Algorithmic Alchemy:
Turning Data into Stock Market Gold

Abstract

The aim of this paper is to present a comprehensive investigation into the predictive accuracy of 14 models in forecasting monthly stock returns, including 13 machine learning techniques and one simple linear model. Adopting a rich dataset spanning over 53 years with 5175 stocks for a total of 758649 observations, we evaluate the models on out-of-sample predictive R^2 and identify Random Forest and specific Neural Networks (NN3 and NN5) as the best performing methods. Our analysis spans six time periods, showing the varying performance of these models and underscoring their adaptability, especially during significant market upheavals like the pandemic. Additionally, the paper identifies key indicators that drive stock returns, including Valuation Ratio, Liquidity, Price Trend, and Chicago Fed National Financial Conditions Index (NFCI). We also reveal that prediction accuracy is primarily driven by data rather than being model-driven. Lastly, we demonstrate that in real-world markets, model-driven portfolios consistently outperform our benchmark, the S&P 500 index return. These results collectively enrich our understanding of machine learning's role in empirical asset pricing and provide practical implications for both scholars and practitioners.

Key words: Machine Learning, Deep Learning, Asset Pricing, Random Forest, Neural Networks

JEL codes: C52, C55, C58, G0, G1, G17

1. Introduction

The application of machine learning algorithms in the field of asset pricing has witnessed a paradigm shift over the past few decades. While traditional empirical asset pricing models have served as valuable tools for investors and researchers alike, they often fall short in capturing the intricacies and non-linear relationships inherent in financial markets especially for return prediction (Gu et al., 2020; Leippold et al., 2022). In contrast, machine learning methods offer richer specifications of functional forms and accommodate a broader array of predictor variables, leveraging the advancements in computational power to provide more precise and effective tools for navigating financial markets (Chen et al., 2023; Gunnarsson et al., 2024).

This shift is further justified by challenges to the credibility of predictive patterns in stock returns based on firm characteristics. For instance, Harvey et al. (2016) raise concerns about the potential for false discoveries among significant factors previously identified. McLean and Pontiff (2016) observe a substantial decline in the out-of-sample profitability of anomaly-based portfolios, while Hou et al. (2020) note that a significant portion of anomalies lose their significance when adjusting for factors such as microcap stocks and using value-weighted returns. The decimalization of U.S. equity markets in 2001 has also contributed to the attenuation of predictable patterns, attributed to increased market liquidity and arbitrage activity (Chordia et al., 2014). Despite these challenges, the integration of machine learning into asset pricing has uncovered new, profitable investment strategies (Bianchi et al., 2021; Feng et al., 2020; Freyberger et al., 2020; Harris et al., 2024). The rise of financial technology has popularized the use of machine learning to develop advanced investment systems capable of outperforming traditional methods and human fund managers (Avramov et al., 2020).

Our research is built upon the work of Gu et al. (2020), a foundational pillar in the integration of machine learning within the asset pricing area. Our primary contributions are fourfold. First, our analysis transcends the framework of Regression models, Tree-based models, and neural networks, by also embracing the advanced field of deep learning. We integrate a total of 14 models within our study, comprising 13 advanced machine learning algorithms alongside one straightforward linear model. This ensemble encompasses nearly the entire spectrum of models prevalent in the asset pricing domain. Specifically, the collection includes Ordinary Least Squares (OLS), Partial

Least Squares (PLS), Principal Components Regression (PCR), Elastic Net (Enet), Decision Tress (DT), Gradient Boosted Regression Trees (GBRT), Random Forest (RF), Neural Networks ranging from one to five layers (NN1-NN5), Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This allows us to present a thorough and comparative evaluation of each model, offering a holistic view of the current landscape. Second, we extend the extant literature on machine learning in asset pricing by providing new evidence on the temporal variation on model performance. Dividing our stock dataset into six distinct time periods, spanning multiple decades, the temporal segmentation reveals the adaptability and resilience of certain models during periods of global upheaval, such as the pandemic, thus highlighting the need to consider external events and evolving data landscapes in predictive modeling. Third, we reveal that prediction accuracy in asset pricing is predominantly data-driven rather than model-driven, with significant trading activities potentially boosting the data's power to reflect market trends. Lastly, our approach expands upon the scope of macroeconomic analysis set out by Gu et al. (2020). While their study includes eight predictors, some of which, such as the dividend to price ratio, earnings to price ratio, and book to market ratio, may not traditionally fall under the umbrella of macroeconomic indicators. Our paper adopts a more expansive and authoritative set of 35 macroeconomic predictors, such as unemployment rate, consumer price index and federal funds rate. These are sourced directly from the Federal Reserve Economic Database, ensuring that our analysis is grounded in robust and widely recognized macroeconomic factors.

The remainder of the paper is organized as followings: Section 2 provides a review of related literature, while the methodology is detailed in Section 3. Section 4 delineates the results and discusses the main findings. Section 5 offers concluding remarks and suggests potential avenues for future research.

2. Literature Review

2.1. Machine learning algorithms in asset pricing

Conventionally, Linear Regression has been widely used thanks to its simplicity and ease of interpretation (Fama & French, 1992; Jegadeesh & Titman, 1993). However, its inability to capture complex and non-linear relationships remains a significant drawback, leading to the exploration of more advanced algorithms. Decision Trees emerge as an alternative, offering flexibility in

modeling non-linear relationships (Dumitrescu et al., 2022). However, Wolff and Neugebauer (2019) argue that they do not always offer better predictability than linear models. Random Forests, an extension of Decision Trees, mitigate overfitting and improve generalization (Breiman, 2001; Liaw & Wiener, 2002). Herrera et al. (2019) also contribute to the robustness of Random Forests, their results show that Random Forests significantly outperform traditional econometric models in forecasting energy commodity prices. Though, Liaw and Wiener (2002) find it less interpretable due to its ensemble nature, which complicates the explanation of the influence of individual predictors. Gradient Boosting further evolves from ensemble techniques, optimizing a loss function for more accurate predictions (Friedman, 2001; Natekin & Knoll, 2013). While Friedman (2001) emphasizes its predictive power, Natekin and Knoll (2013) point out that it is computationally intensive. Furthermore, Support Vector Machines (SVM) are particularly noted for their effectiveness in high-dimensional data spaces and classification tasks, although there are concerns about the choice of kernel affecting model performance (Cristianini & Shawe-Taylor, 2000).

In addition, Neural Networks (NNs) and the associated deep learning algorithms significantly improve the performance over traditional methods, establishing a new standard in the asset pricing domain (Chen et al., 2023; Gao et al., 2020). Bianchi et al. (2021) demonstrate NNs' superior ability to uncover and leverage predictive signals in the prediction of bond premia. Jiang et al. (2023) apply convolutional neural networks (CNNs) to analyze stock price charts, claiming that machine learning can identify price trends and patterns that traditional analysis methods might overlook. The adoption of Long Short-Term Memory (LSTM) networks further enhance the predictive accuracy, particularly for out-of-sample directional movements in major stock indices like the S&P 500 (Fischer & Krauss, 2018). Mehtab and Sen (2020) along with Chen and He (2018) have further substantiated the efficacy of Convolutional Neural Networks in enhancing prediction performance. Additionally, Harris et al. (2024) and Wang et al. (2023) further highlight the superiority of NNs and LSTM in cryptocurrency market, respectively.

These machine learning algorithms often build upon or complement each other, with Gradient Boosting and Random Forests both extending Decision Trees, and deep learning offering a more complex, layered approach. This logical progression in the literature not only highlights the

increasing complexity and sophistication of algorithms applied to asset pricing but also underscores the iterative nature of research in this domain, where each new methodology often builds upon the strengths and limitations of its predecessors, providing a more nuanced toolkit for asset pricing prediction and analysis. In light of the evolving landscape of machine learning algorithms for asset pricing, our paper aims to conduct a comprehensive comparative analysis of all the mainstream algorithms. This endeavor will help practitioners and researchers alike in selecting the most effective machine learning techniques for specific asset pricing challenges.

2.2. Predictive factors and performance of machine learning models

While extant literature has compared the performance of various machine learning algorithms in asset pricing, the results are rather inconclusive. The work of Gu et al. (2020) serve as a cornerstone in this area, employing a comprehensive comparison of various machine learning algorithms for asset pricing. They compare linear models, decision trees, random forests, and neural networks and find that tree-based models, particularly random forests, performed the best in terms of predictive accuracy and out-of-sample performance. This finding is consistent with the work of Krauss et al. (2017) who also claim that tree-based models perform better than the deep neural network in predicting stock returns.

Conversely, Nabipour et al. (2020) argue that LSTM shows more accurate results with the higher model fitting ability than tree-based models. Adding another dimension to the discourse, JingTao and Tan (2001) focus solely on Neural Networks, comparing their performance against traditional methods in predicting asset prices. They conclude that Neural Networks outperform other algorithms but caution that their effectiveness is contingent upon the availability of large datasets for training, echoing the sentiments of LeCun et al. (2015), who also advocate for Neural Networks but highlight the need for large datasets. Drobetz and Otto (2021) show that the Neural Networks outperform other machine learning algorithms even after accounting for transaction costs.

