Pontifícia Universidade Católica
do Rio de Janeiro

Guilherme de Moraes Masuko

Forecasting Returns on High-Frequency Environment: A Comparative Study of Econometric Models and Machine Learning Techniques

Dissertação de Mestrado

Masters dissertation presented to the Programa de Pós–graduação em Economia, do Departamento de Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia.

Advisor : Prof. Nathalie Gimenes
Co-advisor: Prof. Marcelo Medeiros

Rio de Janeiro
April 2024

Guilherme de Moraes Masuko

Forecasting Returns on High-Frequency Environment: A Comparative Study of Econometric Models and Machine Learning Techniques

Masters dissertation presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia. Approved by the Examination Committee:

Prof. Nathalie Gimenes
Advisor
Departamento de Economia – PUC-Rio

Prof. Marcelo Medeiros
Co-Advisor
Department of Economics – UIUC
Departamento de Economia – PUC-Rio

Prof. Christian Montes Schütte
Department of Economics and Business Economics – Aarhus University

Prof. Marcelo Fernandes
EESP – FGV

Rio de Janeiro, April 26th, 2024

Guilherme de Moraes Masuko

B.A. in Economics, Universidade Estadual de Londrina, 2021

Bibliographic data

Masuko, Guilherme de Moraes

Forecasting Returns on High-Frequency Environment: A Comparative Study of Econometric Models and Machine Learning Techniques / Guilherme de Moraes Masuko; advisor: Nathalie Gimenes; co-advisor: Marcelo Medeiros. – 2024.

49 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2024.

Inclui bibliografia

1. Economia – Teses. 2. Previsão. 3. Aprendizado por Máquina. 4. Dados em Alta Dimensão. 5. Alta Frequência. 6. Apreçamento de Ativos. 7. Finanças. I. Gimenes, Nathalie. II. Medeiros, Marcelo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. IV. Título.

CDD: 004

To my daughter, Melina.

Acknowledgments

I thank my advisor, Marcelo Medeiros, for his support and guidance throughout this work and for the opportunities that would not have been possible without his involvement.

Additionally, I am grateful to all the professors and colleagues from PUC-Rio with whom I had the pleasure of sharing this endeavor.

I thank my mother, Marlene, and my sister, Isabella, for their support and encouragement.

A special thanks to my girlfriend at the time, and now wife, Karen, for all her affection and patience, despite the distance.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Masuko, Guilherme de Moraes; Gimenes, Nathalie (Advisor); Medeiros, Marcelo (Co-Advisor). **Forecasting Returns on High-Frequency Environment: A Comparative Study of Econometric Models and Machine Learning Techniques**. Rio de Janeiro, 2024. 49p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Forecasting returns on financial assets has been an important task throughout the history of the financial economy. This study employs machine learning (ML) techniques to predict portfolio returns based on the size factor, aiming to not only improve predictions but also understand the underlying source of predictability. Amid the challenge of identifying relevant predictors in noisy data, this research employs a rolling window approach, incorporating three lags of stock returns as candidate predictors to project returns one minute ahead. Benchmark models, including in-sample averaging and autoregressive approaches, are explored alongside ML techniques such as Ridge, LASSO, AdaLASSO, and Random Forest. We consistently identify the superiority of ML models over benchmark models in terms of predictability, with the Random Forest model standing out as the most effective. Furthermore, analysis of the predictors selected by the models revealed that they are predominantly unexpected, short-lived and sparse.

Keywords

Forecasting; Machine Learning; High-dimensional Data; High-frequency; Asset Pricing; Finance.

Resumo

Masuko, Guilherme de Moraes; Gimenes, Nathalie; Medeiros, Marcelo. **Previsão de Retornos em Ambiente de Alta Frequência: Um Estudo Comparativo de Modelos Econométricos e Técnicas de Aprendizado por Máquina**. Rio de Janeiro, 2024. 49p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

A previsão de retornos sobre ativos financeiros tem sido uma tarefa importante durante toda a história da economia financeira. Este estudo emprega técnicas de aprendizado por máquina (ML) para prever retornos de portfólio com base no fator tamanho, visando não apenas melhorar as previsões, mas também compreender a fonte subjacente de previsibilidade. Em meio ao desafio de identificar preditores relevantes em dados ruidosos, esta pesquisa emprega uma abordagem de janela móvel, incorporando três defasagens de retornos de ações como candidatos a preditores para projetar retornos um minuto à frente. Modelos de benchmark, incluindo a média dentro da amostra e abordagens autorregressivas, são explorados junto com técnicas de ML como Ridge, LASSO, AdaLASSO e Random Forest. Identificamos consistentemente a superioridade dos modelos de ML sobre os modelos benchmark em termos de previsibilidade, com o modelo Random Forest se destacando como o mais eficaz. Além disso, a análise dos preditores selecionados pelos modelos revelou que eles são predominantemente inesperados, de curta duração e esparsos.

Palavras-chave

Previsão; Aprendizado por Máquina; Dados em Alta Dimensão; Alta Frequência; Apreçamento de Ativos; Finanças.

Table of contents

List of figures

List of tables

List of Abreviations

ML - Machine Learning
LASSO - Least Absolute Shrinkage and Selection Operator
AdaLASSO - Adaptive LASSO
CRSP - Center for Research in Security Prices
TAQ - Trade and Quote
NYSE - New York Stock Exchange
AMEX - American Stock Exchange
BIC - Bayesian Information Criterion

1
Introduction

Forecasting returns on financial assets represents an important topic in the field of financial economics. This study aims to address this challenge by employing machine learning (ML) techniques as well as traditional methods as benchmark to predict portfolio returns based on the size factor. Using a broad and comprehensive set of stock returns as predictors, our main objective goes beyond improving forecasts, but also seeking to understand the underlying source of predictability in these returns.

The identification of relevant predictors for predicting returns in financial markets has been one of the main research motivations. This task is notably challenging considering highly noisy data - low signal-to-noise ratio (Timmermann (2018)) - due to the inherently unpredictable nature of returns.

Historically, this identification process has been characterized by a thorough and well-founded analysis of companies, often dependent on the individual experience of the researcher. However, given the growing and complex availability of financial data, there is a demand for models capable of not only efficiently manage the considerable volume of information, but also to accurately identify and utilize the most pertinent data for analysis.

In this context, recent studies, such as Han et al. (2023), integrate ML models into the framework proposed by Fama and MacBeth (1973), analyzing the cross-section of stock returns and using 207 company characteristics as predictors at a monthly frequency. In the study, extensions of penalized regressions and combinatorial strategies (ensembling) were used in response to traditional models, in addition to exploring a Random Features model (Neural Networks with only one hidden layer containing a large number of nodes) as a non-linear alternative. These models have presented significant advances in mitigating overfitting problems found in conventional approaches like Ordinary and Weighted Least Squares methods (OLS and WLS, respectively).

For example, the work of Dong et al. (2022) reveals that, by using 100 portfolios of long-short anomalies as predictors of market excess returns on a monthly basis, a variety of ML techniques, including forecast combination and reduction of dimensionality, efficiently extract predictive signals in a high-dimensional configuration, obtaining out-of-sample R^2 between 0.89% and 2.81%.

Avramov, Cheng and Metzker (2023) show that investment strategies based on deep learning, using neural networks, have proven profitable, especially in periods of high market volatility. This finding supports the idea that ML techniques combine multiple weak and difficult-to-identify sources of information into a meaningful signal.

On the other hand, in an intra-day context, the work of Chinco, Clark-Joseph and Ye (2019) showed, by estimating minute-by-minute forecast models for a set of 250 shares daily, that the use of the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani (1996)) captures a different set of information to those considered in a benchmark model, adding a gain in predictability measured through a combination of predictions (Granger

and Ramanathan (1984)). At the same time, Aleti, Bollerslev and Siggaard (2023) adopted LASSO to regularize predictive regressions of returns in the intraday market, selecting between returns from a varied set of lagged factors, highlighting the robustness and applicability of this technique in high-frequency scenarios.

This high-frequency environment has gained prominence due to significantly better results in terms of predictability compared to lower frequencies. The substantial improvement in the results obtained at high frequency can be attributed to Fama (1970)'s market efficiency hypothesis. In low-frequency scenarios, models that achieve good results are likely to be adopted by the asset management industry due to their easy applicability, leading to competitive pressure and, consequently, the reestablishment of balance in the financial market. In contrast, in high-frequency settings, the consistency and durability of results can be attributed to stronger entry barriers.

The models covered in this work make predictions one minute ahead using a rolling window, thus offering estimates and predictions to ensure their applicability and obtaining profits. This approach entails significant technological costs, thus making the consistent level of predictability more affordable. Aït-Sahalia et al. (2022) add to this debate by exploring different time intervals, showing that predictability of $R^2_{\text{OS}} \approx 15\%$,[1] when examining Intel (INTC), at a frequency of 5 seconds dissipates monotonically when expanding to larger intervals, reaching negative R^2_{OS} already at the interval of 5 minutes.

This study uses benchmark models widely used in the literature, such as the within-sample mean and autoregressive models with h lags (defined by the Bayesian Information Criterion) to predict portfolio returns based on the size factor with a minute in advance. Furthermore, it uses machine learning models in a high-dimensional environment due to the use of three-lag returns from a large set of stocks listed on the New York Stock Exchange (NYSE), Nasdaq (NASDAQ) and American Stock Exchange (AMEX). The model is configured with a rolling window to estimate the parameters for each prediction, employing the Ridge, LASSO and AdaLASSO linear models, in addition to Random Forest as a non-parametric alternative. In addition to forecasting size-based portfolio returns, this study goes further by investigating what type of predictor is selected through model selection performed by LASSO and AdaLASSO, as well as which predictors obtained greater relative importance by Random Forest.

The use of ML methods is crucial in high-frequency environments due to their ability to adapt to a large number of explanatory variables, requiring fewer observations for robust estimates. These models adjust to consider only the set of variables relevant to predicting returns at such high frequencies, focusing only, in our case, on the last few minutes of data through the estimation windows.

The preliminary findings of this study highlight the superior performance of the non-parametric Random Forest model, which excels in both adopted measures: out-of-sample R^2 and Accuracy. The LASSO and AdaLASSO mod-

[1] Aït-Sahalia et al. (2022) uses the infeasible out-of-sample average as a benchmark model compared to forecasts, a much harder measure to beat.

els also demonstrate commendable efficacy, achieving significantly positive results. Interestingly, while the Ridge model shows modest performance in terms of out-of-sample R^2, it delivers surprisingly strong results in Accuracy.

The structure of this study is outlined as follows: Section 2 covers details regarding the data set used in this work, including the set of stocks considered for each trading day and details on the construction of the portfolio based on size factor. In Section 3, we justify the use of ML techniques, in addition to discussing traditional models widely used in predictive environments, aiming to quantify the benefit arising from the use of more advanced methods. Subsequently, in Section 4, results based on widely recognized quality measures are presented, such as the out-of-sample R^2 and the Accuracy operator, to evaluate the capacity of models in predicting values and movements one minute in advance. Section 5 provides a thorough analysis of the selected predictors, while Section 6 concludes.

2
Data

In this section, we deepen the exploration of the data for the analysis carried out in this study. We detail the sources of data on stock returns, highlighting the scope and origin of this information. Furthermore, we thoroughly examined the construction of the portfolio based on the size factor, emphasizing the formulation and strategies adopted in its composition. This analysis offers a panoramic view of the crucial data sets used in the research, highlighting their fundamental importance for subsequent analyses.

2.1
Stock Returns

The stock returns data used in this study were obtained from CRSP (Center for Research in Security Prices) and TAQ (Trade and Quote), covering the extensive range of trading days between January 2005 and December 2019, totaling 3773 trading days. This data incorporates a wide range of stocks listed on major U.S. stock exchanges such as the New York Stock Exchange (NYSE), Nasdaq (NASDAQ), and American Stock Exchange (AMEX). Recorded at one-minute intervals, this data generates 389 daily observations for each traded share, with trading hours from 9:31 am to 3:59 pm, excluding the last minute to mitigate possible distortions arising from the closing auction.

A careful filtering step was implemented, removing stocks with a closing price of less than $5 on the previous day. Furthermore, for computational purposes, stocks with more than 20% zero returns were systematically excluded, resulting in a much more restricted set of candidate predictors for each estimation window, without significantly compromising informational integrity. Figure A.1 illustrates the evolution of the number of companies included in the stock returns data over time, before and after the filtering process.

After this filtering process, 33 trading days without the presence of shares were identified, characterizing days of low liquidity in the US financial market. This resulted in the number of days in our sample being reduced from 3773 to 3740. Notably, we observed a recurring pattern on these specific dates: the nine days before the Independence Day holiday (July 4th), the 15 consecutive days following the Thanksgiving Day (celebrated on the fourth Thursday of November) and, finally, nine days on Christmas Eve were also excluded.

2.2
Size Factor-Based Portfolio Returns

The target portfolio to which we direct our forecasts will be built based on the factor investing approach. Factor investing represents an investment strategy that directs the allocation of a portfolio based on the factors that influence the performance of selected assets. Theoretically, this strategy seeks to expand diversification, generate returns above the market average and manage risks more effectively.

