAE5(2). This suggests that while incorporating transaction costs improves performance in both linear and nonlinear models, the benefits are more substantial when nonlinear relationships are considered.

These findings indicate that both linear and nonlinear components contribute to our TC model's performance improvements, with the nonlinear specification providing more substantial benefits. This aligns with Giglio, Liao and Xiu (2021), who document that model specification significantly affects anomaly detection, and is consistent with Gu, Kelly and Xiu (2021), who document that nonlinear models better capture conditional relationships in asset pricing. The linear component provides a baseline improvement by shifting focus toward stocks with lower transaction costs, while the nonlinear component captures complex interactions between stock characteristics and transaction costs. As DeMiguel, Martín-Utrera and Nogales (2014) demonstrate, portfolio optimization with transaction costs becomes analytically intractable with multiple assets, requiring more sophisticated modeling approaches to derive optimal trading strategies.

4 Conclusion

This paper proposes a transaction cost-adjusted machine learning model that enhances the economic relevance and robustness of stock return prediction. Building on the conditional autoencoder framework of Gu, Kelly and Xiu (2021), we incorporate transaction costs directly into the model's optimization process, penalizing costly-to-trade stocks while retaining the full stock universe. Our approach mitigates the small-stock bias and addresses overfitting to illiquid stocks, which often limits the practical use of standard machine learning models in asset pricing. Empirically, we find that the transaction cost-adjusted model (TC model) consistently improves the out-of-sample performance of portfolios constructed based on machine learning signals, particularly when microcap stocks are excluded. Furthermore, the TC model reduces reliance on a narrow set of trading-friction-related characteristics and significantly lowers the number of characteristics that detract from out-of-sample performance. By embedding economic frictions into model training, our framework better aligns statistical prediction with real-world investment considerations, offering a more practical and economically meaningful application of machine learning to asset pricing.

References

Avramov, D., Cheng, S. and Metzker, L. (2023) Machine learning vs. economic restrictions: Evidence from stock return predictability, *Management Science*, **69**, 2587–2619.

Brandt, M. W., Santa-Clara, P. and Valkanov, R. (2009) Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns, *The Review of Financial Studies*, **22**, 3411–3447.

Campbell, J. Y. and Thompson, S. B. (2008) Predicting excess stock returns out of sample: Can anything beat the historical average?, *The Review of Financial Studies*, **21**, 1509–1531.

Chen, L., Pelger, M. and Zhu, J. (2024) Deep learning in asset pricing, *Management Science*, **70**, 714–750.

Cong, L. W., Tang, K., Wang, J. and Zhang, Y. (2021) Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai, *SSRN Electronic Journal*.

DeMiguel, V., Garlappi, L. and Uppal, R. (2009) Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?, *The Review of Financial Studies*, **22**, 1915–1953.

DeMiguel, V., Martín-Utrera, A. and Nogales, F. J. (2014) Parameter uncertainty in multiperiod portfolio optimization with transaction costs, *Journal of Financial and Quantitative Analysis*, **49**, 1329–1359.

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

Detzel, A., Novy-Marx, R. and Velikov, M. (2023) Model comparison with transaction costs, *The Journal of Finance*, **78**, 1743–1775.

Dimopoulos, Y., Bourret, P. and Lek, S. (1995) Use of some sensitivity criteria for choosing networks with good generalization ability, *Neural Processing Letters*, **2**, 1–4.

Drechsler, I. and Drechsler, Q. F. (2016) The shorting premium and asset pricing anomalies, *Review of Financial Studies*, **29**, 3466–3501.

Engelberg, J. E., Reed, A. V. and Ringgenberg, M. C. (2018) Short-selling risk, *Journal of Finance*, **73**, 755–786.

Frazzini, A., Israel, R. and Moskowitz, T. J. (2018) Trading costs, *SSRN Electronic Journal*.

Gârleanu, N. and Pedersen, L. H. (2013) Dynamic trading with predictable returns and transaction costs, *The Journal of Finance*, **68**, 2309–2340.

Giglio, S., Liao, Y. and Xiu, D. (2021) Thousands of alpha tests, *Review of Financial Studies*, **34**, 3456–3496.

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

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

Gu, S., Kelly, B. and Xiu, D. (2021) Autoencoder asset pricing models, *Journal of Econometrics*, **222**, 429–450.

Hand, J. R. and Green, J. (2011) The importance of accounting information in portfolio optimization, *Journal of Accounting, Auditing, and Finance*, **26**, 1–33.

Hou, K., Xue, C. and Zhang, L. (2020) Replicating anomalies, *The Review of Financial Studies*, **33**, 2019–2133.

Jensen, T. I., Kelly, B. T., Malamud, S. and Pedersen, L. H. (2024) Machine learning and the implementable efficient frontier, *The Review of Financial Studies*.