While machine learning offer advanced modeling techniques, the choice of predictive factors remains crucial for accurate asset pricing. Fama and French (1993) initially proposed three factors—market risk, size, and value—that have been widely used in traditional asset pricing models. However, with the advent of machine learning, the scope for incorporating predictive

factors has expanded significantly. Gu et al. (2020) identify a set of dominant predictive signals in all of their adopted machine learning models, which include variations on momentum, liquidity, and volatility in the U.S. market. However, according to Leippold et al. (2022), liquidity emerges as the most significant predictive factor in the Chinese market, followed closely by predictors related to fundamental factors such as valuation ratios. This observation contrasts with the findings of Gu et al. (2020) in their study on the U.S. market, where classical trend indicators emerge as the primary drivers of asset price predictability.

On the other hand, Drobetz and Otto (2021) argue that the most significant predictors are a combination of the results from both Gu et al. (2020) and Leippold et al. (2022). They find that recent price trends, such as short-term reversals and stock momentum, as well as fundamental indicators like earnings-to-price and book-to-market ratios, are the most impactful predictors. Given this, the effectiveness of these factors appears to be both context and algorithm-dependent, warranting further research to reconcile these varying perspectives.

3. Methodology

3.1 Dataset

Our study encompasses a dataset spanning from January 1970 to August 2023, totaling more than 53 years. The dataset comprises 5175 unique stocks, 758649 observations in total, with their distribution across different time periods summarized in Table 1.

Table 1. Summary of Sample number and stock number across years

Year	Sample number	Stock number
1970-1979	11896	150
1980-1989	49147	619
1990-1999	108097	1552
2000-2009	207205	2553
2010-2019	266787	3427
2020-2023	115517	3822
Total	758649	5175

For the study, we assemble a comprehensive set of 123 indicators, which include 69 financial ratios, 18 trading signals, and 35 macroeconomic indicators. The financial ratios are sourced from the Financial Ratios Firm Level dataset available in the WRDS (Beta) database, listed in Appendix A. We generate trading signals utilizing daily data sourced from Yahoo Finance, which encompasses

six fundamental metrics: open, high, low, close, adjusted close price, and volume. Based on these 6 primary metrics, we derive monthly return as our explained variable and 12 additional trading indicators (elaborated in Appendix B). The dataset is further enriched with 35 macroeconomic indicators, sourced from the Federal Reserve Economic Database (FRED). Among these, 30 indicators are updated on a monthly basis, while the remaining 5 are updated quarterly. We implement a backfilling data resampling methodology to transform quarterly data to monthly data. Furthermore, we manipulate selected variables, such as GDP and money supply into growth rates to eliminate the influence of temporal trends. Appendix C provides the details of these characteristics.

The monthly return can be represented by the following formula:

$$r_t = \frac{p_t}{p_{t-1}} - 1 \tag{1}$$

r_t is stock return in this month, p_t is close price for last day of current month, p_{t-1} is close price for last day of month. As the purpose of this study is to use a collection of indicators to predict monthly returns, there is a one-month lag in the dependent variable. For example, the indicators from Jan 2020 are used to predict the monthly return for Feb 2020.

3.2. Sample

The model utilized in this study incorporates machine learning algorithms for cross-sectional data, as well as deep learning algorithms, like LSTM and CNN, which are capable of capturing the temporal structure of samples. As a result, it is necessary to concatenate the cross-sectional data into time-series format.

We first conduct data merging and resampling. The cross-sectional data consists of a merger between trading data, financial metrics, and macroeconomic indicators. The merger is conducted based on stock codes and dates as the merging indices. Trading data are downsampled from daily to monthly frequencies, using data from the last trading day of each month. For quarterly macroeconomic variables such as GDP and net exports, a backfilling resampling technique is employed. For instance, January's data is used to fill in missing values for February and March, and April's data is used for May and June, and so on.

In the calculation process, instances of infinity and negative infinity, as well as other missing values, are filled using the mean of the respective variable. Monthly returns exceeding 10 (1000%) and falling below -10 (-1000%) are considered outliers and are removed from the dataset. Additionally, the IQR method is applied to remove samples in the predictors that are beyond the upper and lower bounds. This study employs z-score normalization. Each stock's mean and standard deviation are calculated for each factor separately. The transformed factor is given by:

$$x'_{ikt} = \frac{x_{ikt} - \mu_{ik}}{\sigma_{ik}} \tag{2}$$

Here, i represents the stock, k represents the factor, and t represents the month. The target variable, monthly return, is not subject to normalization.

Time-series data is generated by concatenating cross-sectional data in a sequential manner (as illustrated in Figure 1). The predictors and targets from one cross-sectional dataset are combined into a single row, which is then vertically concatenated with corresponding rows from different time points to form the time-series predictors. The next time point's cross-sectional data serves as an integrated whole, constituting the time-series prediction target.

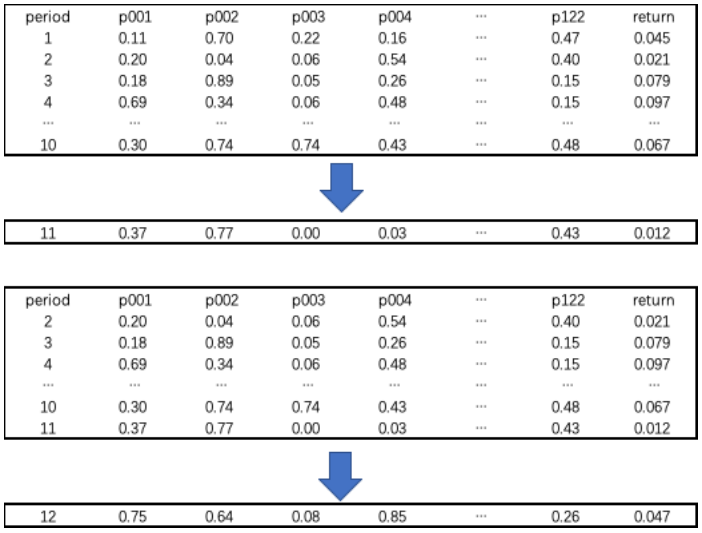

Figure 1. Construction of LSTM and CNN dataset

To elaborate, 122 indicators of each stock are taken into account every month, serving as predictive factors X. Additionally, the stock returns of the subsequent month are used as the output variable y. This forms a sample with a length of 122+1=123. Across 5175 stocks, this results in a total of 708,468 sample pairs. The dataset is then randomly divided, with 80% of the samples designated

for training and the remaining 20% reserved for testing. This method ensures a thorough and robust training process, while also preserving a significant portion of the data for model evaluation and validation. Assuming a time step of 10 (see Figure 1), for a particular stock, the first 10 cross-sectional samples are concatenated to form X_1, and the 11th cross-sectional sample is used as y_1. Similarly, the cross-sectional samples from 2 to 11 are concatenated to form X_2, and the 12th sample is used as y_2, and so on, until all the cross-sectional samples of a particular stock are traversed. The same procedure is then applied to the other stocks. Ultimately, the dimension of X is (708,468, 10, 123), and the dimension of y is (708,468, 1, 123).

3.3. Measurement

3.3.1 Measurement of model performance

Throughout our study, we employ a general additive prediction error model to characterize the relationship between a stock's return and its associated predictors, represented as:

$$r_{i,t+1} = E_t[r_{i,t+1}] + \epsilon_{i,t+1} \tag{3}$$

We also presume that the conditional expectation of a stock's return, given the data available at time t, is a constant function of a set of predictors:

$$E_t[r_{i,t+1}] = g(z_{i,t}) \tag{4}$$

where $\mathbf{z_{i,t}}$ is a P-dimensional predictor vector. Stocks are denoted by i = 1,…, N_t and months by t = 1,…, T. The functional form *g(·)* is intentionally left undefined. Our aim is to identify the best-performing prediction model from a list of potential candidates.

The predictor vector $\mathbf{z_{i,t}}$ incorporates a series of dummy variables and can be formulated as:

$$z_{i,t} = [c_{i,t} \quad x_t \quad d_{i,t}] \tag{5}$$

In this equation, $c_{i,t}$ is a 1×69 vector of stock-level characteristics, x_t is a 1×35 vector of macroeconomic indicators, $d_{i,t}$ is a 1×18 vector of trading signals. Hence, the total number of covariates in $z_{i,t}$ is 1×122.

Following the approach of Gu et al. (2020), we utilize the non-demeaned out-of-sample predictive R^2 for a straightforward model-to-model comparison:

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in T}\left(r_{i,t+1} - \widehat{r_{i,t+1}}\right)^2}{\sum_{(i,t)\in T} r^2_{i,t+1}} \tag{6}$$

Here, T represents the testing subset that is isolated from the data used for model estimation or tuning. The R^2_{oos} value offers a comprehensive, panel-level gauge of each model's predictive efficacy by aggregating prediction errors across multiple firms and time periods.