These factors can cover macroeconomic aspects, aiming to capture systemic and larger-scale events, as well as style factors, categorizing different types of assets. The core of factor investing lies in the formation of a portfolio aligned to a rule based on a specific factor, which allows capturing aggregate movements in the economy and the financial market.

In the context of factor-based portfolio construction, it is essential to highlight that all companies contained in the original data set are considered, without going through the filtering process applied to the returns data used in the models. This procedure encompasses stocks that have counterparts in the dataset covered in Giglio, Kelly and Kozak (2023) and Haddad, Kozak and Santosh (2020). Figure A.2 highlights the good correspondence between the two bases: stock returns and factors.

The ordering of companies in the factor database is based on the values of specific characteristics of each company. This process categorizes shares listed on the NYSE, NASDAQ and AMEX exchanges into ten percentiles. Although there are 55 company characteristics as ranking criteria, for computational power purpose, this work focuses only on the size factor based on the market capitation characteristic. It is important to highlight that the breakpoints for forming these percentiles are established only based on shares listed on the NYSE, following the approach adopted in Fama and French (2016).

In this study, our focus is exclusively on the size characteristic, measured by the market capitalization of each company. The size factor aims to identify a pattern between companies with high and low market value. While Table A.1 presents descriptive statistics of the average daily returns for the ten size factor portfolios, each representing a decile, Figure A.3 provides a visual representation of the distribution of average daily returns of these portfolios.

By following the approach proposed by Kozak, Nagel and Santosh (2020), we build a portfolio composed of long and short positions, specifically incorporating the top and bottom three percentiles of stock returns, respectively, based on the size factor distribution. Figure A.4 illustrates the distribution of companies between long and short positions in this context.

The weighting of the portfolio based on the size factor is carried out using weights derived from market capitalization, adhering to practices established in academic research in finance. In this way, the portfolio based on the size factor is composed of portfolios weighted by the market value considered within each of the long and short positions.

In mathematical terms, the portfolio return associated with the size factor on day d and minute m is calculated by Equation 2-1:

$$f_{m,d} = \sum_{i \in \mathbf{T}_{3,d}} \omega_{i,d} \cdot r_{i,m,d} - \sum_{j \in \mathbf{B}_{3,d}} \omega_{j,d} \cdot r_{j,m,d} \tag{2-1}$$

Here, $\omega_{i,d} = \frac{\text{MarketCap}_{i,d}}{\sum_{i \in \mathbf{T}_{3,d}} \text{MarketCap}_{i,d}}$ and $\omega_{j,d} = \frac{\text{MarketCap}_{j,d}}{\sum_{j \in \mathbf{B}_{3,d}} \text{MarketCap}_{j,d}}$ represent the weights based on market value, for each day, for the set of shares comprised in long and short positions, while $\mathbf{T}_{3,d}$ and $\mathbf{B}_{3,d}$ denote the subsets of companies belonging to the top three deciles and bottom three deciles of the size factor, respectively, on day d.

Figures A.5 and A.6 offer a statistical perspective on portfolio behavior over time, as well as the dispersion of returns in relation to the average,

respectively, through the visualization of the time series and distribution of the average daily returns of the portfolio based on the size factor.

3
Models

The methodology adopted in this study incorporates elements established by Chinco, Clark-Joseph and Ye (2019), employing a rolling window composed of 150 observations. This approach aims to estimate parameters in order to provide projections for a time horizon of one minute ahead, considering three lags of candidate predictors in most cases.

For example, for the first forecast on a specific day the procedure involves using the rolling window containing two and a half hours of data and a setting of 3 lags. In this context, the estimation of parameters occurs considering data between 9:34 am and 12:03 pm, aiming to predict the returns of factors based on characteristics at 12:04 pm. It is important to mention that data from 9:31 am to 9:33 am is excluded due to lags.

Subsequent projections follow an analogous sequence, expanding the rolling window to the range between 9:35 am and 12:04 pm for the next forecast, scheduled for 12:05 pm. This procedure continues until the last forecast, scheduled for 3:59 pm. The visual representation of this process for a specific day is illustrated in Figure A.7.

The methodological approach of this study is based on previously established concepts, such as the use of a rolling observation window and the consideration of time lags. This strategy aims to capture dynamic patterns and relevant characteristics to predict future returns.

3.1
Benchmark Models

In this section, we explore three fundamental benchmark models for financial forecasting. The first model uses the average of the values in the rolling window, providing a direct baseline through simple averaging to generate predictions. The second model adopts an autoregressive approach of order 3 to make predictions. In turn, the third model, also autoregressive, adapts its h order through the Bayesian Information Criterion (BIC), where $h_{\max} = 10$ is the maximum number of lags considered as candidate predictors, offering dynamic flexibility in determining the model order.

These benchmark models offer valuable points of comparison for evaluating the performance of more complex prediction techniques, providing essential context for evaluating the predictive accuracy and robustness of the ML models discussed later.

3.2
Machine Learning Models

In this section, we delve into the world of machine learning models in financial forecasting. Parametric models such as Ridge, LASSO (Least Absolute Shrinkage and Selection Operator) and AdaLASSO (Adaptive LASSO) will be explored, along with the non-parametric Random Forest model.

ML models have gained prominence due to their effectiveness in predictive capacity in financial contexts. The adoption of ML models is based on their suitability for complex predictive scenarios. Its applicability emerges from the intrinsic ability to learn patterns from data, enabling the modeling of linear and non-linear relationships and adaptation to environments with multiple explanatory variables. This flexibility and adaptability are critical in high-frequency contexts, where the nature of data often presents complexities that challenge traditional approaches.

Before diving into the details of these specific models, it is important to understand the concept of parametric models. In summary, parametric models take on a specific functional form, characterized by a fixed number of parameters that are estimated from the available data. The Ridge, LASSO and AdaLASSO models are examples of this type of model and appear as extensions of Penalized Least Squares.

These models respond to the need to deal with high-dimensional environments, a common scenario in this research context, where the number of explanatory variables exceeds the number of observations ($p > T$). In this environment, obtaining the uniqueness of the parameters via Ordinary Least Squares (OLS) becomes unfeasible due to the impossibility of achieving the rank condition, $\text{rank}(\boldsymbol{X}'\boldsymbol{X}) = p$. The inclusion of regularization terms in OLS optimization problems offers the feasibility of unique estimates.

Furthermore, the application of penalized regression techniques aims to mitigate the overfitting often associated with OLS, a phenomenon widely documented in the literature. The essence of the Penalized Least Squares problem is represented by Equation 3-1.

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left[\sum_{t=1}^{T} \left(Y_t - \boldsymbol{\beta}'\boldsymbol{X}_t\right)^2 + \sum_{j=1}^{p} p_\lambda \left(|\beta_j|\,; \boldsymbol{\alpha},\ \text{data}\ \right) \right] \tag{3-1}$$

where $p_\lambda\left(|\beta_j|\,;\boldsymbol{\alpha}, \text{data}\right)$ is a non-negative penalty function indexed by the regularization parameter λ responsible for control the number of parameters in the model.

To determine the regularization parameter in penalized regression models, two approaches are commonly used: Cross-Validation (CV) and Information Criterion (IC). The Cross-Validation method consists of dividing the estimation sample into several subsamples, using part for estimation and another for evaluation. The goal is to identify the regularization parameter that results in the lowest Mean Square Error (MSE) among various combinations of sample divisions for estimation and evaluation. However, this technique may not be the most appropriate in time series environments, leading us to opt for an Information Criterion.

Among the various Information Criteria available, we chose the Bayesian criterion (BIC) to select the regularization parameter. BIC considers both the MSE and the number of parameters added to the model. This selection is performed using an optimization problem expressed in Equation 3-2.

$$\lambda^{\text{BIC}} = \underset{\boldsymbol{\lambda}\in\Lambda}{\arg\min} \left[T \ln[\hat{\sigma}^2(\lambda)] + \text{df}(\lambda) \ln(T) \right] \tag{3-2}$$

Here, $\hat{\sigma}^2(\cdot)$ represents the Mean Square Error (MSE), while $\text{df}(\cdot)$ indi-

cates the number of variables included in the model. The BIC criterion, by simultaneously considering model complexity and data adequacy, offers a robust approach to choosing the regularization parameter.

3.2.1 Ridge

Ridge, as the initial method of penalized regressions to be discussed, was first introduced into the literature by Hoerl and Kennard (1970). This model uses the ℓ_2 norm in the penalty function, defined as $p_\lambda\left(|\beta_j|\,;\boldsymbol{\alpha},\text{data}\right)=\lambda\beta_j^2$. Due to its strictly convex nature, estimates are found directly using the closed expression $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}=(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

3.2.2 LASSO (Least Absolute Shrinkage and Selection Operator)

LASSO (Least Absolute Shrinkage and Selection Operator), introduced by Tibshirani (1996), is a method belonging to penalized regressions. This model incorporates a form of regularization based on the ℓ_1 norm, where the penalty term is defined as $p_\lambda\left(|\beta_j|\,;\boldsymbol{\alpha},\text{data }\right)=\lambda\,|\beta_j|$. This formulation discourages the inclusion of weak or irrelevant coefficients, assuming the assumption of sparsity in the model, that is, only a limited number $s=\sum_j 1_{\beta_j\neq 0}$ of p regressors differs from zero. Unlike Ridge, which only reduces in magnitude parameter estimates, LASSO tends to cancel out many parameters, effectively taking them to zero, thus also performing model selection.

Although it is not a strictly convex problem especially in high-dimensional environments, Tibshirani (2013) demonstrates that, under the assumption of regressors drawn from a continuous probability distribution, the uniqueness of the LASSO solution has probability one. In the context of stock returns, it is reasonable to make this assumption, considering the feasibility that the regressors are in fact derived from a continuous probability distribution.

3.2.3 AdaLASSO (Adaptive LASSO)

AdaLASSO (Adaptive LASSO) Zou (2006), the last parametric model discussed in this section, is an extension of the previously discussed LASSO. This extension introduces an idiosyncratic penalty component to each regressor in the original LASSO penalty function, defined as $p_\lambda\left(|\beta_j|\,;\boldsymbol{\alpha},\text{data}\right)=\lambda w_j\,|\beta_j|$. The definition of this component may vary, including the inverse of the Ridge or OLS estimates, when this is feasible. In this work, we chose to configure it as the inverse of the LASSO estimates itself, $w_j=\left(\left|\widehat{\beta}_{j,\text{LASSO}}\right|+\frac{1}{\sqrt{T}}\right)^{-\tau}$, with $\tau=1$, as frequently found in the literature. To avoid divergences, we added a positive constant term to the calculation, considering that the LASSO sets some regressors of the model to zero.

The inclusion of AdaLASSO is motivated by its model selection consistency feature, achieved through the Weighted Irrepresentable Condition (WIC) assumption. This condition, unlike LASSO, incorporates the penalty compo-

nent for each of the regressors, making it more viable, as demonstrated by Medeiros and Mendes (2016). Considering the objective of model selection analysis, the adoption of this technique appears to be pertinent.

3.2.4
Random Forest

In contrast to parametric models, non-parametric models do not assume a specific functional form defined by a fixed set of parameters. This characteristic gives non-parametric models greater flexibility to adjust to the data, allowing adaptation to complex and non-linear patterns present in the data.

The non-parametric Random Forest model falls into the class of Ensemble Models and shares similarities with Bagging (Bootstrap Aggregating). While in Bagging, several Bootstrap samples are taken and, in parallel, a predictor model is run for each of them, in the case of Random Forest, each predictor model must be set as a regression tree.

Unlike Bagging, in Random Forest only a random subset of q variables is selected, from the total set of p variables for each split node in each regression tree. Similar to Bagging, each regression tree is built from a bootstrap sample $\mathbf{N}$, selecting observations with replacement from the original data set. The two parameters related to this context were set to 0.7 and 0.5, respectively, representing the proportion of variables for each division node and the proportion of data for each regression tree.

Denoting $\mathbf{Q}$ as the set of variables selected at each division node, the Random Forest method searches, for each $k \in \mathbf{Q}$, a threshold x that divides the sample from Bootstrap $\mathbf{N}$ in two subsets

$$\mathbf{N}^+(k,x) = \{t \in \mathbf{N} : X_{t,k} > x\} \quad \text{e} \quad \mathbf{N}^-(k,x) = \{t \in \mathbf{N} : X_{t,k} \leq x\}$$

where $X_{t,k}$ is the t-th observation of the k-th variable.

Denoting $\bar{Y}^+$ by the mean of Y_t in the subset $\mathbf{N}^+(k,x)$ and $\bar{Y}^-$ by the mean of Y_t in the subset $\mathbf{N}^-(k,x)$. The division node (k,x) is chosen in order to minimize the problem:

$$\widehat{k}, \widehat{x} = \text{argmin}_{k,x} \left\{ \sum_{t \in \mathbf{N}^+(k,x)} \left(Y_t - \bar{Y}^+_{k,x}\right)^2 + \sum_{t \in \mathbf{N}^-(k,x)} \left(Y_t - \bar{Y}^-_{k,x}\right)^2 \right\}$$

The tree construction process continues recursively, repeating this process Z times (number of division nodes). The stopping criterion adopted in this work is the maximum number of tree depths, set to 5.