Jiang, J., Kelly, B. and Xiu, D. (2023) (re-)imag(in)ing price trends, *The Journal of Finance*, **78**, 3193–3249.

Jo, Y. and Kim, Y. (2025) Rethinking variable importance in machine learning: An economic perspective on empirical asset pricing, working Paper.

Kelly, B. T., Pruitt, S. and Su, Y. (2019) Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics*, **134**, 501–524.

Kozak, S., Nagel, S. and Santosh, S. (2020) Shrinking the cross-section, *Journal of Financial Economics*, **135**, 271–292.

Novy-Marx, R. and Velikov, M. (2016) A taxonomy of anomalies and their trading costs, *The Review of Financial Studies*, **29**, 104–147.

Patton, A. J. and Weller, B. M. (2020) What you see is not what you get: The costs of trading market anomalies, *Review of Financial Studies*, **33**, 4069–4117.

Stambaugh, R. F., Yu, J. and Yuan, Y. (2012) The short of it: Investor sentiment and anomalies, *Journal of financial economics*, **104**, 288–302.

Figure 1: **Distribution of Transaction Costs**

This figure illustrates the distribution of monthly transaction costs, as described in Section 2.1. Transaction costs are assumed to decline linearly over time and inversely with firm size. The solid line indicates the median, while additional percentiles are also shown.

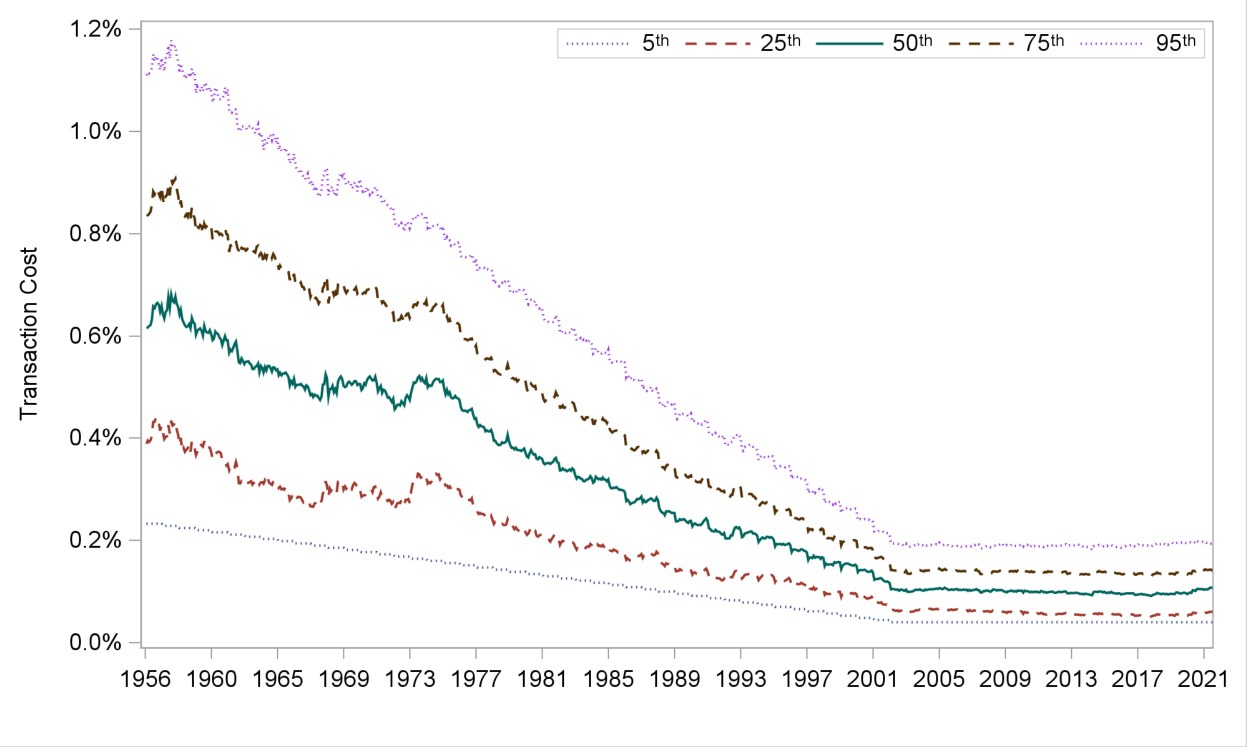


Figure 2: **Out-of-Sample Predictive R^2 Across Transaction Cost Deciles**
This figure illustrates the out-of-sample predictive R^2 across transaction cost quintiles for both the plain model and the transaction cost model.