One notable feature of our R^2 metric is that it doesn't demean the sum of squared excess returns in the denominator. In many forecasting contexts, predictions are typically benchmarked against historical average returns. While this method may be appropriate for assessing aggregate indices or long-short portfolios, it is not reliable for evaluating individual stock returns. Using historical averages as a predictive basis usually results in substantially poorer performance compared to a simple zero forecast. This is because the historical mean return for individual stocks introduces too much noise, thus lowering the standard for what is considered 'good' forecasting. To sidestep this issue, we set our R^2 benchmark against a forecast value of zero.

3.3.2 Measurement of importance of predictor

The feature importance in Random Forest is calculated based on Gini Impurity, a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the set.

The Gini Impurity of a node is calculated as:

$$Gini = 1 - \sum p_i^2 \tag{8}$$

where p_i is the probability of choosing an element of class i at the current node.

The importance of a feature is then calculated as the total decrease in Gini Impurity brought about by splits over that feature, averaged over all trees in the forest. More formally, the importance of feature f can be computed as:

$$Importance(f) = \frac{\sum_{t=1}^{T} \sum_{n \in node_t} \varpi_{tn} \Delta Gini_{tn}}{\sum_{t=1}^{T} \sum_{n \in node_t} \varpi_{tn}} \tag{9}$$

T is the number of trees in the forest, $node_t$ is the set of nodes in tree *t*, ϖ_{tn} is the proportion of samples reaching node *n* in tree *t, and* $\Delta Gini_{tn}$ is the decrease in Gini Impurity brought about by the split at node *n* in tree *t*. This yields a score for each feature, with a higher score indicating greater importance of the feature towards the prediction task.

4. Result

4.1. Model Performance across diverse models

We conduct a comparative analysis for out-of-sample predictive performance across all 14 models based on their R^2 values, the Random Forest model emerges as the top performer with the highest R^2 value of 0.1251. This suggests that it is the most adept model at predicting the outcome variable among all models considered. Close contenders include Neural Network 3 and Gradient Boosting, which deliver R^2 values of 0.1200 and 0.0966, respectively. Although they are strong performers, they do not surpass the Random Forest model. On the other hand, traditional statistical methods such as Linear Regression and PCR yield moderate predictive power with an R^2 value of 0.0556, but they are outperformed by more complex algorithms.

Notably, models like the Decision Tree, Elastic Net, and Neural Network 1 either performed poorly or failed to offer predictive power better than a naive model that predicts the mean outcome. Their R^2 values range from -0.9315 for Decision Tree to 0.0038 for Elastic Net. Additionally, despite their capability to capture temporal structures, deep learning models like LSTM and Conv1D performed disappointingly, offering R^2 values close to zero (0.0082 and 0.0075, respectively). Therefore, based on the R^2 performance metric, the Random Forest model would be the most suitable choice for this specific problem based on the available data.

Table 2. Comparison of Model Performance

Model	R^2
Linear Regression	0.0556
PLS	0.0184
PCR	0.0556
Elastic Net	0.0038
Decision Tree	-0.9315
Random Forest	0.1251
Gradient Boosting	0.0966
Neural Network 1	-0.1016
Neural Network 2	0.0952
Neural Network 3	0.1200
Neural Network 4	0.1146
Neural Network 5	0.1134
LSTM	0.0082
Conv1D	0.75

4.2. The dependency of Model Performance on Temporal Variation

We segment our dataset into six unique time frames. The initial five periods each covers a decade, while the period from 2020 to 2023 is treated as a separate entity due to the exceptional circumstances it encompasses, specifically the COVID-19 pandemic and the onset of a significant regional conflict. Table 3 and Figure 2 illustrate how the models perform across these varying years, using the R^2 value as the performance metric.

Table 3. Comparison of Model Performance across years

Model	1970-1979	1980-1989	1990-1999	2000-2009	2010-2019	2020-2023
Linear Regression	0.0579	0.0447	0.0276	0.0210	0.0417	0.1389
PLS	0.0208	0.0259	0.0133	0.0106	0.0105	0.0292
PCR	0.0579	0.0447	0.0260	0.0210	0.0417	0.1389
Elastic Net	0.0098	0.0154	0.0082	0.0031	0.0063	0.0021
Decision Tree	-1.0012	-0.7641	-1.1336	-0.9694	-0.8917	-0.8373
Random Forest	0.2131	0.1351	-0.0306	0.0621	0.0636	0.1398
Gradient Boosting	0.1789	0.1390	0.0270	0.0614	0.0656	0.1534
Neural Network 1	-0.1841	-0.1179	0.0086	-0.0651	-0.7850	-0.0217
Neural Network 2	-0.0618	0.0676	0.0319	0.0541	0.0053	0.1119
Neural Network 3	0.0187	0.0880	0.0419	0.0596	0.0560	0.1542
Neural Network 4	0.0402	0.0887	0.0477	0.0625	0.0721	0.1618
Neural Network 5	0.0344	0.0670	0.0413	0.0630	0.0646	0.1575
LSTM	0.0357	0.0274	0.0167	0.0201	0.0190	0.0642
Conv1D	0.0393	0.0192	0.0124	-0.0008	0.0120	0.0312

In assessing the R^2 values of various models over different time periods, it becomes evident that certain models demonstrate more consistent performance, while others exhibit greater variability. Linear Regression, PCR, Neural Network 3, 4, and 5, Conv1D (CNN), and LSTM have all shown enhanced predictive capabilities during the 2020-2023 period. Notably, Linear Regression and PCR, with R^2 values ranging from 0.02 to 0.14 in earlier years, have seen a significant boost, with values both reaching up to 0.1389 during the pandemic. This increase, particularly during a period of global upheaval, indicates their adaptability and resilience to sudden changes in data patterns. Similarly, Neural Networks 3 to 5 and LSTM have also displayed their peak performances during this period, suggesting their adeptness in capturing both linear and non-linear data relationships intensified by the pandemic's impact. In contrast, PLS has consistently maintained relatively stable

but low R^2 values across all years, such as 0.0259 in 1980-1989 and 0.0292 in 2020-2023, indicating its consistent yet limited predictive power. On the other end, the Decision Tree model's consistently negative R^2 values, dropping as low as -1.1336 in the 1990s, point to potential issues like overfitting or an inability to grasp genuine data patterns during the examined periods.

The variability observed in models like Random Forest and Gradient Boosting suggests their sensitivity to evolving data landscapes. For instance, the R^2 values of Random Forest swung from -0.0306 in the 1990s to 0.2131 in the 1970s, showcasing its fluctuating performance across years. Such performance variations could be attributed to the inherent characteristics of the models. Simple linear models, like Linear Regression, make fewer assumptions about the data and are less susceptible to overfitting, which can result in consistent results. In contrast, more complex models, with their ability to capture intricate data patterns, might be sensitive to changes in data distributions or underlying relationships.