The prediction function of the b-th regression tree is therefore given by the following expression:

$$\widehat{f}_b(\boldsymbol{X}_{\text{new}}) = \sum_j \widehat{\beta}_{j,b} 1_{\{\boldsymbol{X}_{\text{new}} \in \mathcal{R}_{j,b}\}}$$

where $\widehat{\beta}_{j,b}$ is simply the average of the observations of the dependent variables in the estimation data located in the partition $\mathcal{R}_{j,b}$.

Finally, the prediction of the Random Forest model with B trees is the average of these predictions:

$$\widehat{Y}_{\text{new}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_b(\boldsymbol{X}_{\text{new}})$$

where $B = 500$ is adopted in this study.

Just like in the LASSO and AdaLASSO models, where we employ their model selection characteristics, in the non-parametric Random Forest model, we will use the feature importance as a measure to evaluate the main predictors in the process of forecasting the portfolio returns based on the size factor. At each division node z, the variable k is selected to maximize the reduction in MSE. By denoting $\widehat{\delta}_z^2$ as the gain in MSE with respect to no splitting, the relative importance of a predictor k in a given tree f is computed by the following formula.

$$\mathcal{I}_k(f) = \sum_{z=1}^{Z} \widehat{\delta}_z^2 1_{\{i_z = k\}}$$

where Z is the number of splitting nodes and i_z is the index of the splitting variable.

In summary, for the Ridge, LASSO and AdaLASSO models, the prediction of portfolio returns based on the size factor is modeled using Equation 3-3.

$$f_{m,d} = \alpha + \sum_{j=1}^{3} \boldsymbol{\beta}_j' \boldsymbol{r}_{m-j,d} + \eta_{m,d} \tag{3-3}$$

where $f_{m,d}$ represents the size factor-based portfolio returns for minute m and day d, and $\boldsymbol{r_{m-j,d}}$ is a vector of the j-th lagged stock returns described in Section 2.

While for the Random Forest model, the prediction of portfolio returns based on the size factor is expressed through Equation 3-4.

$$f_{m,d} = \frac{1}{B}\sum_{b=1}^{B} f_b((\boldsymbol{r}_{m-1,d}, \boldsymbol{r}_{m-2,d}, \boldsymbol{r}_{m-3,d})') \tag{3-4}$$

It is important to note that three lags of all stock returns will be employed as candidate predictors.

4
Results

To evaluate each model's prediction, we rely on two widely used measures, the out-of-sample coefficient of determination (R^2_{OS}) and Accuracy. Among the results obtained in this study, we have predictions from the models discussed in the Chapter 3 and the true values, between the period from 12:04 pm to 3:59 pm, totaling 236 predictions for each of the days in our sample. This way, both measurements will be calculated at daily frequency, that is, for each day, we calculate R^2_{OS} and Accuracy for each model using the 236 predictions. It is also worth highlighting that the results will be presented as the monthly average of the results computed on a daily basis, thus obtaining a statistic for standard deviation, enabling the representation of a confidence interval for these results.

4.1
Out-of-sample R^2 (R^2_{OS})

To assess how well models predict, empirical finance research basically follows two out-of-sample measures of R^2. The first is the measurement in its centered version, that is, when computing the predictive quality of a model, what we are actually doing is comparing this prediction with a benchmark model, often set as the in-sample mean of the predicted variable, as discussed in Campbell and Thompson (2008). Thus, the measure of R^2_{OS} represented by Equation 4-1 presents positive values when the model predictions are better than the prediction calculated using the in-sample mean, while negative values of R^2_{OS} determine worse model predictions in relation to the computation of the average values within the sample.

$$R^2_{\text{OS}}(\boldsymbol{f}_d, \hat{\boldsymbol{f}}_d) = 1 - \frac{\sum_m (f_{m,d} - \hat{f}_{m,d})^2}{\sum_m (f_{m,d} - \frac{1}{L}\sum_l f_{m-l,d})^2} \tag{4-1}$$

where $\boldsymbol{f}_d$ and $\hat{\boldsymbol{f}}_d$ represent, respectively, the vectors of the true returns of the size-based portfolio and the returns predicted by a given model for day d. The terms $f_{m,d}$ and $\hat{f}_{m,d}$ denote, respectively, the actual values of the size-based portfolio return and the values predicted by the model for the minute m and day d. The parameter m runs through all the minutes in which model predictions occurred, covering the interval from 12:04 pm to 3:59 pm. In turn, the parameter l runs through the minutes of the rolling window, made up of L observations.

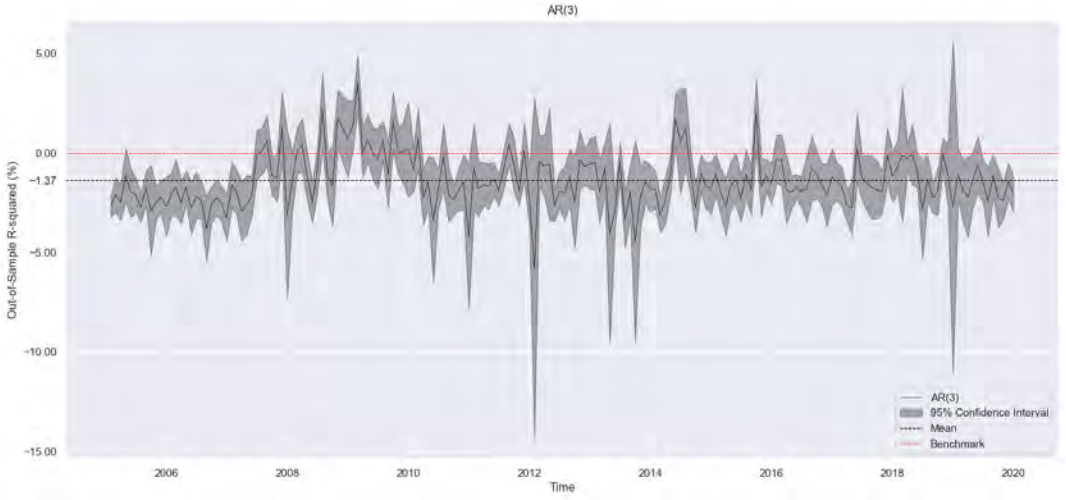
The second measurement of out-of-sample R^2 differs from the first in relation to the benchmark model to be compared. Setting this model to simply the value zero, the measure, used mostly in more recent research, is known as the non-centered version of R^2_{OS}. The choice for the centered version of R^2_{OS} is made for two reasons: firstly, one of the benchmark models addressed by this work is the historical in-sample mean of the rolling window, and secondly because the rolling window contains only 150 return observations at a frequency of one minute, thus causing the sample average to obtain values

very close to zero. Therefore there is no great relevance in the choice between the two measures.

Exploring the performance of the AR(3) and AR(h) models, presented in Figures 4.1 and 4.2, in high-frequency financial environments reveals a significant challenge in overcoming the predictive accuracy of simple benchmarks as the in-sample mean. Despite the theoretical sophistication of these autoregressive models, which are designed to leverage historical data points to predict future returns, their effectiveness in the context of our study was limited. On average, both models exhibited negative out-of-sample values of R^2, indicating a struggle to consistently provide predictive power beyond what could be inferred from the historical average of the data.

Figure 4.1: Out-of-Sample R-squared of AR(3) Model

This figure presents the results in terms of R^2_{OS} for the AR(3) predictive model. The R^2_{OS} is calculated from Equation 4-1 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model represented by the in-sample mean.



This result serves as a powerful reminder of the complexities inherent in modeling financial markets at high frequencies. It suggests that the dynamic and often unpredictable nature of these environments can make traditional econometric approaches less effective, especially when trying to capture the nuances of market movements over very short time frames.

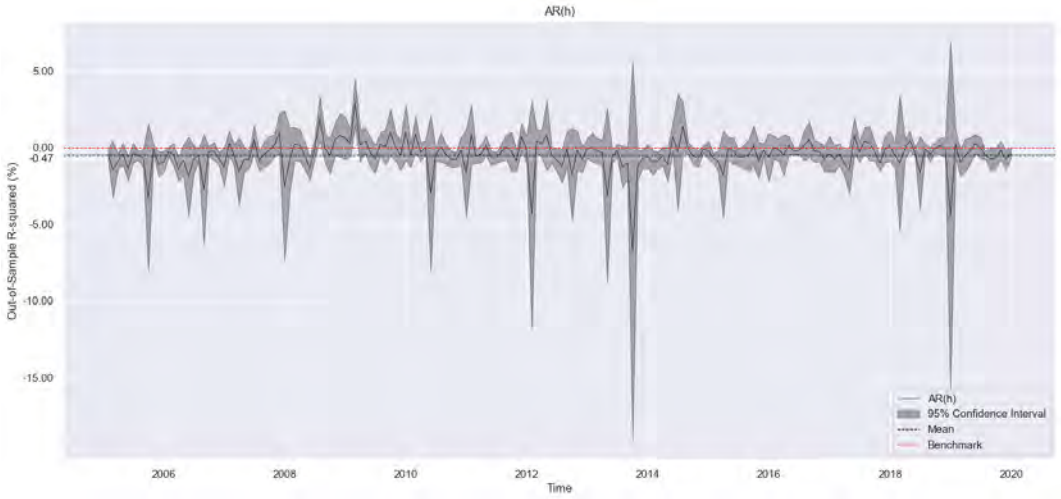
In essence, the modest performance of the AR(3) and AR(h) models was already expected since financial markets are approximately efficient and, therefore, finding predictability in them is a challenging task.

These results altogether highlight the need for financial modelers to consider alternative strategies, potentially incorporating more elaborate techniques and different data sources capable of adapting more dynamically to rapid market changes, to improve pr'edictive performance in high-frequency trading contexts.

The performance of the Ridge regression model, as shown in Figure 4.3, highlights the challenges in high-frequency financial forecasting, especially when the model incorporates predictors that may not significantly impact

Figure 4.2: Out-of-Sample R-squared of AR(h) Model

This figure presents the results in terms of R^2_{OS} for the predictive model AR(h) (h defined using the Bayesian Information Criterion). The R^2_{OS} is calculated from Equation 4-1 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model represented by the in-sample mean.



portfolio returns. Initially, the model presents poor performance, a trend that continues with high volatility in its predictive accuracy. This volatility suggests that, although Ridge regression attempts to mitigate overfitting through regularization, it may erroneously diminish the influence of important predictors or, conversely, fail to adequately penalize predictors that contribute little to portfolio predictability.

Such results reflect the critical balance needed in the selection and weighting of predictors in financial models. In the case of Ridge regression, considering irrelevant predictors can lead to an underestimation of essential market signals or an over-reliance on less impactful variables. This limitation points to the need for models capable of selecting truly important predictors or also capable of estimating their nonlinear structure. Models that prioritize predictor selection and capture the nonlinear dynamics of financial markets can offer more effective forecasting tools in the fast-paced environment of high-frequency trading.

When examining the performance of the linear machine learning models LASSO and AdaLASSO, which are detailed in Figures 4.4 and 4.5, a distinct pattern emerges that differentiates these models in our analysis. During certain intervals, both models demonstrated positive and remarkably significant results, underlining their potential in effectively predicting high-frequency financial returns. This performance contrasts with some of the traditional models we examined, showing the strengths of these regularization techniques in increasing predictive power.

The LASSO model, with its inherent ability to perform variable selection by reducing less important predictor coefficients to zero, offers an approach capable of dealing with high-dimensional data from financial markets. This feature not only improves the interpretability of the model, but also reduces

Figure 4.3: Out-of-Sample R-squared of Ridge Model

This figure presents the results in terms of R^2_{OS} for the Ridge predictive model. The R^2_{OS} is calculated from Equation 4-1 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model represented by the in-sample mean.

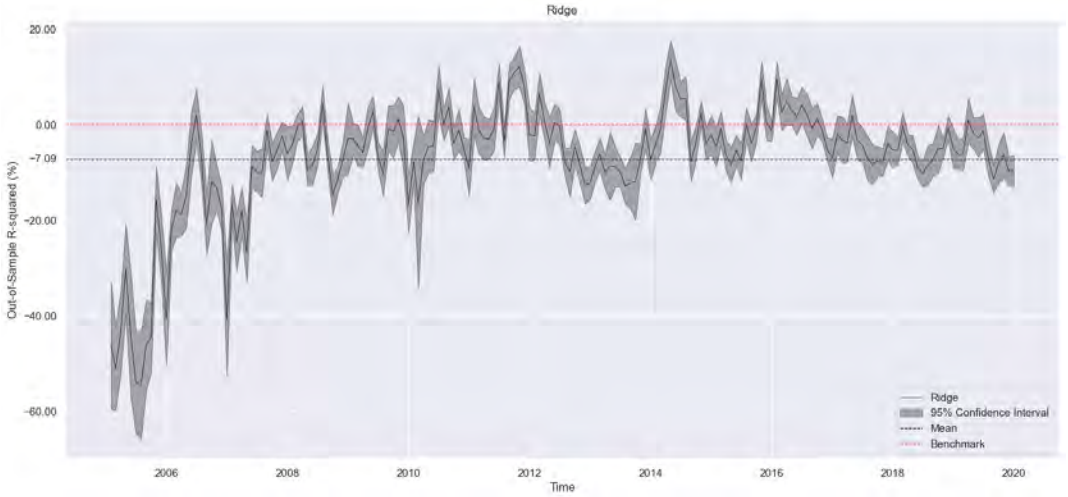


Figure 4.4: Out-of-Sample R-squared of LASSO Model

This figure presents the results in terms of R^2_{OS} for the LASSO predictive model. The R^2_{OS} is calculated from Equation 4-1 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model represented by the in-sample mean.