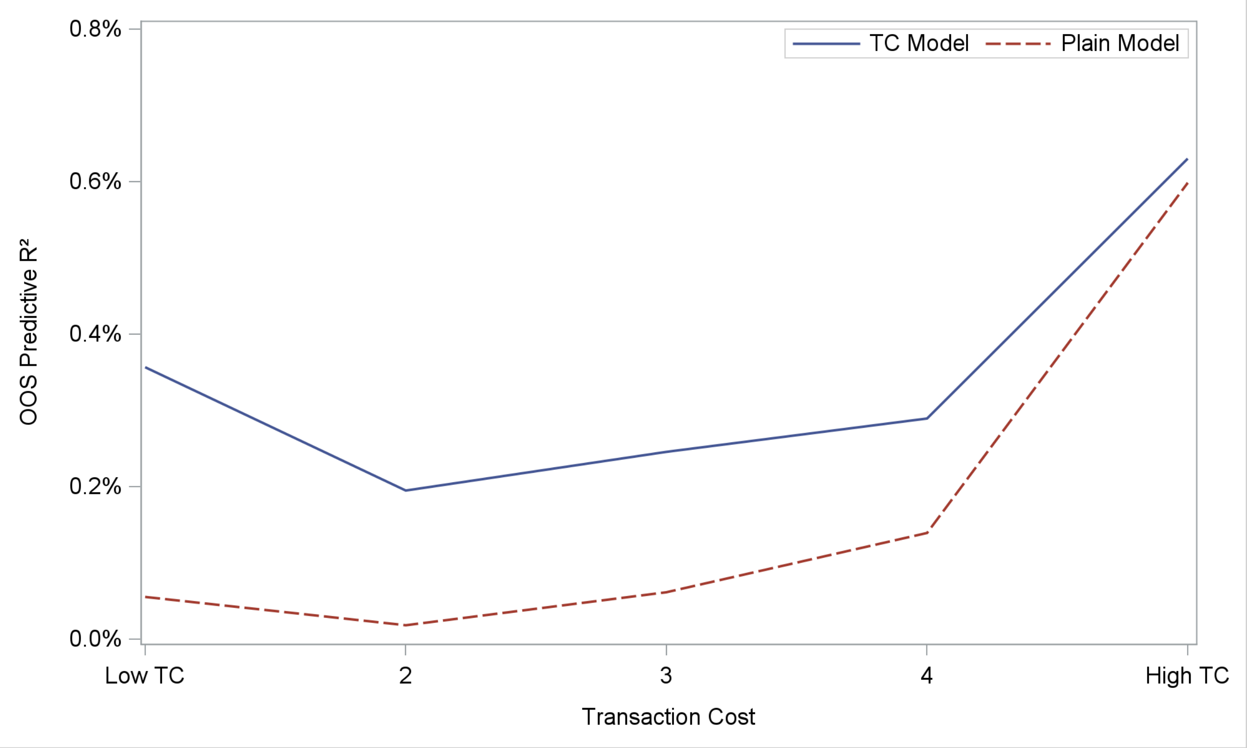
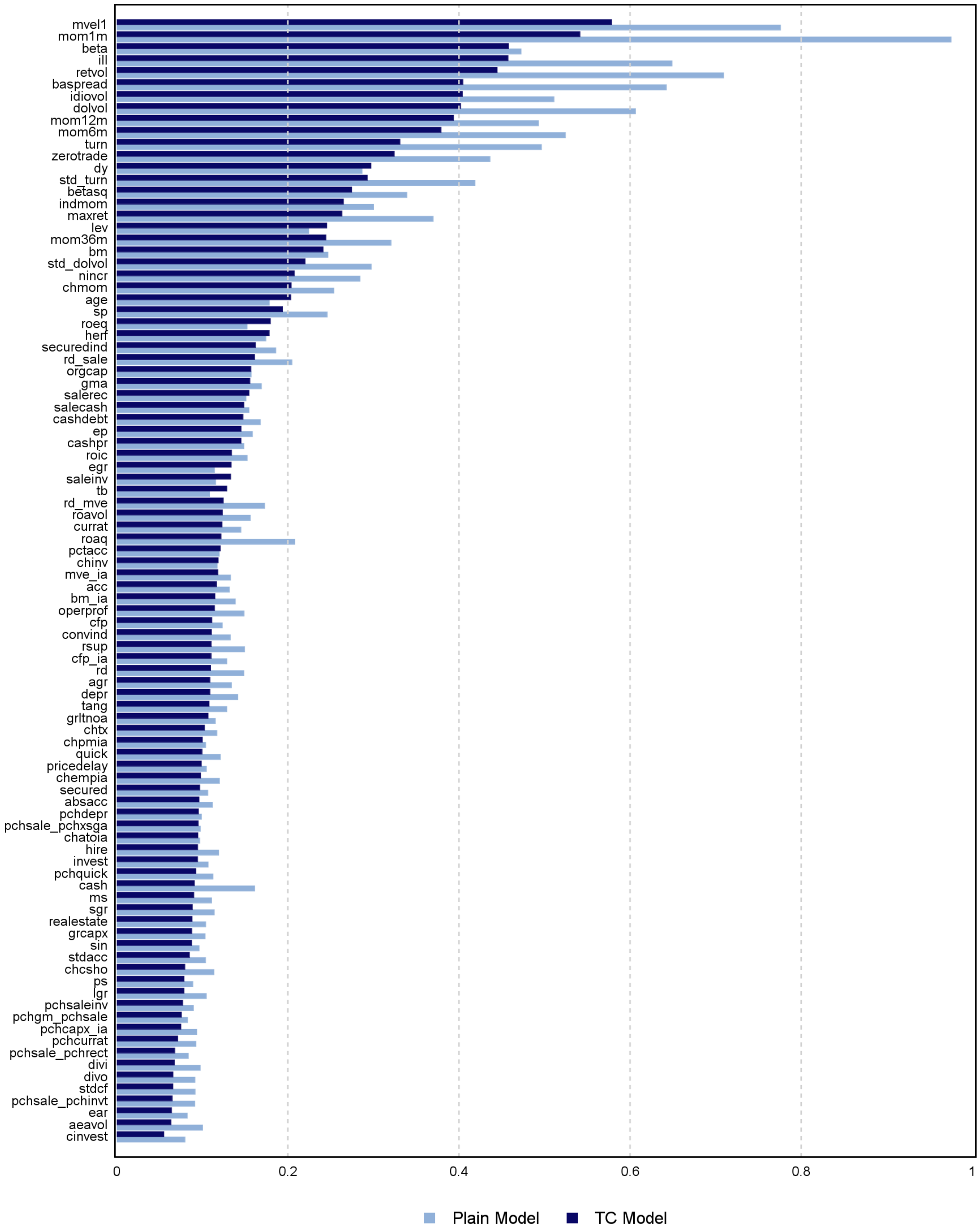