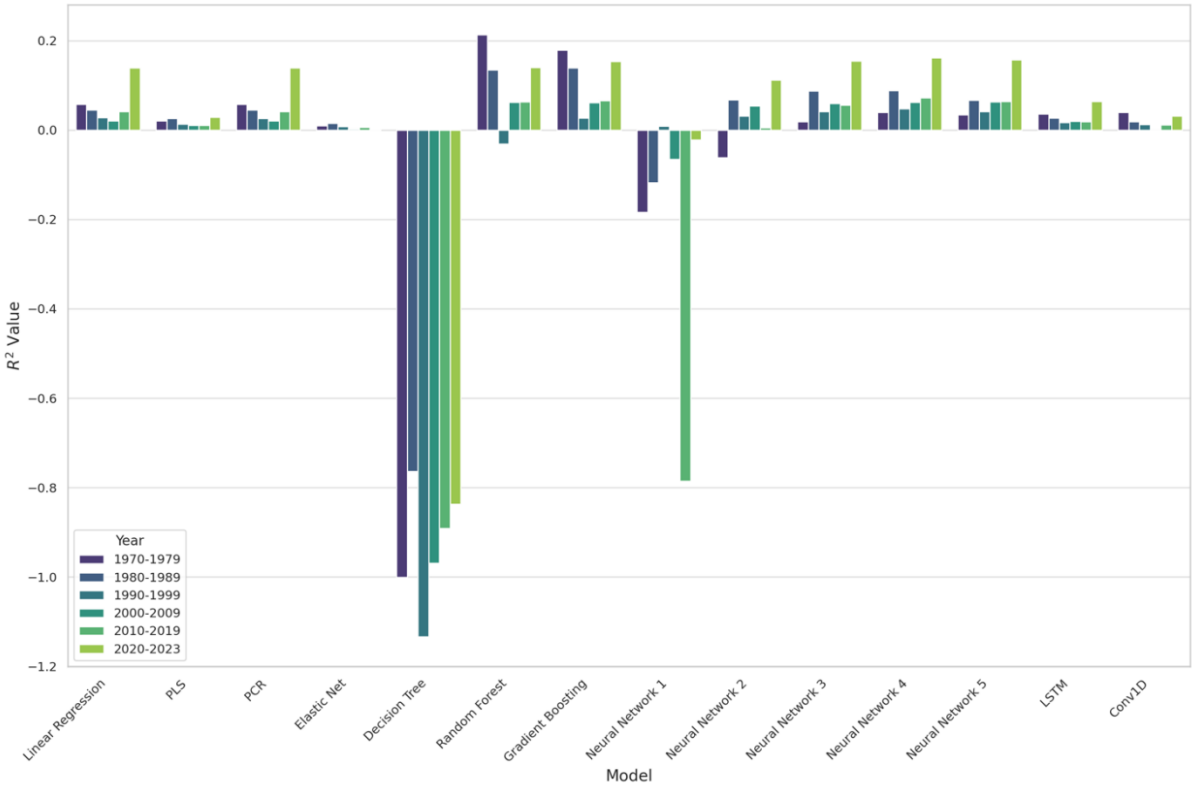

Figure 2. Temporal Model Performance

From a future research perspective, delving deeper into the specific data traits from different periods and understanding how external events influence model performance is crucial. The pandemic era, in particular, has emphasized the importance of developing models resilient to abrupt changes. Hybrid models, combining the strengths of both simple and complex structures, might be a promising avenue. Additionally, with the rise of interpretable machine learning, exploring model decisions can provide deeper insights, guiding further model refinements and innovations.

4.3. Importance of Financial Characteristics in Explaining Returns

Figure 3 provides the overall ranking of predictors by their importance. The spotlight falls primarily on three salient categories within the top 15 influential factors: 'Valuation Ratios,' 'Liquidity,' and 'Price Trend.'

Taking the lead is the 'Valuation Ratios' category, comprising 5 key factors out of the top 15 factors, including price to cash flow (pcf), price to book (ptb), price to sales (ps), book to market (bm), and enterprise value multiple (evm). These ratios are vital in assessing how the market values a company in relation to its financial standing. Coming in second, the ‘Liquidity’ category includes volatility of liquidity based on both dollar trading volume for the past 21 days (std_dolvol) and share turnover for the past 21 days (std_turn), Trading volume in a trading day (Vloume) and Dollar trading volume (dolvol). The third tier 'Price Trend' category includes influential factors like maximum daily return (maxret) the 36-month (mom36m) and 1-month momentum (mom1m), offering important insights into the directional movement of stock prices over both short and long terms. Notably, Chicago Fed National Financial Conditions Index (NFCI) emerges as the sole macroeconomic indicator within the top 15 factors. Its unique inclusion signifies an imperative need for the integration of macroeconomic conditions in asset pricing models, underscoring the multifactorial and dynamic nature of financial markets.

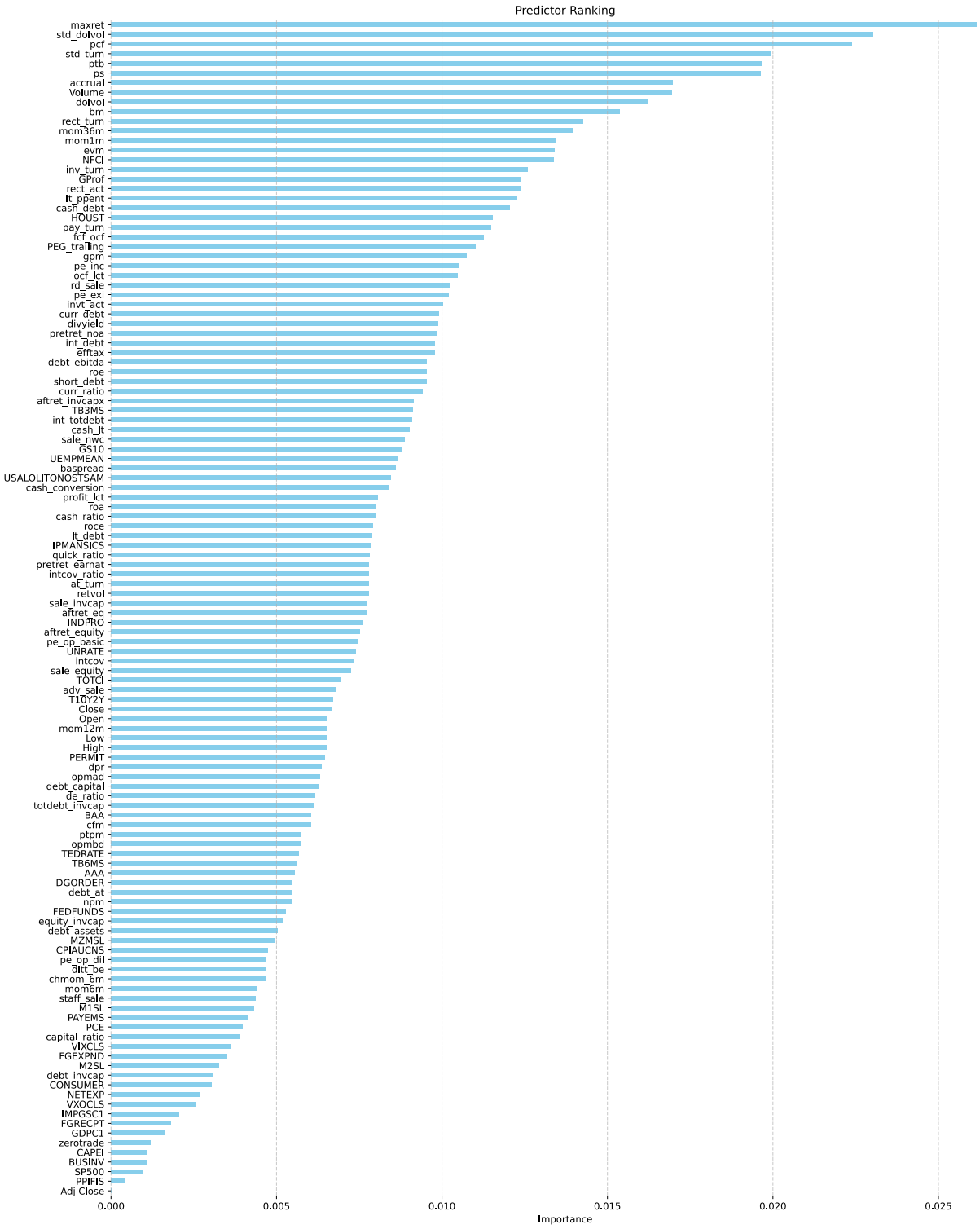

Figure 3. Overall ranking of Predictors by Importance

Figure 4 furnishes a temporally segmented analysis of factor importance, the most influential characteristics at the top and the least influential at the bottom. Each column, representing a five-year period, provides separate statistics for each factor. Within a column, a darker color indicates a larger contribution of that factor during the period, while a lighter color indicates a smaller contribution. A salient observation emanates from this longitudinal evaluation: factors that exhibit robust performance on an aggregate level do not necessarily maintain that vigor uniformly across all time periods. For instance, 'Maximum Daily Return' (MaxRet), which tops the overall ranking, manifests a pronounced influence exclusively during two non-contiguous decades: 1995-2004 and 2010-2019. Similarly, 'Price to Cash Flow' (PCF), another preeminent metric ranking third overall, exerts an exceptionally high degree of influence solely during the 1990-1994 quinquennium. Adding another layer of nuance is the noteworthy performance of the 'Chicago Fed National Financial Conditions Index' (NFCI), the only macroeconomic indicator to secure a position among the top 15 factors. Its predictive prowess stands out particularly during the decade spanning 1970-1979.

To conclude, the fluctuating significance of various factors across different time periods prompts the question of what drives these changes. It may be attributable to structural shifts in the market, policy changes, or even broader economic cycles. This forms a fertile ground for future research. Investigating the underlying mechanisms that account for the time-varying relevance of these factors could provide invaluable insights into more adaptive and resilient asset pricing models. Furthermore, the apparent influence of macroeconomic indicators, like the NFCI, even if sporadic, suggests that integrating such elements into multi-factor models could be a promising avenue for future inquiries.