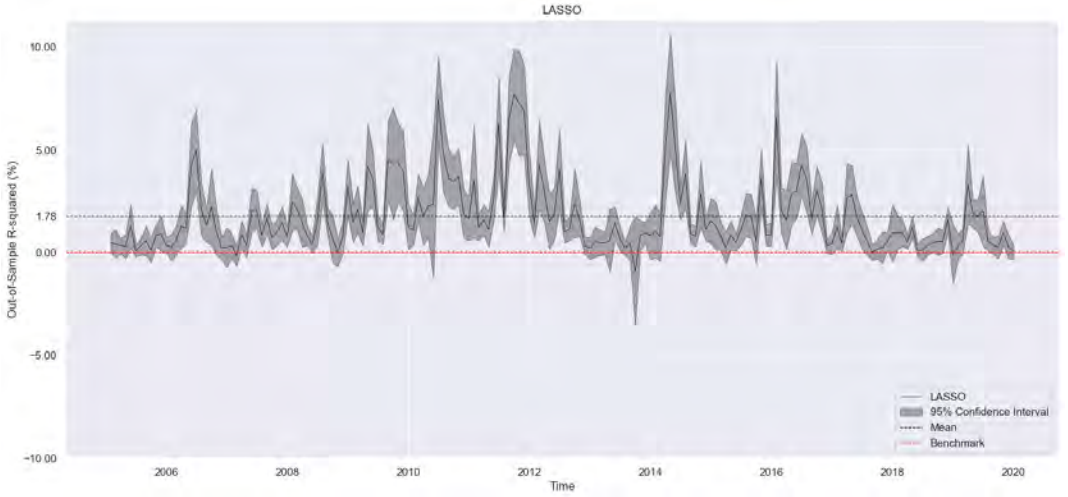
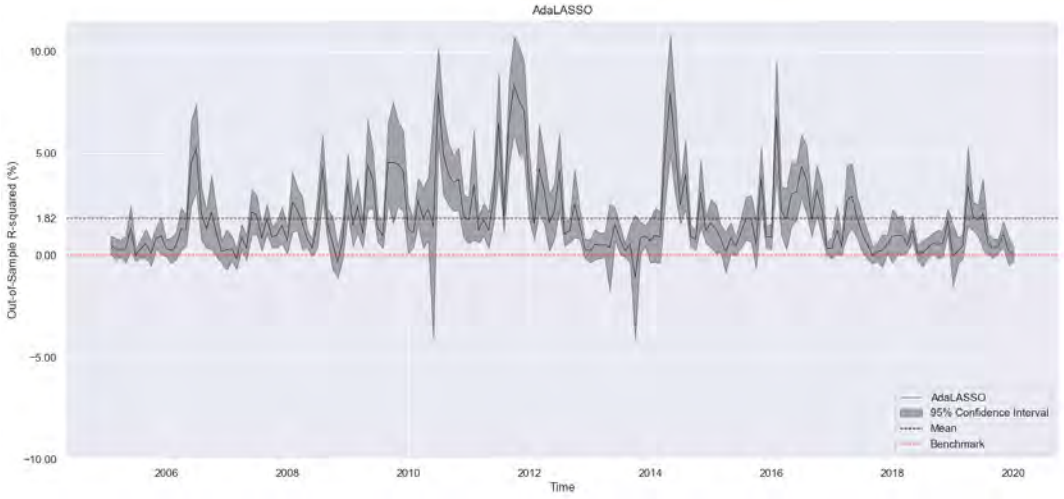


Figure 4.5: Out-of-Sample R-squared of AdaLASSO Model

This figure presents the results in terms of R^2_{OS} for the AdaLASSO predictive model. The R^2_{OS} is calculated from Equation 4-1 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model represented by the in-sample mean.



the risk of overfitting, which is crucial in a high-frequency context where the signal-to-noise ratio tends to be particularly low.

Likewise, the AdaLASSO model, an adaptive version of LASSO, further refines this process by assigning different weights to the regularization of each coefficient. This adaptability makes AdaLASSO particularly effective in environments where the predictive relevance of variables can change over time, as is often the case in financial markets. This improvement in the penalty process, carried out in two stages, significantly reinforces the robustness of the model.

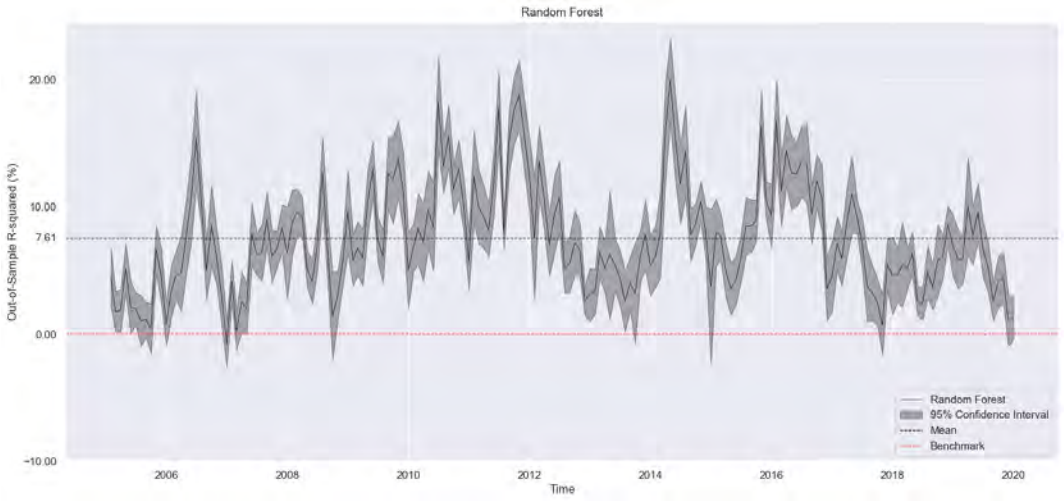
The positive results observed during specific periods for both models highlight their ability to not only adapt to the complex dynamics of financial data, but also to extract meaningful predictive signals from a vast data set. This suggests that the incorporation of such machine learning techniques can significantly improve the toolkit for financial analysts looking to predict asset returns with greater accuracy and reliability.

The non-parametric Random Forest model clearly shows its superiority in the domain of high-frequency financial forecasting, as evidenced by its performance depicted in Figure 4.6. The adaptability of this model to the dynamic and complex environment of financial markets highlights its robustness, particularly in identifying non-linear relationships between predictors that traditional linear models often ignore. The Random Forest approach, which integrates multiple decision trees to make more accurate and stable predictions, inherently takes into account interactions and dependencies between variables, allowing for a more complex understanding of the factors relevant in predicting portfolio returns based on the factor size.

This methodology not only improves the predictive power of the model, but also offers a quantitative analysis of the importance of individual predic-

Figure 4.6: Out-of-Sample R-squared of Random Forest Model

This figure presents the results in terms of R^2_{OS} for the AdaLASSO predictive model. The R^2_{OS} is calculated from Equation 4-1 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model represented by the in-sample mean.



tors, facilitating a deeper analysis of the dynamics that govern this market. The Random Forest model's ability to perform well in diverse market conditions without the need for a pre-defined functional form of the non-linear interactions of predictors is a testament to its versatility and effectiveness in financial forecasting.

Furthermore, the success of the Random Forest model in this study illustrates the potential benefits of applying machine learning techniques to analyzing financial data. Encourages further exploration of non-parametric models and advanced computational methods to improve the predictive power of forecasts. The model's success in capturing nonlinear patterns presents a compelling argument for its inclusion in the arsenal of tools of modern financial researchers.

The results in terms of R^2_{OS} over the entire study period are briefly encapsulated in Table 4.1, offering an overview of the comparative performance of various forecasting models . This table aggregates the mean and standard deviation of the R^2_{OS} values, providing a statistical summary that elucidates the relative effectiveness and consistency of each model in predicting the size-based portfolio.

Table 4.1: Descriptive Results of Out-of-Sample R^2 (%)

This table presents the descriptive statistics of the mean and standard deviation for the results of the AR(3), AR(h) models with h determined by the Bayesian information criterion, Ridge, LASSO, AdaLASSO and Random Forest, in terms of R^2_{OS}. The R^2_{OS} was calculated using Equation 4-1.

	AR(3)	AR(h)	Ridge	LASSO	AdaLASSO	Random Forest
Mean	-1.3680	-0.4730	-7.0858	1.7784	1.8245	7.6053
Standard Deviation	4.1137	4.3211	16.5859	3.2705	3.4889	7.0627

When summarizing these results, it becomes clear that forecasting financial markets with high accuracy remains a challenge, as evidenced by the varying degrees of success in the models tested. However, the analysis also high-

lights the potential for certain models, especially machine learning approaches like Random Forest, to adapt and function robustly within the complex and often unpredictable picture of financial data. These findings underscore the importance of continued exploration in model choice, emphasizing the need for financial modelers to employ a diverse toolkit that can accommodate the complex dynamics of market behavior.

4.2 Accuracy

The Accuracy measure, also described in Aït-Sahalia et al. (2022), provides valuable information about the predictive performance of a model by quantifying the proportion of predicted returns $\hat{f}_{m,d}$ that share the same sign as the realized returns $f_{m,d}$. The measure, represented by Equation 4-2, is much less ambitious than R^2_{OS}. While the latter tells us how close the model's prediction came to the true value, the measure of Accuracy simply concerns whether the model has predictive power over the direction of returns for each minute. Even so, the measure is extremely important in a finance scenario, where knowing significantly about the movements of an asset can generate profitability.

$$\text{Accuracy}(\boldsymbol{f}_d, \hat{\boldsymbol{f}}_d) = \frac{1}{M} \sum_m 1_{\{\hat{f}_{m,d} \cdot f_{m,d} > 0\}} \tag{4-2}$$

where $1_{\{\hat{f}_{m,d} \cdot f_{m,d} > 0\}}$ is an indicator function that evaluates to 1 when the product of predicted and actual returns is greater than zero, indicating accurate predictions, and 0 otherwise.

The initial evaluation of the models in terms of Accuracy begins with the in-sample mean model, as illustrated in Figure 4.7. This model demonstrates superior performance when compared to the AR(3), AR(h) and Ridge models in the context of R^2_{OS}. However, when it comes to outperforming a simple benchmark – conceptually similar to flipping a coin, where heads predict a positive movement and tails a negative one – the in-sample mean model does not present a clear advantage.

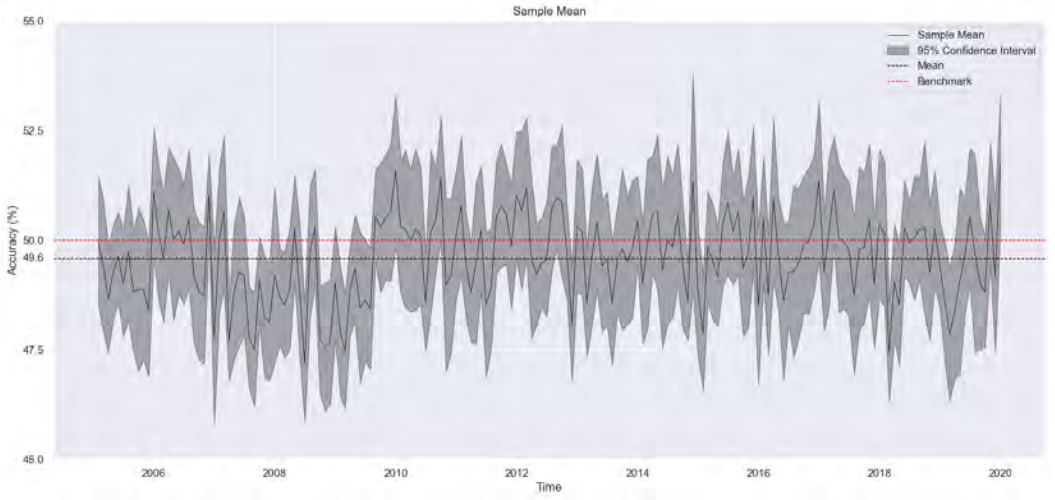
This phenomenon can largely be attributed to the inherently conservative nature of forecasts through the in-sample mean, especially in high-frequency trading environments. Here, predictions are typically very close to null values, making any substantial prediction error a significant obstacle to beating the in-sample mean in terms of R^2_{OS}. Despite this, the challenge presented by the in-sample mean's performance does not extend to its accuracy in predicting the direction of portfolio movements minute-by-minute.

This discrepancy highlights the crucial distinction between the ability to predict accurate future values and the ability to correctly anticipate only the direction of a portfolio's movements, highlighting the usefulness of the in-sample mean in certain contexts despite its limitations in others.