Figure 3: **Characteristic Importance Based on Absolute Partial Derivatives**
This figure illustrates the importance of 94 characteristics based on absolute partial derivatives from the full sample (Panel A) and the nonmicrocap sample (Panel B). Annual absolute partial derivatives are normalized so that the most important characteristic has a value of one and then averaged over time. Dark blue bars represent the transaction cost model, and light blue bars represent the plain model. Characteristics are ordered by their importance in the TC model, with the most important at the top.

Panel A. Full Sample



Panel B. Nonmicrocap Sample

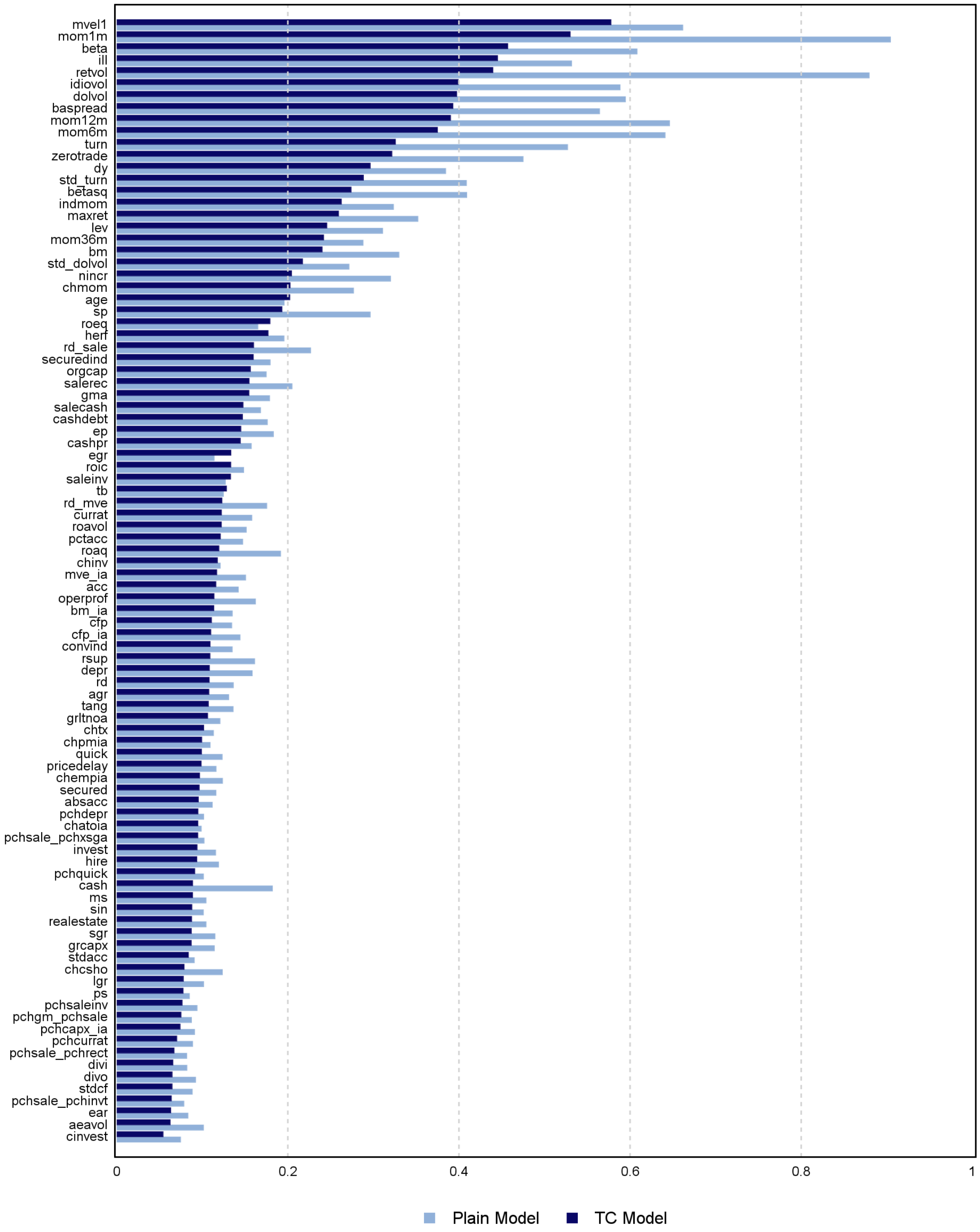
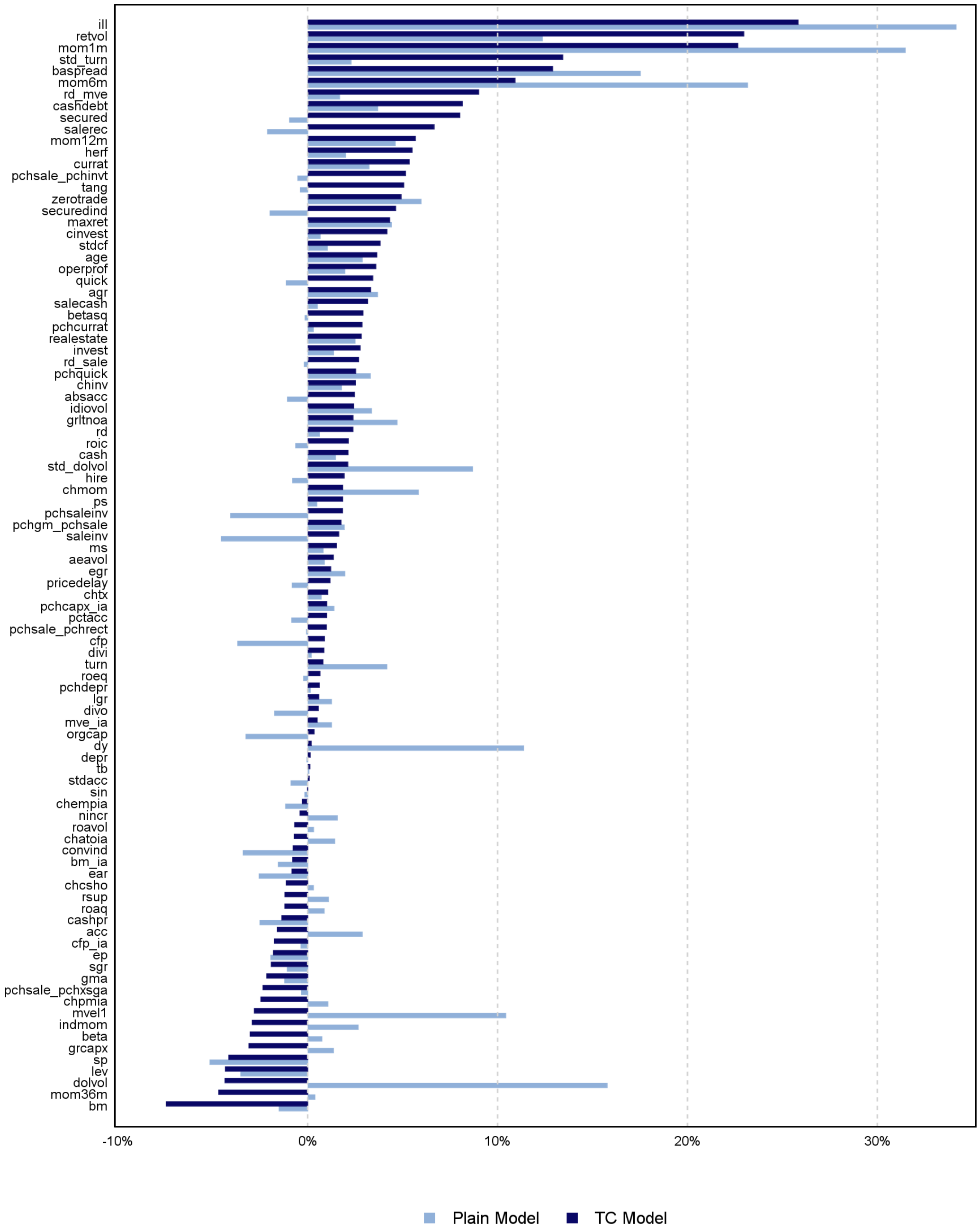