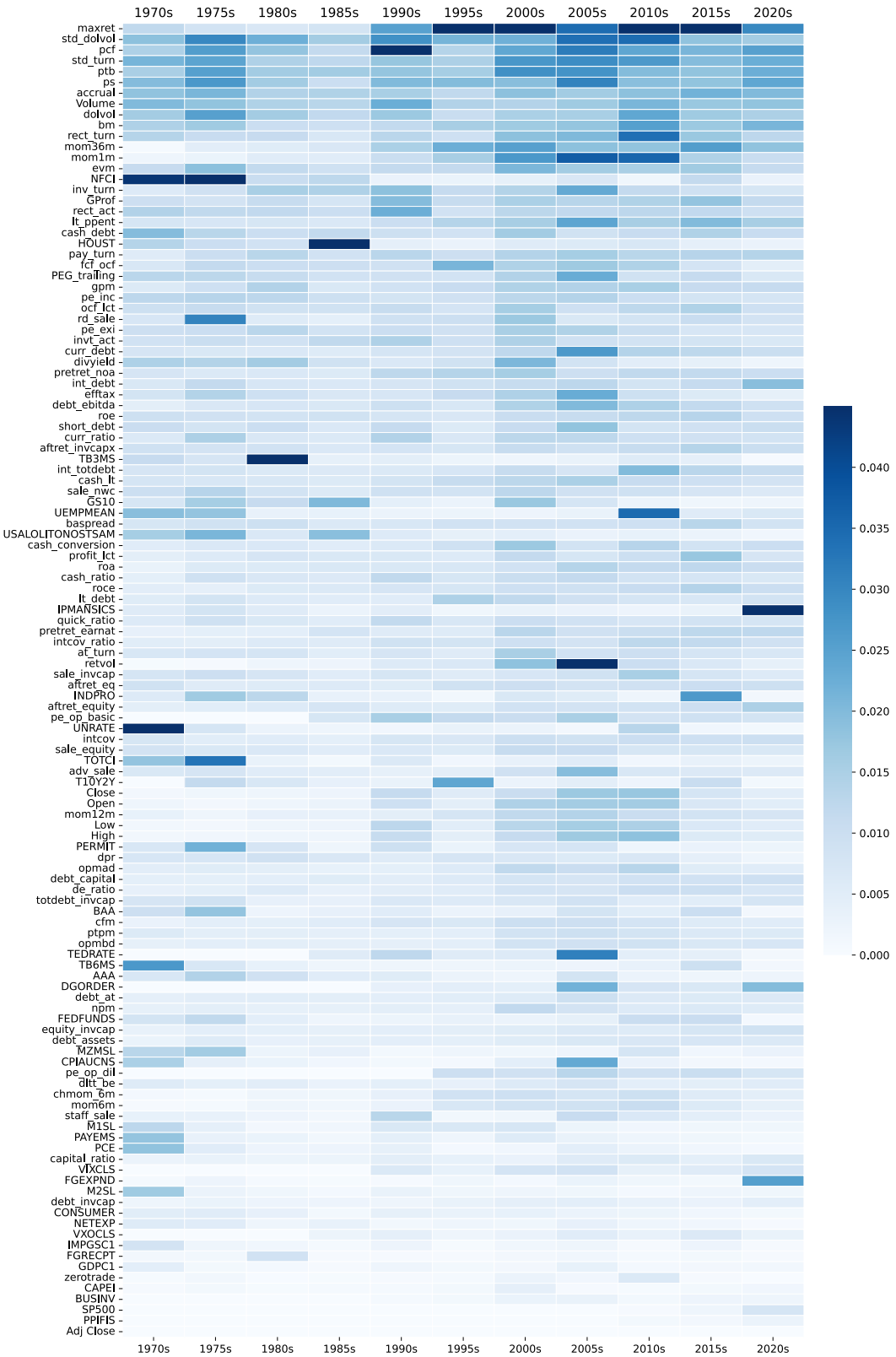

Figure 4. Characteristic importance

4.4. Marginal association between characteristics and expected returns

Figure 5 elucidates the marginal influence of specific characteristics on expected asset returns across 12 models. Given that the deep learning models in this study utilize non-cross-sectional data, they are deemed inappropriate for factor analysis. Consequently, the CNN and LSTM models are excluded from this section. In this visualization, we have normalized these characteristics to fall within the (-1,1) range while maintaining all other variables at their median value, effectively set to zero.

We have selected the eleven most salient characteristics for illustration, in addition to the sole macroeconomic indicator—Chicago Fed National Financial Conditions Index (NFCI)—that ranks among the top 15 factors. The eleven characteristics selected in this section encompass maximum daily return (maxret), two distinct measures of liquidity volatility based on 21-day dollar trading volume (std_dolvol) and share turnover (std_turn), valuation ratios including price to cash flow (pcf), price to book (ptb), and price to sales (ps), accruals to average assets (accrual), daily trading volume (Volume), dollar trading volume (dolvol), book to market (bm), and a 36-month momentum indicator (mom36m).

First, our findings indicate that linear models, including PLS, and Elastic Net, struggle to capture the marginal association between all these characteristics and expected returns. In essence, they present an almost exact zero relationship. Furthermore, Neural Networks 2 through 5 depict patterns that resonate with some established empirical phenomena. A case in point is the inverse correlation of expected stock returns with volatility, specifically liquidity volatility based on a 21-day dollar trading volume (std_dolvol). For instance, a firm experiencing a volatility of liquidity rise from the 20th percentile to the median sees an approximate annual return decline of 4.8% (0.4%×12). Neural Networks 2-5 also showcase similar trends, displaying a negative marginal association between dollar trading volume (dolvol) and expected returns. Yet, when it comes to the marginal association between expected returns and Chicago Fed National Financial Conditions Index (NFCI), a clear divergence in model performance emerges. For instance, both Neural Network 5 and PCR consistently reveal a positive correlation, implying that returns ascend with NFCI. Neural Network 3 presents an intriguing pattern: an initial positive relation that transitions to a zero relation upon reaching the median. Contrarily, Neural Network 4 starts with a zero

relationship, shifting to a positive correlation post-median. One potential reason for this could be the underlying model architecture of Neural Networks, where certain layers or nodes might be more sensitive to specific data patterns or ranges, leading to these shifts.

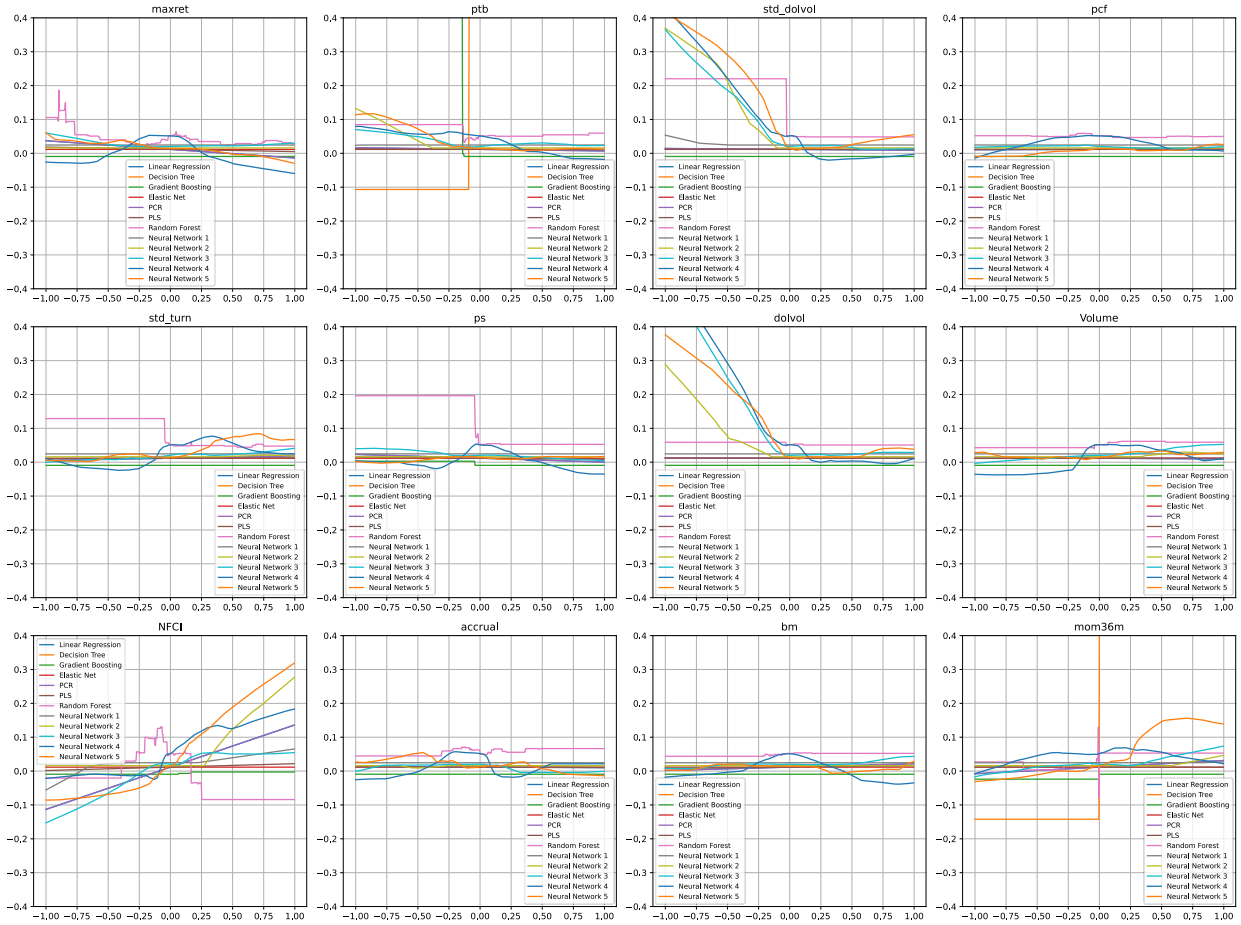

Figure 5. Marginal association between expected returns and characteristics

The panels show the sensitivity of expected monthly percentage returns (vertical axis) to the individual characteristics (holding all other covariates fixed at their median values).