When analyzing the performance of the AR(3) and AR(h) models through the lens of Accuracy, they demonstrate an intriguing pattern of results. These models, serving as additional benchmarks, reveal a different picture of prediction accuracy when compared to their results in the context of R^2_{OS}. The AR(3) model, in particular, performs in certain segments of the sample,

Figure 4.7: Accuracy of In-Sample Mean Model

This figure presents the results in terms of Accuracy for the In-Sample Mean predictive model. The Accuracy is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.



providing significantly positive Accuracy rates, as illustrated in Figure 4.8. This suggests that, despite the simplicity of the model, it has the ability to accurately predict the direction of portfolio movements based on the size factor during specific intervals, highlighting its value in certain market conditions.

On the other hand, the performance of the AR(h) model, although designed to adapt its complexity based on the Bayesian Information Criterion (BIC) to potentially improve predictive accuracy, appears to fail significantly over parts of the period under analysis. The figure 4.9 illustrates that this model often produces results that are not statistically significant. The reasonably different success between these two autoregressive models highlights the complexity of financial market data and the need for models that balance adaptability with predictive accuracy.

Furthermore, the comparative improvement in Accuracy for these models over the performance of R^2_{OS} also highlights the essential aspect of financial forecasting, where the ability to correctly forecast the direction of movement of a portfolio may be different from the accuracy of predictions in terms of magnitude of return. This is particularly relevant in trading strategies where the main objective is to capitalize on directional movements rather than predicting specific return values.

Among the models evaluated, Ridge regression stands out for its unexpected Accuracy results, as evidenced by the data illustrated in Figure 4.10. Despite its depressing performance in terms of out-of-sample R^2, suggesting a lesser ability to accurately predict portfolio returns, the model demonstrates a surprising ability to predict the direction of movements of the portfolio based on size. This peculiar result suggests that although Ridge regression may not be excellent at capturing the exact magnitude of returns due to its regularization mechanism that potentially smooths out some predictive power, it retains

Figure 4.8: Accuracy of AR(3) Model

This figure presents the results in terms of $\mathrm{Accuracy}$ for the AR(3) predictive model. The $\mathrm{Accuracy}$ is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.

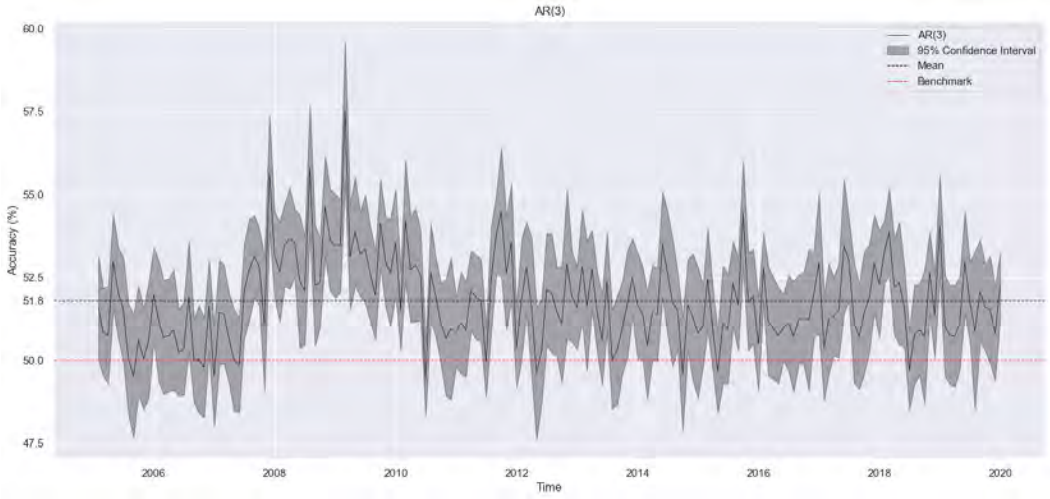
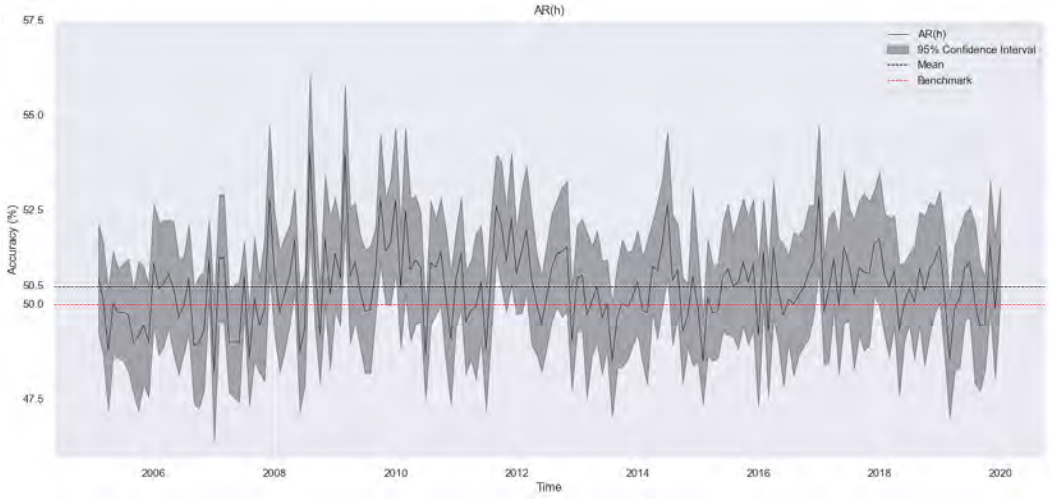


Figure 4.9: Accuracy of AR(h) Model

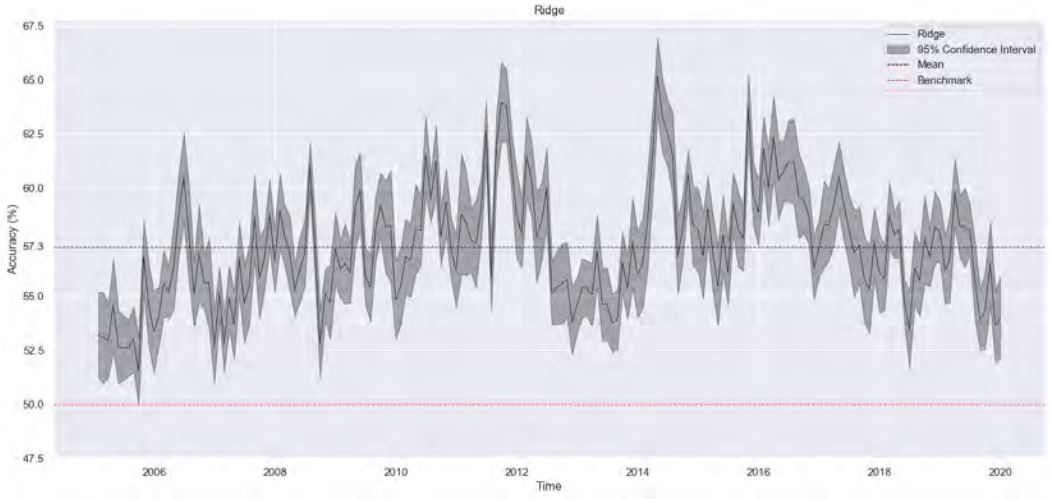
This figure presents the results in terms of $\mathrm{Accuracy}$ for the predictive model AR(h) (h defined using the Bayesian Information Criterion). The $\mathrm{Accuracy}$ is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.



a surprising ability in understanding the trend of portfolio movements.

Figure 4.10: Accuracy of Ridge Model

This figure presents the results in terms of Accuracy for the Ridge predictive model. The Accuracy is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.



The fact that the model has a higher Accuracy again implies that, in the context of high-frequency financial data, the ability to predict directional movements does not necessarily correlate with the model's ability to predict return values accurately. This discrepancy highlights the usefulness of Ridge regression in scenarios where the primary objective is to identify directions of movements rather than quantifying specific future returns, offering valuable implications for trading strategies that rely more on market dynamics than exact return projections.

The LASSO and AdaLASSO models stand out for their consistent forecasting effectiveness, evidenced by their performance in both Accuracy and R^2_{OS} measures. These models have demonstrated their ability to achieve significantly positive accuracy results also in terms of Accuracy over several months, a testament to their robustness in high-frequency financial market predictions, as presented in Figures 4.11 and 4.12.

The LASSO model, known for its ability to select and regularize variables, effectively reduces the complexity of models by penalizing the absolute size of regression coefficients. Doing so not only mitigates the risk of overfitting, but also improves the interpretability of the model by keeping only the variables with the greatest predictive power. This feature is particularly beneficial in the context of financial data, where the large number of potential predictors can easily lead to complex and difficult to interpret models.

Adaptive LASSO, based on the fundamentals of the LASSO model, by adjusting the penalty applied between different coefficients based on the parameters estimated by LASSO, allows for more flexibility and potentially improves forecast accuracy. The AdaLASSO feature significantly increases its

Figure 4.11: Accuracy of LASSO Model

This figure presents the results in terms of Accuracy for the LASSO predictive model. The Accuracy is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.

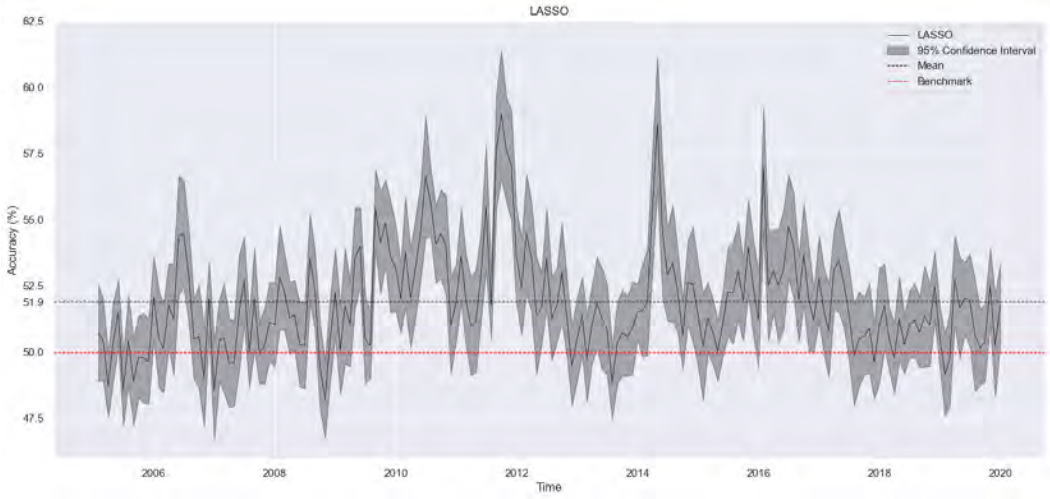
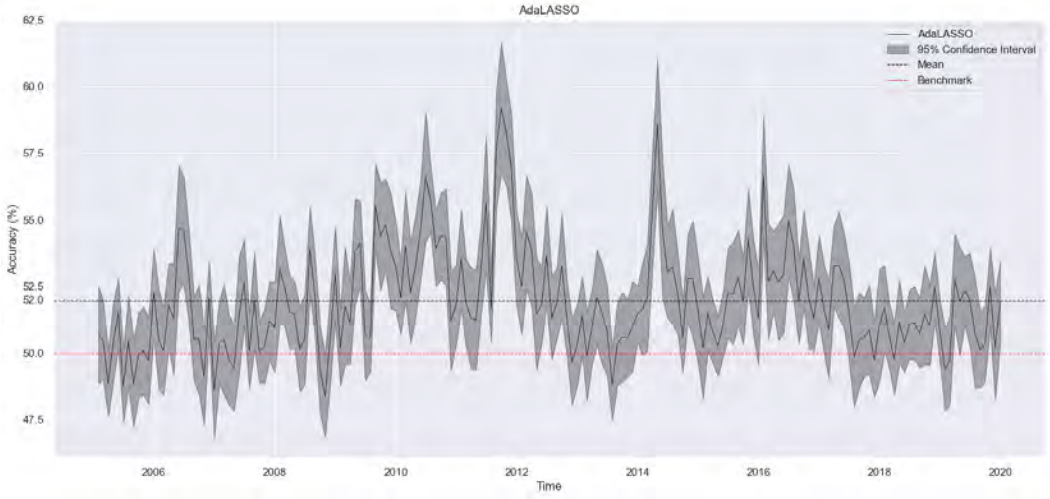


Figure 4.12: Accuracy of AdaLASSO Model

This figure presents the results in terms of Accuracy for the AdaLASSO predictive model. The Accuracy is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.



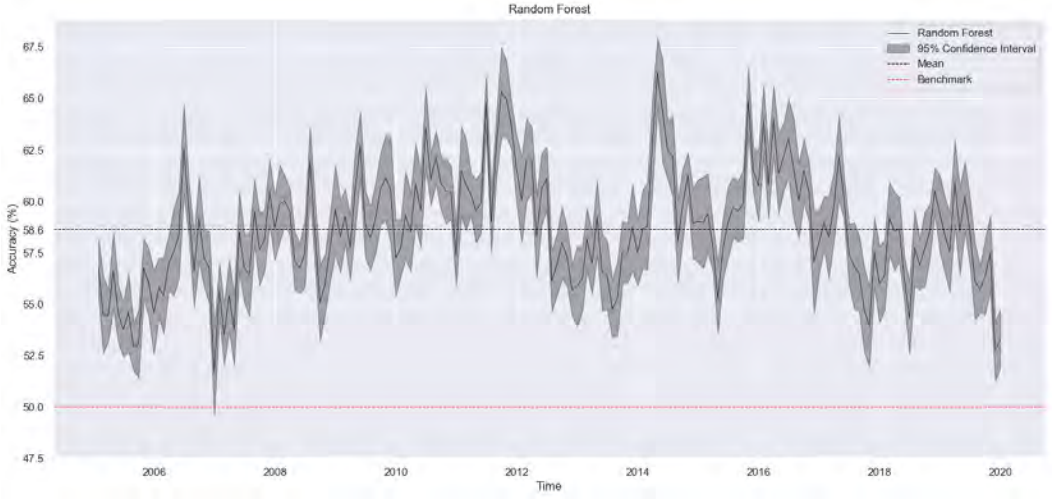
robustness in distinguishing between relevant and irrelevant predictors, further refining the model's predictions.

