Firm Characteristics and Stock Returns in Brazil[*]

Ruy M. Ribeiro[†] Josué P. A. Costa[†] Mohammed M. Kaebi[†]

Tomas Nóbrega[†] Igor F. B. Martins[†]

This version: March 2024

Abstract

We investigate the impact of firm characteristics on stock returns in the Brazilian financial market, considering a long list of characteristics found be relevant in the U.S. market. Employing Fama-MacBeth regressions, alongside machine learning techniques, we examine over 24 firm-level characteristics. Our findings highlight the stronger influence of price-related metrics, such as momentum, liquidity, size and volatility, over accounting variables. We also explore the robustness of these characteristics through the construction of various portfolios, revealing significant alphas in multiple portfolio construction methods and substantial out-of-sample performance.

Keywords: Asset pricing, stock characteristics, portfolio choice.

JEL Classification: G12; G14.

[*] This project was implemented by members of the Center for Finance and Macroeconomics at Insper. We thank seminar participants at Insper for valuable discussion, comments, and suggestions. E-mail address: ruymr@insper.edu.br (Ribeiro).

[†] Insper, Rua Quatá, 300 – Vila Olímpia, São Paulo – SP, 04546-042, Brazil.

1 Introduction

This paper investigates the influence of various firm characteristics on stock returns in the Brazilian financial market. Focusing on over 24 firm-based characteristics, we seek to understand their influence on stock returns in this emerging market, drawing on parallels and contrasts with more established markets like the United States. Our methodology encompasses multiple Fama-MacBeth regressions across different data subsets, supplemented by advanced techniques such as LASSO, non-parametric LASSO (as proposed by Freyberger et al. (2020)) and Random Forest analysis. Additionally, we assess portfolios sorted on these characteristics to identify those with significant alphas and create long-short portfolios with the identified characteristics.

Here we evaluate whether the characteristics that have been identified in the U.S. context are indeed relevant in the Brazilian market. We do not propose any new characteristic that could be particular to the Brazilian market. Hence, we only explore the specificity of the Brazilian market by testing whether well-known characteristics are robust when analyzed in a different market.

Our results indicate that price-related metrics, specifically 12-month momentum, liquidity, size and volatility, are more closely associated with stock returns than accounting variables, although the latter show significance in certain cases. This study not only contributes to the understanding of the Brazilian financial market but also establishes a foundational dataset for further research. We anticipate the creation of country-specific risk factors based on these characteristics and plan to make our dataset available to the academic community, thereby fostering additional research in this area.

This paper follows a long literature identifying characteristics related with returns. This has been a heated topic in financial research with new characteristics and factors brought into attention and has been particularly prominent after Cochrane's (2011) presidential address posing the challenge of mapping the 'veritable zoo' of factors, which are fundamentally related to firm characteristics. In this paper, we closely follow Green et al. (2017) and then enhance the analysis with additional machine learning techniques.

Besides their relevance, our study diverges from U.S. market research in some key aspects. First, we face limited availability of comprehensive data for the companies listed in the Brazilian financial market. Our preliminary data analysis revealed the presence of extreme values that could potentially skew our results. To address this issue, we undertake a rigorous validation process, which includes cross-referencing our data with alternative data-sources, including primary sources. In this analysis, we separate our outliers in errata and anomalies, and then apply a quantitative criterion that aims to exclude erratas and keep anomalies. This step is crucial to ensure the accuracy and reliability of our findings. Additionally, we employ winsorization and standardization techniques to refine our dataset, effectively mitigating the impact of these outliers and enhancing the robustness of our analysis.

In our study, we begin by employing the Fama and MacBeth (1973) regression framework to examine the relation between various asset characteristics and stock returns. This approach begins by assessing the impact of individual characteristics on returns, and then expands to analyze the combined effect of multiple characteristics. By doing so, we capture a detailed picture of how these signals interact to shape stock returns. Our analysis progressively controls for additional benchmark characteristics, such as market beta and key attributes such as the characteristics used to compute the factors in Fama and French (1993) and Fama and French (2015). We also employ a suite of machine learning algorithms, such as the standard LASSO proposed by Tibshirani (1996), non-parametric group LASSO approach proposed by Freyberger et al. (2020) and Random Forest, proposed by Breiman (2001). These methods allow us to discern which asset characteristics maintain their explanatory power under various regularization constraints and allows us to test non-linearity and variable importance.

In the last section, we shift focus to portfolio construction using the characteristics. First, we examine long-short portfolios through a single-sort method. We implement different sorting criteria to form both value-weighted and equal-weighted portfolios. To gauge the performance of these portfolios, we calculate CAPM alphas. Second, we employ predictive models to estimate the expected returns of each asset, using a 60-month rolling window. This method allows us to establish long positions in assets with the highest expected returns and short positions in those with the lowest. The success of this strategy is evaluated by contrasting these expected returns with the actual returns generated by the assets, thus offering an insight into the effectiveness of our long-short strategy in reflecting the influence of individual stock characteristics.