Interestingly, our analysis found that the Decision Tree model exhibits extreme sensitivity at the median of the price to book (ptb) and 36-month momentum indicator (mom36m) characteristics. One plausible explanation might tie back to our earlier discussions regarding Decision Trees. Their inherent nature to form splits based on feature values might cause them to be particularly sensitive to median values of certain characteristics, especially if those values represent significant decision boundaries in the data. The previously observed inferior predictability of the Decision Tree might be intertwined with this sensitivity, as overfitting to specific feature boundaries can lead to poor generalization and predictive performance on new or unseen data.

It's noteworthy to highlight the behavior of the Random Forest model. The model exhibits a pronounced shift in its predictions around the median of specific characteristics, including price to book (ptb), liquidity volatility based on a 21-day dollar trading volume (std_dolvol) share turnover (std_turn), and price to sales (ps). This distinct shift in the marginal association between expected returns and these characteristics sheds light on the model's nuanced comprehension of the underlying data dynamics. This non-linear association signifies that as certain characteristics change, even slightly, the expected returns adjust in a notably different manner before and after the median threshold. Random Forests, by their very nature, excel at discerning non-linear relationships in data. The pronounced shift in predicted returns at the median implies that the model identifies the median as a critical pivot point. This might suggest that for values of the characteristic below this median, the marginal impact on returns is consistent and follows a specific pattern. However, upon crossing this median, the marginal impact on returns sees a significant transformation, reflecting another consistent pattern, albeit different from the first. The model's consistency in its predictions, both before and after this median threshold, underscores its confidence. It suggests that within these two brackets, the marginal effects of other characteristics on expected returns remain relatively stable. Such behavior not only amplifies the Random Forest's robust predictive capabilities but also emphasizes its adeptness in identifying pivotal transitions in the marginal associations between returns and characteristics.

While other models might not detect or might even downplay such transitions, the Random Forest's nuanced understanding of these shifts highlights its superior capability to gauge the marginal effects of characteristics on expected returns. In the grand scheme of financial analysis, recognizing these non-linear marginal associations can be pivotal. Such insights can guide investors to better understand how slight changes in specific characteristics can lead to significantly different expected returns, especially when these characteristics hover around critical thresholds like the median.

4.5. Data Driven vs. Model Driven

In this section, our objective is to explore whether the accuracy of predictions in asset pricing is primarily driven by the data or by the models. We implement an annual cycle of training and

testing on our asset pricing models, using the respective data from each year to assess the performance of various models. Figure 6 displays the training years along the horizontal axis and the testing years along the vertical axis. Predictive performance along the diagonal is expectedly superior since the training and testing datasets are identical. An interesting observation from Figure 6 is that between 1993 and 2016, predictive performance is substantially driven by the data, as demonstrated by the consistently strong results from all models across every training year.

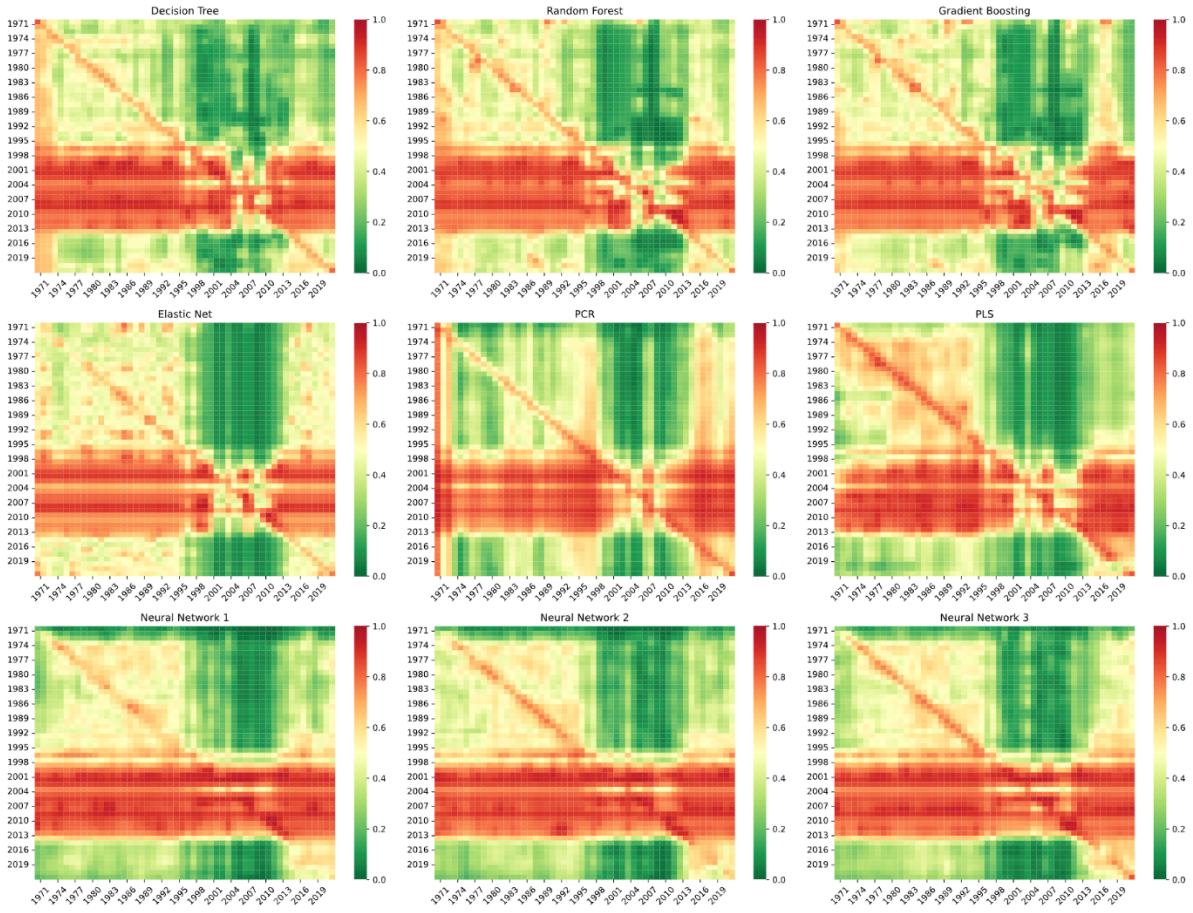

Figure 6. Model Performance of Data year vs Model year

To understand the drivers of this trend, we investigate potential structural changes that may have occurred during this time. Our objective is to identify whether there have been shifts in market mechanisms or external economic factors that could have impacted the performance of asset pricing models in the periods before or after this interval. In Figure 7, we note that financial characteristics are the main contributors to predictive accuracy before 1993 and after 2016. However, in the span from 1993 to 2016, the relative impact of financial and trading factors is

more variable, indicating that robust trading factors can enhance the data's ability to explain market movements.

Thus, the analysis presented in this section suggests that prediction accuracy in asset pricing is predominantly data-driven rather than model-driven, with significant trading activities potentially boosting the data's power to reflect market trends.

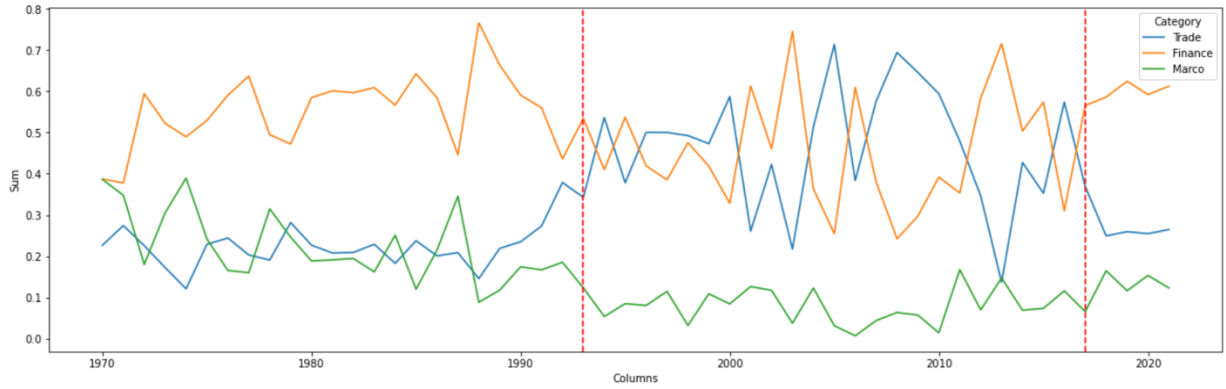

Figure 7. Total Contribution of Trade, Finance and Macro Indicators Across Years

4.6. Portfolio Performance Across Models

In a pursuit to understand the investment efficacy of various predictive models, we embarked on an empirical study to assess their ability to inform profitable stock portfolio strategies, as shown in Figure 8. For the period spanning 2003 to 2021, we constructed model-driven portfolios wherein stocks forecasted to yield positive returns each month were equally weighted and incorporated into the respective model's portfolio. The monthly return for a given model was thus derived as the mean return of its selected stocks, with portfolio constituents recalibrated monthly based on the latest predictions.