The positive results observed with the LASSO and AdaLASSO models during certain months highlight their effectiveness in predicting movements of the portfolio based on the size factor. Its performance highlights the potential of regularization techniques to increase the predictive power of traditional statistical models, especially in environments characterized by a lot of noise and an abundance of data.

The Random Forest model once again establishes itself as the best in adapting to obtaining accurate predictions also in terms of Accuracy within the set of models evaluated in this study. This non-parametric approach has consistently demonstrated superior performance, achieving significantly positive results that underline its effectiveness in predicting size-based portfolio directional movements. As depicted in Figure 4.13, the model's ability to excel not only in terms of out-of-sample R^2 but also in Accuracy highlights its comprehensive predictive capabilities.

Figure 4.13: Accuracy of Random Forest Model

This figure presents the results in terms of Accuracy for the Random Forest predictive model. The Accuracy is calculated from the Equation 4-2 based on the predicted and true values from the first forecast at 12:04 pm to the last one at 3:59 pm for a given day. The solid black line represents the monthly average of these results, while the gray band illustrates the 95% confidence interval of these results. The black dashed line represents the average of the model results for the entire analyzed period, while the red dashed line represents the benchmark model interpreted as flipping a coin and assigning a positive movement in case of heads and a negative movement in case of tails.



Random Forest effectively addresses the limitations of individual models by ensambling predictions from multiple decision trees. This method reduces the risk of overfitting by balancing the trade-off between bias and variance more effectively than other statistical methods. Furthermore, its inherent mechanism for dealing with non-linear relationships and interactions between predictors without the need for explicit specification makes it particularly suitable for the complex dynamics of financial markets.

The results in terms of Accuracy, covering the entire study period, are compiled in Table 4.2. This table serves as a quantitative summary, presenting mean and standard deviation metrics, to provide a comparative perspective on the performance of various models. The diversity in model performance

underscores the importance of selecting the appropriate forecasting tool based on the specific needs and dynamics of the financial markets in question.

Table 4.2: Descriptive Results of Accuracy (%)

This table presents the descriptive statistics of the mean and standard deviation for the model results In-sample mean, AR(3), AR(h) with h determined by the Bayesian information criterion, Ridge, LASSO, AdaLASSO and Random Forest, in terms of Accuracy. The Accuracy was calculated using Equation 4-2.

	In-Sample Mean	AR(3)	AR(h)	Ridge	LASSO	AdaLASSO	Random Forest
Mean	49.5837	51.8029	50.4688	57.2871	51.8964	51.9627	58.6372
Standard Deviation	3.5227	3.7208	3.7125	4.6235	4.3526	4.3848	4.7558

In conclusion, it becomes clear that forecasting in high-frequency financial markets presents a major challenge, with machine learning models, especially the non-parametric Random Forest model, standing out for its forecasting ability both in terms of R^2_{OS} as in Accuracy. The AR(3) and AR(h) models have limitations in their predictions, highlighting the complexity of capturing financial market dynamics with traditional econometric approaches. Interestingly, the Ridge regression model, despite its poor result in terms of R^2_{OS}, was shown to be capable of accurately predicting the direction of portfolio movements, suggesting its potential usefulness in certain predictive contexts. Meanwhile, the regularization techniques employed by the LASSO and AdaLASSO models have produced promising results, emphasizing the importance of models that perform predictor selection among a huge set of candidates.

5
Predictor Analysis

The predictor analysis in this study, inspired by work carried out in Chinco, Clark-Joseph and Ye (2019), adopts three distinct approaches to examining the selected predictors. Initially, we investigated whether there is any prior bias in the model selection performed by LASSO and AdaLASSO, as well as in the relative importance attributed by the Random Forest model. Furthermore, for the LASSO and AdaLASSO models, we explore the average duration that a predictor remains selected during daily forecasts, along with their respective sparsities.

The results are recorded at one-minute intervals for each day covered by this research. In other words, for each minute between 12:04 and 15:59, for LASSO and AdaLASSO models that perform model selection, we assign values of 0, 1, 2 or 3 to each company, indicating respectively whether none, one, two or three lags of this company were selected. In the case of the Random Forest model, in the same way, we record the importance of the predictor in an aggregated manner for each minute of a given day. This means that if, for example, a company obtains importance of 0.1, 0.2 and 0.3 for its three lags in a given minute, the aggregate value of importance for that company is 0.6 on that specific day.

5.1
Unexpected

In this section, we seek to identify predictors that may stand out among market capitalization percentiles and industry classification in predicting size-based portfolio returns. To this end, we investigated whether there is a significantly greater proportion of companies selected by LASSO and AdaLASSO, as well as a greater relative importance attributed by Random Forest, in any of these segmentations.

To consolidate these results, we aggregated the one-minute frequency results for a daily perspective. Thus, in the LASSO and AdaLASSO models, we count the number of times any lag of an action was selected as a predictor within a specific day. In the case of the Random Forest model, we first normalize the importance of each company for each minute, considering its proportion within that period. We then aggregate these relative importance values to a daily level, calculating the average. This way, we obtain an average value of relative importance for each company during a given day.

The classification of companies is initially segmented by size factor percentiles and, subsequently, by industry classification. This allows us to check whether stocks in a specific percentile or industry are more frequently selected by the LASSO and AdaLASSO models, or whether, in the case of the Random Forest model, there is a different relative importance for stocks between these classifications.

In the first approach, when grouping the predictors into their respective percentiles, we consider the number of candidate predictors in each percentile.

For the LASSO and AdaLASSO models, we calculate the ratio of selected predictors from companies belonging to a percentile to the total number of candidates in that same percentile and normalize this value for each day. In the case of the Random Forest model, we followed a similar procedure, computing the ratio between the aggregate mean of relative importance and the number of candidates for each percentile. Again, these values were normalized in proportion for each day.

The results of this analysis are presented in Table 5.1. We observed that, in the case of Random Forest, no percentile demonstrated significantly different behavior in terms of importance in portfolio prediction. As for the LASSO and AdaLASSO linear models, the results indicate a tendency to select larger capitalization stocks more frequently, although the considerable errors make any conclusive statement difficult.

Table 5.1: Predictor Analysis by Factor Size Percentile

This table presents the descriptive statistics of the mean and standard deviation (in parentheses) for the model selection results of the Random Forest, LASSO and AdaLASSO models using the size factor percentiles as the segmentation criterion. The numbers represent the proportion of relative importance (Random Forest) and selected predictors (LASSO and AdaLASSO) within each of the size factor percentiles. The results of relative importance (Random Forest) and selected predictors (LASSO and AdaLASSO) were computed for each of the minutes in which there was a prediction. These results were aggregated for each percentile taking into account the number of candidates within each percentile and finally aggregated at a daily level. The mean and standard deviation were calculated from the results on a daily basis.

	Percentile 1	Percentile 2	Percentile 3	Percentile 4	Percentile 5	Percentile 6	Percentile 7	Percentile 8	Percentile 9	Percentile 10
Random Forest	11.8011	10.5406	9.7506	9.7024	10.1966	10.2876	10.2920	10.0616	9.6704	8.9649
	(1.9611)	(1.7242)	(1.5513)	(1.6937)	(2.0933)	(2.3683)	(2.5508)	(3.2169)	(3.9348)	(4.3135)
LASSO	14.0076	13.1883	12.7703	11.1647	11.3704	11.1841	10.5646	7.3570	5.0503	3.9662
	(12.9231)	(11.9625)	(12.0027)	(11.5449)	(12.5764)	(13.1492)	(14.3474)	(12.9902)	(12.4439)	(11.7470)
AdaLASSO	13.8832	13.1147	12.724765	11.1785	11.4250	11.2277	10.5676	7.3723	5.1539	3.9803
	(12.8807)	(12.025844)	(12.031016)	(11.6196)	(12.6630)	(13.2226)	(14.4273)	(13.0266)	(12.5969)	(11.8519)

Similarly, predictors are grouped by industry. The results presented in Table 5.2 corroborate previous analyses. In the Random Forest model, only the Information sector seems to be slightly more relevant in predicting portfolio returns compared to the other sectors. In the LASSO and AdaLASSO models, we again observed that some industries seem to have a greater average importance, however, again the substantial deviations make any definitive conclusion difficult.

Table 5.2: Predictor Analysis by Industry Classification

This table presents the descriptive statistics of the mean and standard deviation (in parentheses) for the model selection results of the Random Forest, LASSO, and AdaLASSO models using industry classification as the segmentation criterion. The numbers represent the proportion of relative importance (Random Forest) and selected predictors (LASSO and AdaLASSO) within each of the industries. The results of relative importance (Random Forest) and selected predictors (LASSO and AdaLASSO) were computed for each of the minutes in which there was a prediction. These results were aggregated for each industry taking into account the number of candidates within each industry and finally aggregated at a daily level. The mean and standard deviation were calculated from the results on a daily basis.

	Manufacturing	Finance and Insurance	Information	Retail Trade	Professional, Scientific, and Technical Services	Utilities
Random Forest	6.6649 (1.2340)	6.3807 (1.4659)	8.2987 (2.0984)	5.9129 (1.3517)	6.8842 (2.1575)	5.7251 (2.1039)
LASSO	9.6659 (9.7015)	9.4400 (12.1061)	7.6289 (10.5553)	7.7805 (10.8311)	7.2039 (11.3003)	3.2432 (8.5857)
AdaLASSO	9.6986 (9.8022)	9.3984 (12.0621)	7.6270 (10.5464)	7.8162 (10.8750)	7.2571 (11.3777)	3.2560 (8.6314)
	Wholesale Trade	**Mining, Quarrying, and Oil and Gas Extraction**	**Transportation and Warehousing**	**Accommodation and Food Services**	**Administrative and Support and Waste Management and Remediation Services**	**Health Care and Social Assistance**
Random Forest	6.0770 (1.9459)	5.7613 (1.4507)	6.0700 (1.8360)	5.9481 (2.1633)	5.8238 (2.1317)	6.1561 (2.6520)
LASSO	6.4419 (11.8066)	7.2438 (10.9843)	6.5888 (11.6697)	5.3149 (11.1046)	5.4033 (11.8202)	3.9777 (10.6343)
AdaLASSO	6.4553 (11.9398)	7.2275 (10.9580)	6.6040 (11.7162)	5.3207 (11.2189)	5.3834 (11.8971)	4.0022 (10.7195)
	Real Estate and Rental and Leasing	**Arts, Entertainment, and Recreation**	**Construction**	**Other Services (except Public Administration)**	**Agriculture, Forestry, Fishing and Hunting**	**Educational Services**
Random Forest	6.2540 (2.9573)	5.5482 (2.4261)	6.2153 (2.8424)	5.0595 (3.5694)	5.4551 (3.6353)	5.1500 (3.3097)
LASSO	5.9867 (14.1194)	4.5539 (11.8491)	7.3169 (14.5914)	2.6826 (10.7555)	2.1930 (10.6381)	2.3785 (9.7413)
AdaLASSO	5.9672 (14.1705)	4.5303 (11.9134)	7.2949 (14.6604)	2.6432 (10.7419)	2.0932 (10.5109)	2.3940 (9.8149)

In summary, we conclude that, overall, it is challenging to make definitive statements about what type of predictor may be playing a more important role in predicting size-based portfolio returns. This highlights the complexity of the model identification problem, which transcends conventional approaches. Therefore, a first characteristic we find among candidate predictors in size-based portfolio forecasting is that they are unexpected.

5.2 Short-Lived

In this section, we highlight the importance of adopting a rolling window modeling approach, which estimates parameters for each subsequent forecast. To do this, we focus on the duration that a predictor remains selected by the LASSO and AdaLASSO models.

We approach this question in the following way: for each company considered as a candidate predictor, we check whether any of its lags were selected at any minute during a specific day. If a company had any of its lags selected at one minute, we begin to count how long that selection persists. For each subsequent minute in which a delay remains selected, we increase this duration parameter by one unit. Counting stops when no lag is selected anymore. If any lag is selected again, we start a new independent count. We repeat this procedure for each stock.

This gives us data on the duration of various selections. We calculate the probability between these selections, obtaining the probability of a duration being greater than x minutes. It is important to highlight that we only performed this calculation for stocks that were selected during the day in question, which gives a specific interpretation. Thus, the result obtained is the probability that the duration of an action remains selected for more than x minutes, given that this action was selected on that day. This probability is expressed in Equation 5-1.