Figure 4: **Characteristic Importance Based on Out-of-Sample Abnormal Returns** This figure illustrates the importance of 94 characteristics from the full sample (Panel A) and the nonmicrocap sample (Panel B). Importance is measured by the percentage reduction in out-of-sample abnormal returns of long–short portfolios when the value of a given characteristic is set to zero. Abnormal returns are estimated using the Fama-French six-factor model. Dark blue bars represent the transaction cost model, and light blue bars represent the plain model. Characteristics are ordered by their importance in the TC model, with the most important at the top.

Panel A. Full Sample



Panel B. Nonmicrocap Sample

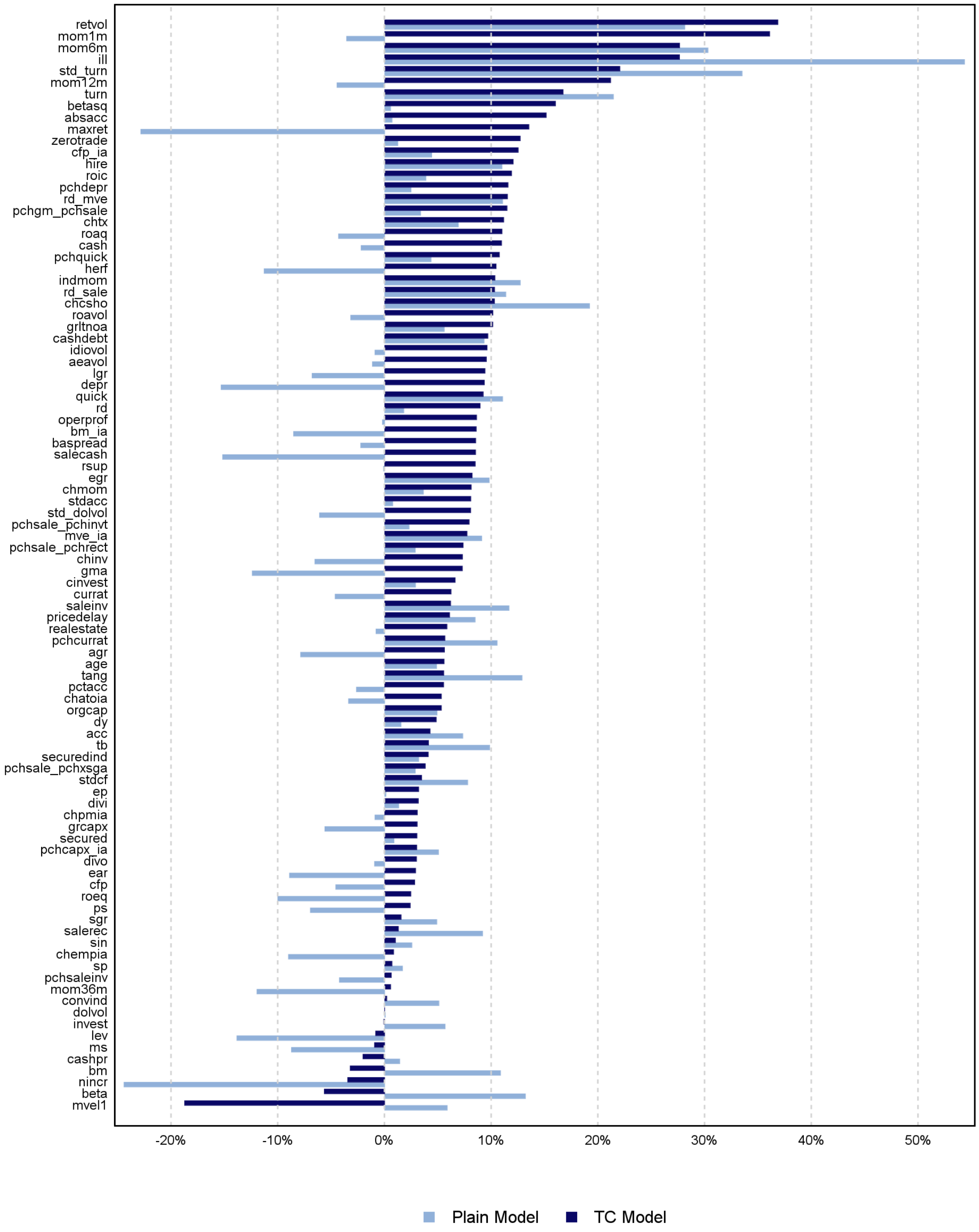


Table 1: **Out-of-Sample Predictive R^2**

This table presents out-of-sample predictive R^2 for excess returns on individual stocks, comparing the plain model and the TC model. Average R^2 values are reported for the full and nonmicrocap samples, partitioned by the sign of predicted excess returns (positive or negative).

Sample	Plain Model	TC Model	Difference
Full	0.27%	0.44%	0.16%
Full ($\hat{r} > 0$)	1.11%	1.29%	0.18%
Full ($\hat{r} < 0$)	0.13%	0.31%	0.19%
Nonmicrocaps	0.03%	0.27%	0.24%
Nonmicrocaps ($\hat{r} > 0$)	0.64%	1.08%	0.43%
Nonmicrocaps ($\hat{r} < 0$)	-0.69%	-0.57%	0.12%

Table 2: **Performance of Portfolios Sorted by Returns Predicted by Conditional Autoencoder Models**

This table presents the performance of value-weighted portfolios sorted by expected returns from conditional autoencoder (CA) models from 1987 to 2021. CA models with five factors and two hidden layers (CAE5(2)) are estimated under the plain model and the transaction cost model, respectively. At the beginning of each month *t*, portfolios are formed by sorting into deciles based on these predicted returns. Performance is reported as gross returns (before transaction costs) and net returns (after transaction costs). FF6 