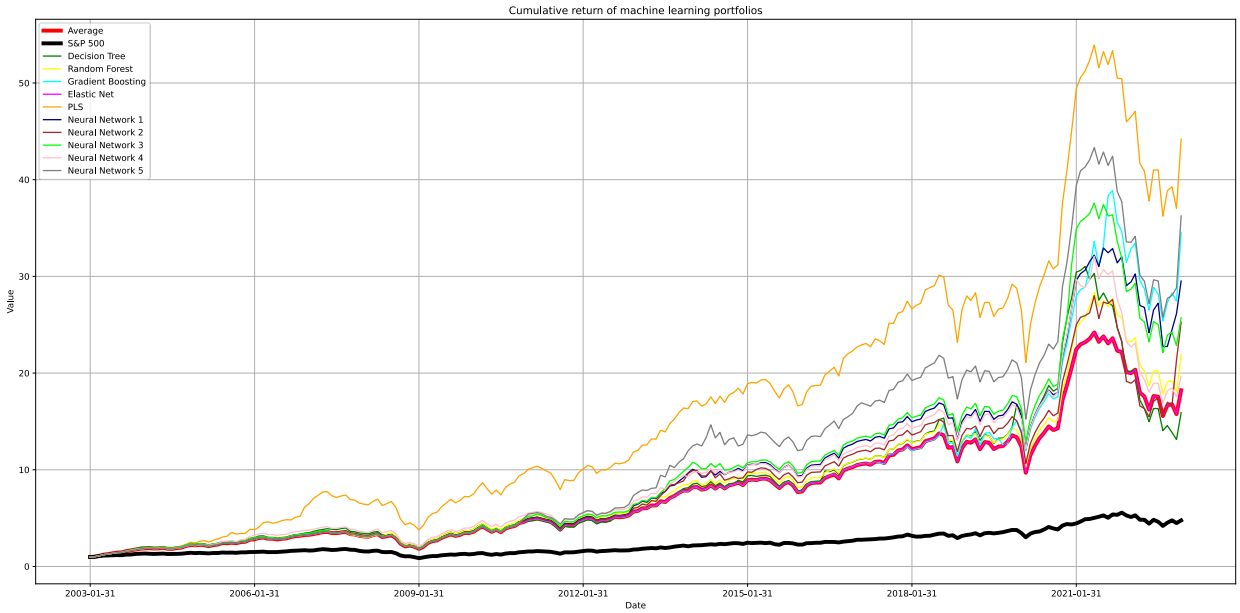

Figure 8. Model driven portfolio performance

A salient observation from our analysis was the remarkable trend consistency across all models, underscoring the potential universal underlying market dynamics captured by these predictors. Surprisingly, the PLS model emerged as a stellar performer; the portfolio steered by its predictions consistently outperformed those from all other models, including sophisticated machine learning algorithms. Another model deserving notable mention for its robust performance was Neural Network 5. Intriguingly, every model-driven portfolio surpassed the S&P 500 index returns, a conventional market benchmark. To further refine our comparative lens, we introduced a secondary benchmark -the average return of portfolios constructed by all models. Here too, the majority of model-driven portfolios outpaced this benchmark, with a singular exception being the Decision Tree model, which lagged behind during the concluding periods of our observation window.

5. Conclusion

Our findings illuminate a complex performance landscape where Random Forest and specific Neural Networks like Neural Network 3 and Neural Network 5 emerge as top performers based on the R^2. These models not only outperform traditional statistical models like Linear Regression but also provide valuable insights into their applicability in predicting asset returns. One of the most intriguing observations is the temporal dependency in model performance. Models like

Linear Regression and PCR, traditionally considered less powerful, demonstrated notable resilience and adaptability during turbulent market conditions, such as the pandemic era. This underscores the importance of considering the temporal dynamics when selecting models for financial prediction tasks. Our study also contributes to the ground by identifying and ranking the characteristics that significantly influence asset returns. The most influential factors fall primarily within the categories of 'Valuation Ratios,' 'Liquidity,' and 'Price Trend'. Chicago Fed National Financial Conditions Index (NFCI) emerges as the sole macroeconomic factors in the top 15 predictive factors also draws our attention in an imperative need for the integration of macroeconomic conditions in asset pricing models. Moreover, we find that prediction accuracy in asset pricing is primarily influenced by the data rather than the models, with substantial trading activities potentially enhancing the data's capacity to mirror market trends. Lastly, the real-world applicability of our research is validated through the construction of model-driven portfolios.

In addition, the findings of our study carry profound implications across the board for investors, managers, and policymakers. For investors, leveraging sophisticated models like Random Forest and Neural Networks can offer a competitive edge in predicting asset returns, highlighting the importance of valuation ratios, liquidity, and price trends. Managers, on the other hand, must cultivate agility and an adaptive toolkit to navigate the ever-changing market dynamics, as highlighted by the resilience of certain models during turbulent times like the pandemic. This adaptability is crucial, not just in model selection but in understanding the broader market and economic indicators such as the Chicago Fed National Financial Conditions Index (NFCI). For policymakers, the integration of macroeconomic conditions into financial analysis and the predictive power of such factors suggest a need for policies that support economic stability and innovation in financial technologies. Ensuring transparent, stable economic conditions can enhance the accuracy and robustness of financial predictions, benefiting the entire financial ecosystem. Thus, the practical application of our research, evidenced by the success of model-driven portfolios in outperforming traditional benchmarks, signifies the collective move towards a more informed, data-driven approach in the financial sector, emphasizing the synergy between technological advancement, economic policy, and investment strategy.

Our study highlights the significant role of data quality over the choice of models in determining the accuracy of asset pricing predictions. However, it falls short of investigating the root causes behind the fluctuations in data quality and the structural market changes identified within specific timeframes. We pinpointed when these changes occurred but not why the data might have been compromised or what exactly triggered these market shifts. Moreover, the direct impact of such changes on the predictive performance of asset pricing models was not fully explored. Future research should aim to address these gaps by examining the specific events or conditions leading to variations in data quality and structural market changes. This could involve analyzing the impact of regulatory adjustments, economic milestones, or shifts in trading behaviors to understand their correlation with the periods of change we've identified. Additionally, there's a need for developing asset pricing models that can better accommodate these data and market dynamics. Future efforts could focus on creating models that automatically adjust to evolving data patterns or enhancing model robustness to maintain performance despite data volatility.

References:

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1046-1089.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chen, L., Pelger, M., & Zhu, J. (2023). Deep learning in asset pricing. *Management Science*.

Chen, S., & He, H. (2018). Stock prediction using convolutional neural network. IOP Conference series: materials science and engineering,

Chordia, T., Subrahmanyam, A., & Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, *58*(1), 41-58.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Drobetz, W., & Otto, T. (2021). Empirical asset pricing via machine learning: evidence from the European stock market. *Journal of Asset Management*, *22*, 507-538.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European journal of operational research*, *297*(3), 1178-1192.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, *47*(2), 427-465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *the Journal of Finance*, *75*(3), 1327-1370.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, *270*(2), 654-669.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326-2377.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gao, P., Zhang, R., & Yang, X. (2020). The application of stock index price prediction with neural network. *Mathematical and Computational Applications*, *25*(3), 53.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.

Gunnarsson, E. S., Isern, H. R., Kaloudis, A., Risstad, M., Vigdel, B., & Westgaard, S. (2024). Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *International Review of Financial Analysis*, 103221.

Harris, R. D., Mazibas, M., & Rambaccussing, D. (2024). Bitcoin replication using machine learning. *International Review of Financial Analysis*, 103207.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5-68.

Herrera, G. P., Constantino, M., Tabak, B. M., Pistori, H., Su, J.-J., & Naranpanawa, A. (2019). Long-term forecast of energy commodities price using machine learning. *Energy*, *179*, 214-221.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, *33*(5), 2019-2133.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *the Journal of Finance*, *48*(1), 65-91.

JingTao, Y., & Tan, C. L. (2001). Guidelines for financial forecasting with neural networks. *Neural Inf. Process. Shanghai*.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European journal of operational research*, *259*(2), 689-702.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of financial economics*, *145*(2), 64-82.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *the Journal of Finance*, *71*(1), 5-32.