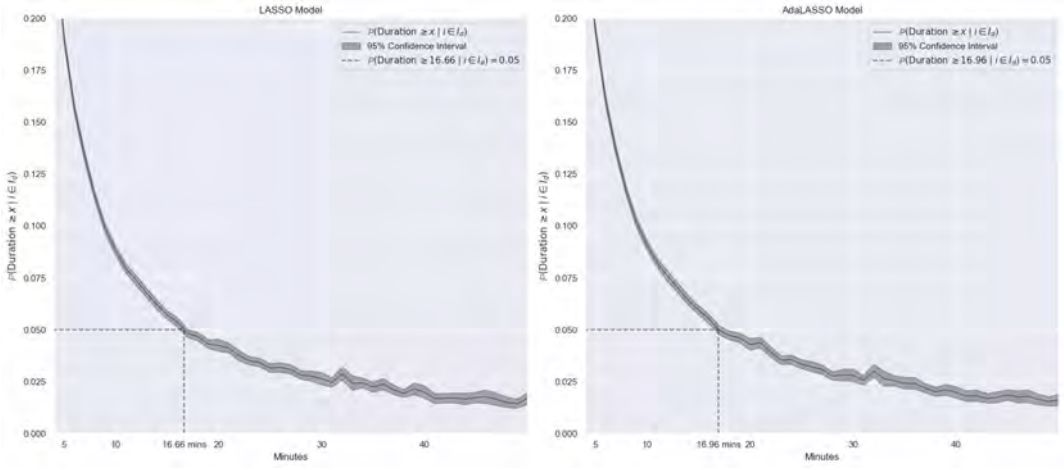
$$\mathbb{P}(\text{Duration} \geq x \mid i \in I_d) \tag{5-1}$$

where x represents the minutes, i is the company index and I_d is the set of companies selected on day d.

In Figure 5.1, we present the results for both the LASSO and AdaLASSO models. As the results vary for each day, that is, the probability of the duration being greater than or equal to 2 minutes on the first day of the sample, January 3, 2005, may not necessarily be equal to the probability on the last day, December 31, 2019, then we can calculate a 95% confidence interval for these results. The graph highlights the point commonly used in the literature, with $\alpha = 0.05$, showing that the probability of the duration of an action remaining selected for more than 16.66 and 16.96 minutes by the LASSO and AdaLASSO models, respectively, is 5%.

Figure 5.1: Probability of Duration being greater than x minutes

This figure presents the probability that each company will remain selected as a predictor for more than x minutes, given that these companies were selected on this day. This is represented by $\mathbb{P}(\text{Duration}_i \geq x \mid i \in I_d)$, where I_d denotes the set of companies selected on day d. To calculate this number, we look at each of the companies that had one of their lags selected during the day and count the duration while any lag from the same company is being selected. The count stops when none of its lags are selected, and if there is another selection ahead, a new count begins. Thus, we obtain the duration (possibly more than one) for each of the companies, allowing us to calculate the probability of these durations for the day in question. The graph on the left presents the results for the LASSO model, while the graph on the right presents the results for the AdaLASSO model. In the figure, each solid black line expresses the probability that the duration of a company i is greater than x minutes, given that this company was selected on day d. The gray band around each line represents the 95% confidence interval. Furthermore, the dashed lines indicate the points at which the probability of the duration of a given company being greater than 16.66 and 16.96 minutes, given that this company was selected on this day, is 5%, for the LASSO and AdaLASSO models, respectively.



With this analysis, we can conclude that, even among only companies that were selected on a given day, the duration of their selection is low. Thus, we find evidence to say that size-based portfolio predictors are short-lived.

5.3 Sparse

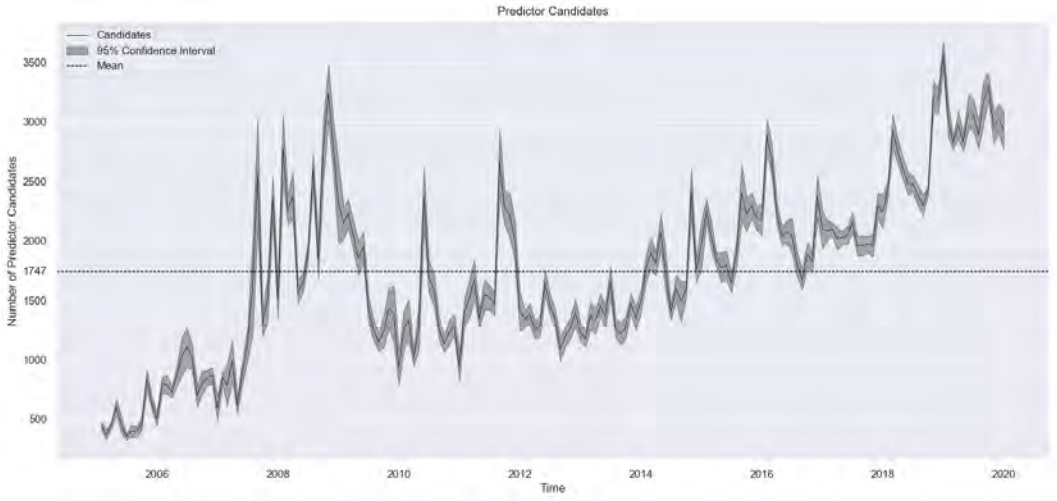
In the analysis of sparsity over selected predictors in the context of forecasting the one-minute-ahead returns of a portfolio based on the size factor, the LASSO and AdaLASSO models show remarkable selectivity among a substantial set of candidate predictors. Given an average of 1747 candidate

predictors, both models discern and use, on average, only 1.4 predictors per minute, concluding the sparse nature of this predictive environment in finance. This minimal selection highlights the models' efficiency and accuracy in isolating the most influential predictors from a vast data set. Such sparsity is not only indicative of the robustness of the models against overfitting, but also highlights their ability to identify the most significant features amidst so much noise, which is fundamental in high-dimensional data scenarios typical of financial time series.

The methodological approach, which involves counting selected predictors every minute and subsequent aggregation into monthly averages, allows for an understanding of model selection behavior over time. The resulting numbers, which trace the evolution of the number of candidates and the average number of selected predictors, as presented in Figures 5.2 and 5.3, offer a visual representation of the behavior of the LASSO and AdaLASSO in terms of sparsity.

Figure 5.2: Number of Predictor Candidates

This figure shows the evolution of the number of candidate predictors over the years. The solid black line represents the monthly average of this number, while the gray band around it indicates the 95% confidence interval. Furthermore, the dashed line shows the average number of candidate predictors for the entire period covered.



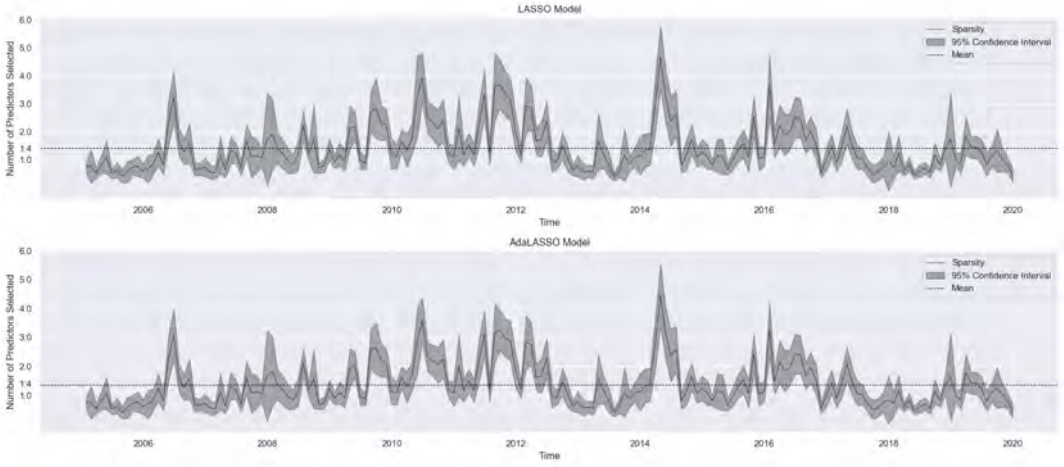
The result observed in the LASSO and AdaLASSO models aligns with the theoretical expectations of these regularization techniques, designed to improve model interpretability and prediction accuracy by imposing restrictions that limit the number of variables included in the final model. This feature is particularly beneficial in financial econometrics, where large volumes of data can easily lead to complex models that may not generalize well to a predictive environment.

Overall, the findings from this sparsity analysis contribute to our understanding of the predictive dynamics at play in size-based portfolio returns. They reflect the critical balance between model complexity and predictive power, affirming the value of the LASSO and AdaLASSO models in effectively addressing the high-dimensional space of financial predictors.

The comprehensive analysis of predictors in this study elucidates three principal attributes of the predictors when forecasting size-based portfolio

Figure 5.3: Number of Predictors Selected

This figure shows the evolution of the number of predictors selected over the years. The upper graph presents the results for the LASSO model, while the lower graph presents the results for the AdaLASSO model. The solid black line represents the monthly average of this number, while the gray band around it indicates the 95% confidence interval. Furthermore, the dashed line shows the average number of predictors selected for the entire period covered.



returns: they are unexpected, short-lived, and sparse, features also found by Chinco, Clark-Joseph and Ye (2019). This triad of characteristics underscores the unpredictable essence of financial markets, as delineated by the inability to consistently identify specific predictors or sectors that uniformly exhibit higher importance or selection frequency across models. The short-lived attribute of predictor selection further accentuates the dynamic and volatile financial market environment, highlighting that even the predictors that are selected maintain their relevance for merely fleeting moments. The sparsity observed in the model's selection process by LASSO and AdaLASSO models—which select, on average, merely 1.4 predictors out of thousands—underscores that just a small handful of variables are really important in forecasting this portfolio. This sparsity indicates a focused approach in cutting through the vast data noise to pinpoint the genuinely impactful predictors. Together, the unexpectedness, short-lived nature, and sparsity of predictors shine a light on the complex dynamics of financial time series forecasting. These findings pose both challenges and opportunities for the ongoing development and refinement of econometric models, aiming for enhanced predictability and understanding of market behaviors.

6
Conclusion

Forecasting returns on financial assets in high-frequency environments represents a critical challenge in contemporary financial economics. This study was dedicated to significantly advancing this area by adopting an approach that employs machine learning (ML) techniques to forecast size-based portfolio returns one minute in advance. More than just improving forecasts, our goal was to understand the underlying source of predictability in returns.

Throughout this research, we conducted a detailed exploration, covering both traditional econometric models and advanced ML techniques. We use a broad dataset of stock returns as predictors, consistently identifying the superiority of ML models over benchmark models in terms of predictability. Notably, the Random Forest model emerged as the most effective among them.

Furthermore, the investigation of the predictors selected by the models revealed important characteristics. These predictors are predominantly unexpected, short-lived and sparse, highlighting the need for advanced approaches to deal with the complexity and volume of data in modern financial markets.

In summary, this research represents a significant contribution to the field of predicting returns on financial assets in high-frequency environments. It reinforces the argument in favor of using machine learning models in financial forecasting and highlights the complexity underlying forecasting returns in dynamic, high-dimensional environments.

7
Bibliography

AÏT-SAHALIA, Y. et al. **How and When are High-Frequency Stock Returns Predictable?** [S.l.], 2022.

ALETI, S.; BOLLERSLEV, T.; SIGGAARD, M. Intraday market return predictability culled from the factor zoo. **Available at SSRN 4388560**, 2023.

AVRAMOV, D.; CHENG, S.; METZKER, L. Machine learning vs. economic restrictions: Evidence from stock return predictability. **Management Science**, INFORMS, v. 69, n. 5, p. 2587–2619, 2023.

CAMPBELL, J. Y.; THOMPSON, S. B. Predicting excess stock returns out of sample: Can anything beat the historical average? **The Review of Financial Studies**, Society for Financial Studies, v. 21, n. 4, p. 1509–1531, 2008.

CHINCO, A.; CLARK-JOSEPH, A. D.; YE, M. Sparse signals in the cross-section of returns. **The Journal of Finance**, Wiley Online Library, v. 74, n. 1, p. 449–492, 2019.

DONG, X. et al. Anomalies and the expected market return. **The Journal of Finance**, Wiley Online Library, v. 77, n. 1, p. 639–681, 2022.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **The journal of Finance**, JSTOR, v. 25, n. 2, p. 383–417, 1970.

FAMA, E. F.; FRENCH, K. R. Dissecting anomalies with a five-factor model. **The Review of Financial Studies**, Oxford University Press, v. 29, n. 1, p. 69–103, 2016.

FAMA, E. F.; MACBETH, J. D. Risk, return, and equilibrium: Empirical tests. **Journal of political economy**, The University of Chicago Press, v. 81, n. 3, p. 607–636, 1973.

GIGLIO, S.; KELLY, B. T.; KOZAK, S. **Equity term structures without dividend strips data**. [S.l.], 2023.

GRANGER, C. W.; RAMANATHAN, R. Improved methods of combining forecasts. **Journal of forecasting**, Wiley Online Library, v. 3, n. 2, p. 197–204, 1984.

HADDAD, V.; KOZAK, S.; SANTOSH, S. Factor timing. **The Review of Financial Studies**, Oxford University Press, v. 33, n. 5, p. 1980–2018, 2020.