represents the abnormal return relative to the Fama-French six-factor model. The Sharpe ratio (SR) is annualized. Panel A uses the full sample, and Panel B restricts the sample to nonmicrocaps. *t*-statistics, adjusted using Newey-West standard errors with four lags, are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Panel A. Full Sample

	Expected Return	Gross Return			Net Return		
		Return	FF6	SR	Return	FF6	SR
		Portfolio Performance Based on the Plain Model					
Low	-0.019	-0.745	-1.091***	-0.282	-0.621	-0.967***	-0.235
		(-1.61)	(-4.78)		(-1.34)	(-4.27)	
High	0.033	2.349***	1.740***	0.960	2.148***	1.538***	0.879
		(5.42)	(5.33)		(4.98)	(4.78)	
H−L	0.052	3.095***	2.831***	1.442	2.769***	2.505***	1.298
		(7.56)	(6.28)		(6.88)	(5.65)	
		Portfolio Performance Based on the Transaction Cost Model					
Low	-0.018	-1.077**	-1.196***	-0.337	-0.950**	-1.069***	-0.286
		(-2.32)	(-5.09)		(-2.05)	(-4.60)	
High	0.032	2.172***	1.668***	0.969	1.964***	1.461***	0.887
		(4.84)	(4.03)		(4.39)	(3.56)	
H−L	0.050	3.249***	2.864***	1.512	2.914***	2.530***	1.363
		(7.30)	(5.01)		(6.64)	(4.48)	

Panel B. Nonmicrocap Sample

	Expected Return	Gross Return			Net Return		
		Return	FF6	SR	Return	FF6	SR
		Portfolio Performance Based on the Plain Model					
Low	-0.012	0.089	-0.302*	0.039	0.167	-0.226	0.073
		(0.22)	(-1.68)		(0.42)	(-1.25)	
High	0.018	1.378***	0.359***	0.845	1.292***	0.273**	0.793
		(4.88)	(2.84)		(4.59)	(2.18)	
H-L	0.031	1.288***	0.661***	0.758	1.125***	0.499**	0.662
		(4.30)	(3.00)		(3.76)	(2.26)	
		Portfolio Performance Based on the Transaction Cost Model					
Low	-0.011	-0.271	-0.417**	-0.014	-0.189	-0.335*	0.023
		(-0.67)	(-2.36)		(-0.47)	(-1.90)	
High	0.018	1.480***	0.683***	0.989	1.398***	0.601***	0.943
		(4.66)	(4.98)		(4.40)	(4.40)	
H-L	0.029	1.751***	1.100***	1.071	1.587***	0.937***	0.971
		(5.65)	(4.37)		(5.13)	(3.72)	

Table 3: **Performance of Portfolios Sorted by Returns Predicted by Alternative Conditional Autoencoder Models**