Mehtab, S., & Sen, J. (2020). Stock price prediction using convolutional neural networks on a multivariate timeseries. *arXiv preprint arXiv:2001.09769*.

Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., & Salwana, E. (2020). Deep learning for stock market prediction. *Entropy*, *22*(8), 840.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Wang, Y., Andreeva, G., & Martin-Barragan, B. (2023). Machine learning approaches to forecasting cryptocurrency volatility: Considering internal and external determinants. *International Review of Financial Analysis*, *90*, 102914.

Wolff, D., & Neugebauer, U. (2019). Tree-based machine learning approaches for equity market predictions. *Journal of Asset Management*, *20*(4), 273-288.

Appendix A. Financial Ratios

Table A1. Valuation Ratios

Acronym	Firm Characteristic
dpr	Dividend Payout Ratio
PEG_trailing	Trailing P/E to Growth (PEG) ratio
bm	Book/Market
capei	Shillers Cyclically Adjusted P/E Ratio
divyield	Dividend Yield
evm	Enterprise Value Multiple
pcf	Price/Cash flow
pe_exi	P/E (Diluted, Excl. EI)
pe_inc	P/E (Diluted, Incl. EI)
pe_op_basic	Price/Operating Earnings (Basic, Excl. EI)
pe_op_dil	Price/Operating Earnings (Diluted, Excl. EI)
ps	Price/Sales
ptb	Price/Book

Table A2. Profitability Ratios

Acronym	Firm Characteristic
efftax	Effective Tax Rate
gprof	Gross Profit/Total Assets
aftret_eq	After-tax Return on Average Common Equity
aftret_equity	After-tax Return on Total Stockholders Equity
aftret_invcapx	After-tax Return on Invested Capital
gpm	Gross Profit Margin
npm	Net Profit Margin
opmad	Operating Profit Margin After Depreciation
opmbd	Operating Profit Margin Before Depreciation
pretret_earnat	Pre-tax Return on Total Earning Assets
pretret_noa	Pre-tax return on Net Operating Assets
ptpm	Pre-tax Profit Margin
roa	Return on Assets
roce	Return on Capital Employed

roe	Return on Equity

Table A3. Capitalization Ratios

Acronym	**Firm Characteristic**
capital_ratio	Capitalization Ratio
equity_invcap	Common Equity/Invested Capital
debt_invcap	Long-term Debt/Invested Capital
totdebt_invcap	Total Debt/Invested Capital

Table A4. Financial Soundness

Acronym	**Firm Characteristic**
invt_act	Inventory/Current Assets
rect_act	Receivables/Current Assets
fcf_ocf	Free Cash Flow/Operating Cash Flow
ocf_lct	Operating CF/Current Liabilities
cash_debt	Cash Flow/Total Debt
cash_lt	Cash Balance/Total Liabilities
cfm	Cash Flow Margin
short_debt	Short-Term Debt/Total Debt
profit_lct	Profit Before Depreciation/Current Liabilities
curr_debt	Current Liabilities/Total Liabilities
debt_ebitda	Total Debt/EBITDA
dltt_be	Long-term Debt/Book Equity
int_debt	Interest/Average Long-term Debt
int_totdebt	Interest/Average Total Debt
lt_debt	Long-term Debt/Total Liabilities
lt_ppent	Total Liabilities/Total Tangible Assets

Table A5. Solvency Ratios

Acronym	**Firm Characteristic**
de_ratio	Total Debt/Equity

debt_assets	Total Debt/Total Assets
debt_at	Total Debt/Total Assets
debt_capital	Total Debt/Capital
intcov	After-tax Interest Coverage
intcov_ratio	Interest Coverage Ratio

Table A6. Liquidity Ratios

Acronym	**Firm Characteristic**
cash_conversion	Cash Conversion Cycle (Days)
cash_ratio	Cash Ratio
curr_ratio	Current Ratio
quick_ratio	Quick Ratio (Acid Test)

Table A7. Efficiency Ratios

Acronym	**Firm Characteristic**
at_turn	Asset Turnover
inv_turn	Inventory Turnover
pay_turn	Payables Turnover
rect_turn	Receivables Turnover
sale_equity	Sales/Stockholders Equity
sale_invcap	Sales/Invested Capital
sale_nwc	Sales/Working Capital

Table A8. Other Ratios

Acronym	**Firm Characteristic**
accrual	Accruals/Average Assets
rd_sale	Research and Development/Sales
adv_sale	Advertising Expenses/Sales
staff_sale	Labor Expenses/Sales

Appendix B. Macroeconomics Indicators

Acronym	Description	Frequency	Treatment
UNRATE	Civilian Unemployment Rate	Monthly	
INDPRO	Industrial Production Index	Monthly	Rate
CPIAUCNS	Consumer Price Index for All Urban Consumers: All Items	Monthly	Rate
FEDFUNDS	Federal Funds Rate	Monthly	
T10Y2Y	10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity	Monthly	
USALOLITONOSTSAM	Loan Officer Opinion Survey on Bank Lending Practices	Monthly	
GDPC1	Real Gross Domestic Product	Quarterly	Rate
PCE	Personal Consumption Expenditures	Monthly	Rate
SP500	S&P 500 Stock Market Index	Monthly	Rate
VIXCLS	CBOE Volatility Index: VIX	Monthly	
M1SL	M1 Money Stock	Monthly	Rate
M2SL	M2 Money Stock	Monthly	Rate
PAYEMS	All Employees, Total Nonfarm	Monthly	Rate
UEMPMEAN	Average Weeks Unemployed	Monthly	
IPMANSICS	Industrial Production: Manufacturing (SIC)	Monthly	
HOUST	Housing Starts: Total: New Privately Owned Housing Units Started	Monthly	
PERMIT	New Private Housing Units Authorized by Building Permits	Monthly	
PPIFIS	Producer Price Index: Finished Goods	Monthly	
DGORDER	Manufacturers' New Orders: Durable Goods	Monthly	Rate
BUSINV	Total Business Inventories	Monthly	Rate
NETEXP	Net Exports of Goods and Services	Quarterly	
IMPGSC1	Imports of Goods and Services	Quarterly	Rate

FGEXPND	Federal Government: Expenditures	Quarterly	Rate
FGRECPT	Federal Government: Receipts	Quarterly	Rate
TB3MS	3-Month Treasury Bill: Secondary Market Rate	Monthly	
TB6MS	6-Month Treasury Bill: Secondary Market Rate	Monthly	
GS10	10-Year Treasury Constant Maturity Rate	Monthly	
TOTCI	Total Consumer Credit Owned and Securitized, Outstanding	Monthly	Rate
CONSUMER	University of Michigan: Consumer Sentiment	Monthly	Rate
MZMSL	MZM Money Stock	Monthly	
VXOCLS	CBOE S&P 100 Volatility Index: VXO	Monthly	
NFCI	Chicago Fed National Financial Conditions Index	Monthly	
TEDRATE	TED Spread	Monthly	
AAA	Moody's Seasoned Aaa Corporate Bond Yield	Monthly	
BAA	Moody's Seasoned Baa Corporate Bond Yield	Monthly	

*For variables labeled as 'Rate' under 'Treatment', it indicates that the variable should be transformed into a growth rate to eliminate time trends. Variables without this label do not require any processing. For data marked as 'Quarterly' under 'Frequency', we have converted them to monthly data using a backward resampling method.

Appendix C. Trading Signals

Acronym	Description	Formula
chmom_6m	6 month change momentum	$chmom_{6m} = close_t - close_{t-126}$
mom12m	12 month momentum	$mom_{12m} = close_t - close_{t-252}$
mom1m	1 month momentum	$mom_{1m} = close_t - close_{t-21}$
mom36m	36 month momentum	$mom_{36m} = close_t - close_{t-756}$
mom6m	6 month momentum	$mom_{6m} = close_t - close_{t-126}$
retvol	14-day Average True Range (Return volatility)	$retvol_t$ = AVG(True Range over 14 days)
baspread	Bid-ask spread	$baspread_t = close_t - open_t$
dolvol	Dollar trading volume	$dolvol_t = close_t * volume_t$
maxret	Maximum daily return	$maxret_t = (close_t - open_t) / open_t$
std_dolvol	Volatility of liquidity (dollar trading volume for past 21 days)	std_dolvol_t = STD(dolvol over past 21 days)
std_turn	Volatility of liquidity (share turnover for past 21 days)	std_turn_t = STD(volume over past 21 days)
zerotrade	Zero trading days	$zerotrade_t$ = 1 if $volume_t = 0$, else 0
open	Open price	Downloaded from yahoo
high	Highest price in a trading day	Downloaded from yahoo
low	Lowest price in a trading day	Downloaded from yahoo
close	Close price	Downloaded from yahoo
adj_close	Adjust close price	Downloaded from yahoo
volume	Trading volume in a trading day	Downloaded from yahoo