HAN, Y. et al. Cross-sectional expected returns: New fama-macbeth regressions in the era of machine learning. **Available at SSRN 3185335**, 2023.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

KOZAK, S.; NAGEL, S.; SANTOSH, S. Shrinking the cross-section. **Journal of Financial Economics**, Elsevier, v. 135, n. 2, p. 271–292, 2020.

MEDEIROS, M. C.; MENDES, E. F. 1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. **Journal of Econometrics**, Elsevier, v. 191, n. 1, p. 255–271, 2016.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 58, n. 1, p. 267–288, 1996.

TIBSHIRANI, R. J. The lasso problem and uniqueness. 2013.

TIMMERMANN, A. Forecasting methods in finance. **Annual Review of Financial Economics**, Annual Reviews, v. 10, p. 449–479, 2018.

ZOU, H. The adaptive lasso and its oracle properties. **Journal of the American statistical association**, Taylor & Francis, v. 101, n. 476, p. 1418–1429, 2006.

A
Appendix

A.1
Figures

Figure A.1: Number of Firms Before and After Filtering Process in Returns Data set

This figure shows the evolution of the number of firms before and after the filtering process of the stock return data set. The initial stock returns dataset incorporates companies listed on the United States stock exchanges New York Stock Exchange (NYSE), Nasdaq (NASDAQ), and American Stock Exchange (AMEX). The filtering process applied to this dataset removes stocks with a closing price of less than U$5 on the previous day. Also, stocks with more than 20% zero returns were systematically excluded.

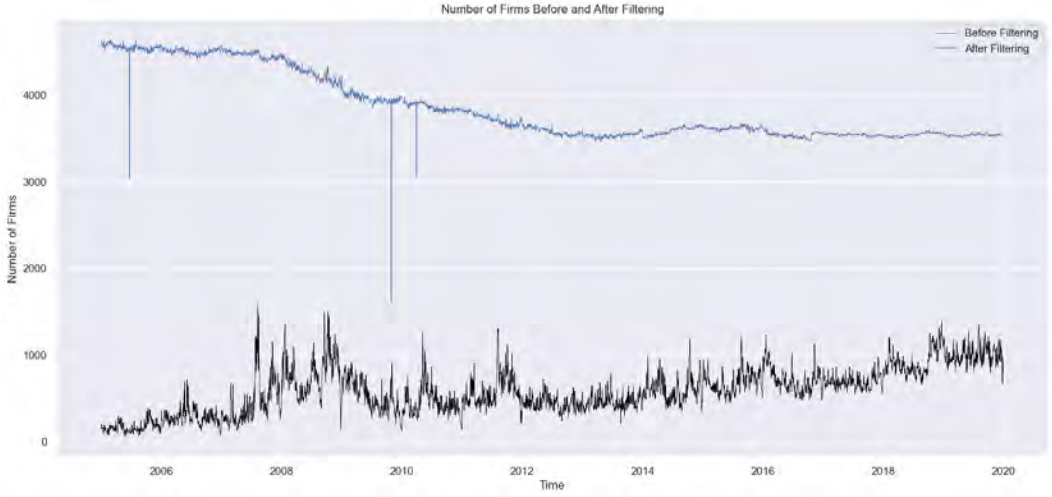


Figure A.2: Number of Firms Matching on Returns and Factors Data Sets

This figure presents the number of firms in the two data sets: Returns and Factors. To construct the portfolio based on size, it was necessary to join these two sets of data using the PERMNO identifier variable for each company. The figure also illustrates the number of companies that contain this variable declared in both databases.

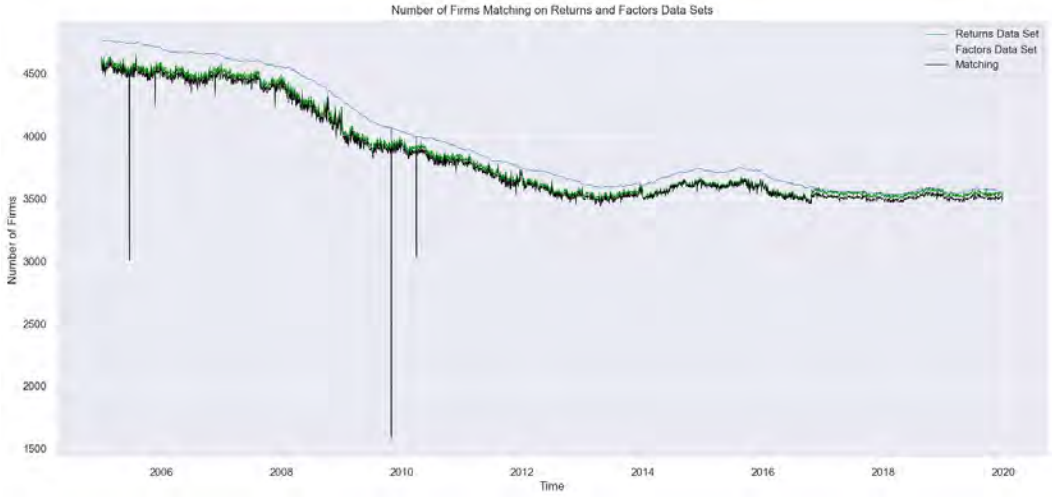


Figure A.3: Empirical Distributions of 10 Percentiles Size Factor-Based Portfolios

This figure presents the empirical probability distribution of the average daily return of portfolios based on the k-th decile. The portfolios were constructed from the set of initial stock returns (without the filtering process presented in Figure A.1) using market value as the portfolio weighting criterion. The portfolios were constructed at a one-minute frequency, but the empirical probability distributions of these portfolios consider the average of daily returns.

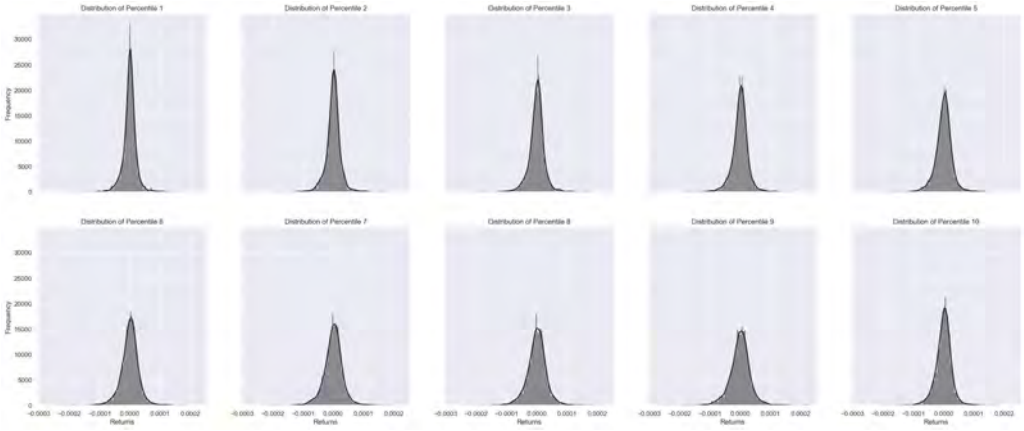


Figure A.4: Number of Firms on Long and Short Position

This figure presents the number of firms that are in the top three deciles (long) and the bottom three deciles (short) based on the size factor.

Figure A.5: Series of Size Factor-Based Portfolio Returns

This figure presents the time series of portfolio returns based on the size factor.

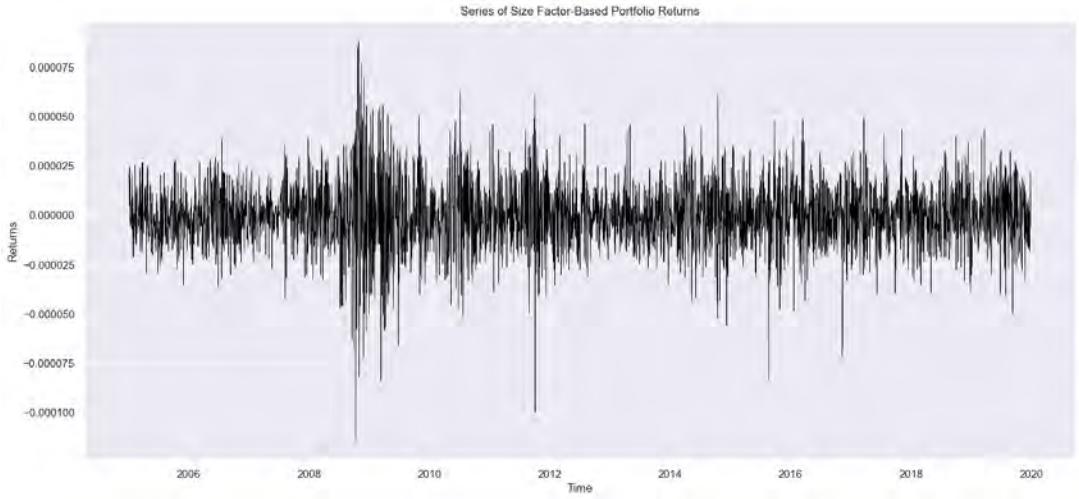


Figure A.6: Distribution of Size Factor-Based Portfolio Returns

This figure presents the empirical probability distribution of portfolio returns based on the size factor.

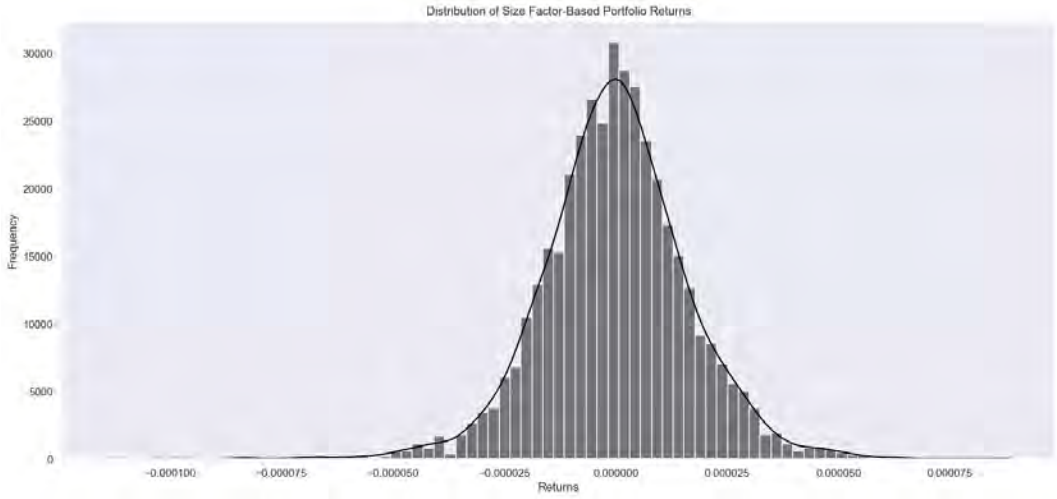
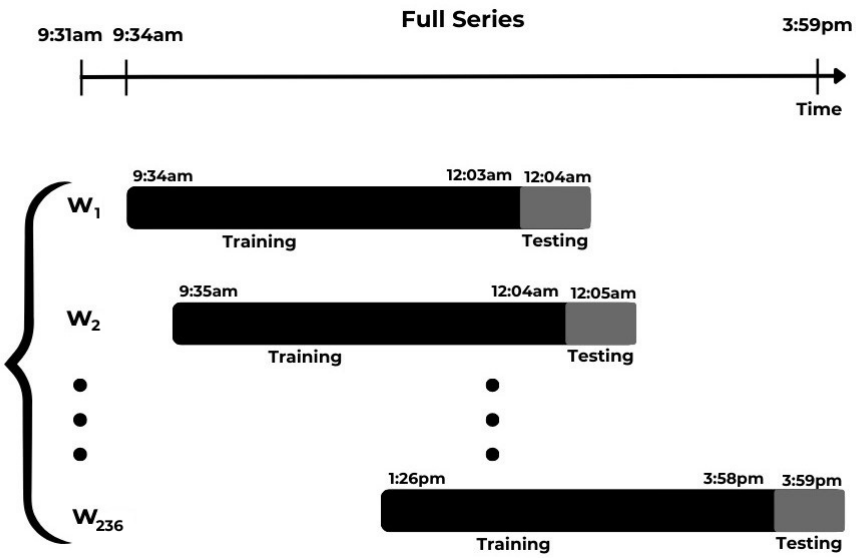


Figure A.7: Rolling Window Scheme

This figure presents the forecast scheme using the rolling window for a specific day.



A.2
Tables

Table A.1: Descriptive Statistics of 10 Percentiles Size Factor-Based Portfolios

This table presents the estimated descriptive statistics of the mean and standard deviation for portfolios based on the size factor. Each portfolio was constructed using the set of shares of companies belonging to the respective decile, operating in the long position and using market capitalization as the portfolio weighting criterion.

	Percentile 1	Percentile 2	Percentile 3	Percentile 4	Percentile 5	Percentile 6	Percentile 7	Percentile 8	Percentile 9	Percentile 10
Mean	0.0	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001	0.0	0.0
Standard Deviation	0.000023	0.000026	0.000027	0.000027	0.000028	0.000031	0.000032	0.000033	0.000036	0.000027