This table presents the performance of value-weighted portfolios sorted by expected returns from alternative conditional autoencoder (CA) models from 1987 to 2021. CA models with six factors and no hidden layer (CAE6(0)) are estimated under the plain model and the transaction cost model, respectively. At the beginning of each month *t*, portfolios are formed by sorting into deciles based on these predicted returns. Performance is reported as gross returns (before transaction costs) and net returns (after transaction costs). FF6 represents the abnormal return relative to the Fama-French six-factor model. The Sharpe ratio (SR) is annualized. Panel A uses the full sample, and Panel B restricts the sample to nonmicrocaps. *t*-statistics, adjusted using Newey-West standard errors with four lags, are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Panel A. Full Sample

	Expected Return	Gross Return			Net Return		
		Return	FF6	SR	Return	FF6	SR
Portfolio Performance Based on the Plain Model							
Low	-0.011	0.201	-0.203	0.118	0.286	-0.120	0.167
		(0.66)	(-1.44)		(0.93)	(-0.84)	
High	0.024	1.496***	0.498***	0.815	1.378***	0.380**	0.751
		(4.64)	(2.93)		(4.29)	(2.25)	
H−L	0.036	1.295***	0.701***	0.796	1.092***	0.499**	0.672
		(4.48)	(2.97)		(3.79)	(2.12)	
Portfolio Performance Based on the Transaction Cost Model							
Low	-0.008	0.187	-0.191	0.115	0.264	-0.115	0.162
		(0.63)	(-1.25)		(0.89)	(-0.75)	
High	0.022	1.832***	0.632***	0.906	1.711***	0.510***	0.846
		(5.13)	(3.71)		(4.79)	(3.02)	
H−L	0.030	1.645***	0.823***	0.937	1.447***	0.625**	0.824
		(5.34)	(3.31)		(4.70)	(2.51)	

Panel B. Nonmicrocap Sample

	Expected Return	Gross Return			Net Return		
		Return	FF6	SR	Return	FF6	SR
			Portfolio Performance Based on the Plain Model				
Low	-0.010	0.252	-0.196	0.154	0.337	-0.111	0.206
		(0.87)	(-1.48)		(1.16)	(-0.84)	
High	0.021	1.322***	0.331**	0.821	1.228***	0.237*	0.763
		(4.72)	(2.37)		(4.39)	(1.71)	
H−L	0.031	1.071***	0.527***	0.769	0.891***	0.348*	0.640
		(4.44)	(2.83)		(3.70)	(1.86)	
			Portfolio Performance Based on the Transaction Cost Model				
Low	-0.008	0.333	-0.055	0.212	0.414	0.026	0.263
		(1.18)	(-0.42)		(1.47)	(0.19)	
High	0.018	1.539***	0.457***	0.907	1.449***	0.366***	0.854
		(5.31)	(3.58)		(5.00)	(2.89)	
H−L	0.026	1.206***	0.512***	0.832	1.035***	0.341*	0.713
		(5.07)	(2.69)		(4.35)	(1.78)	

Table 4: **Performance of Portfolios Sorted by Returns Predicted by Conditional Autoencoder Models Trained on Nonmicrocaps**

This table presents the performance of value-weighted portfolios sorted by expected returns from conditional autoencoder (CA) models from 1987 to 2021. CA models with five factors and two hidden layers (CAE5(2)) are estimated on only nonmicrocap stocks under the plain model and the transaction cost model, respectively. At the beginning of each month *t*, portfolios are formed by sorting into deciles based on these predicted returns. Performance is reported as gross returns (before transaction costs) and net returns (after transaction costs). FF6 represents the abnormal return relative to the Fama-French six-factor model. The Sharpe ratio (SR) is annualized. *t*-statistics, adjusted using Newey-West standard errors with four lags, are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

	Expected Return	Gross Return			Net Return		
		Return	FF6	SR	Return	FF6	SR
Portfolio Performance Based on the Plain Model							
Low	-0.009	0.028	-0.211	0.012	0.092	-0.148	0.039
		(0.066)	(-1.202)		(0.214)	(-0.838)	
High	0.018	1.474***	0.398***	0.880	1.397***	0.320**	0.834
		(5.035)	(2.770)		(4.777)	(2.241)	
H−L	0.027	1.446***	0.609***	0.826	1.305***	0.468**	0.746
		(4.449)	(2.879)		(4.016)	(2.214)	
Portfolio Performance Based on the Transaction Cost Model							
Low	0.000	0.257	0.000	0.117	0.344	0.086	0.157
		(0.647)	(0.002)		(0.866)	(0.433)	
High	0.015	1.325***	0.303**	0.692	1.251***	0.228	0.653
		(4.040)	(2.010)		(3.813)	(1.516)	
H−L	0.014	1.069***	0.302	0.583	0.907***	0.141	0.495
		(3.495)	(1.317)		(2.967)	(0.613)	