To check the robustness of our findings, we perform Fama-McBeth analysis on a diverse range of sub-samples, dividing the sample by size and only Ibovespa member companies. Changing the subsamples did not substantially change the conclusions, with the exception of small firms, that are particularly challenging to explain. Overall, our analysis of the Brazilian equity market, through both single and multi-characteristic, underscores important roles of liquidity and price, with some accounting metrics explaining some sub-samples. These results are further corroborated by machine learning techniques, with certain economic variables consistently significant across models, particularly those related to liquidity, momentum, and volatility.

Transitioning from model-specific results to portfolio performance, we observe that portfolios increase in average annual returns and volatility as the sorting becomes more extreme (from median to quintile). This implies that portfolios with extreme characteristic values enhance performance metrics. However, equal-weighed portfolios have, in general, higher returns and higher volatility, that could be attributed to the higher influence of smaller stocks, which tend to exhibit higher volatility and are given equal prominence in such portfolios. Most portfolios had negative CAPM alphas, indicating returns below market levels after adjusting for risk. However, some, like 12-month momentum-based portfolios, had positive alphas in certain sorts, suggesting potential excess market returns when adjusted for risk.

The long-short portfolios analysis corroborates previous findings of the paper. Models that jointly select characteristics generally outperform single selection of characteristics. Our best performing model for selection of characteristics is the Random Forest, with its characteristics generating an annualized return of 20% and Sharpe Ratio of 1.30. This model included 1-, 6- and 12-month momentum, return volatility and change in returns. Overall, models that select more price-based characteristics tend to outperform other selection methods.

The present paper is organized as follows. Section 2 provides a detailed description of the data used, including the data cleaning processes employed to ensure accuracy and reliability. Section 3 is dedicated to our methodology, outlining the analytical techniques and approaches we have adopted. Section 4 presents the empirical analysis, where we delve into the results and interpretations derived from our data. Lastly, Section 5 concludes the paper, summarizing our key findings and offering insights into their implications.

2 Data description

To construct monthly stock characteristics, we get data from the Eikon Refinitiv database, focusing on the universe of stocks domiciled in Brazil for the period extending from the year 2000 to June 2023. This dataset incorporates both actively traded securities and those that have been delisted at any point within the specified time-frame. To avoid forward looking bias, we use data as it was originally published, ignoring post-publication data amendments [1]. All accounting related data is yearly and set to end-of-year, while all price and volume related data

[1] Eikon allows us to select variables ”As reported”, which is the first released information.

is daily.

Subsequent to data aggregation, we implement a liquidity filter across the entire cohort of firms to refine our sample. Specifically, for each month in the study period, a company is deemed eligible for inclusion if it meets the following criteria over the trailing 12-month period: the ticker must have been traded on more than 80% of the eligible trading days; the average daily trading volume must have exceeded 1 million reais; and the median daily trading volume must have surpassed half a million reais. Additionally, the security must have been listed for a minimum duration of six months and must have been available for trading in the month immediately preceding the evaluation. This liquidity filter serves to ensure that the stocks included in our analysis exhibit sufficient market activity to warrant empirical investigation.

We initially selected 24 characteristics from the international literature, which are listed in Table 1 with details of how it is calculated and the authors of the underlying academic study. Table 2 report some descriptive statistics for the computed firm characteristics across the whole sample, along with statistics for the monthly stock returns. Figures 1 and 2 visually display the dispersion observed in the firm characteristics. For instance, asset growth has a right-skewed distribution, indicating that while most firms have low to moderate asset growth, there's a tail of firms with very high growth. *earn_pr* (earnings to price ratio) shows a concentration of values around zero but with a significant spread, implying a varied landscape of company earnings relative to their stock prices. Volume measures like *vlm* (trading volume) show a right-skewed distribution, indicating most stocks have lower volume, with fewer stocks having high volume. Importantly, these visualizations can help identify the nature of the distributions, whether they are somewhat normally distributed, skewed, or have heavy tails.

2.1 Data cleaning and standardization

In finance, it is common to encounter extreme values due to various reasons, such as extreme market events, data entry errors, or other anomalies. These extreme values can skew the results and lead to misleading conclusions. In our analysis we separate this into *errata* and *anomalies*. Errata are extreme values that are incorrect and could potentially lead our analysis to wrong conclusions. These values should be disregarded. After cleaning the data of potential errata, we address anomalies. Anomalies, while potentially correct, are excessively extreme and can disproportionately influence our statistical procedures.

To identify errata, we cross-check extreme values with data from reliable information providers. For accounting data, we refer to investor relations websites, and for other data types, we consult sources like Economatica, Bloomberg, and Yahoo Finance. We found that accounting characteristics seem to exhibit more extreme values. Liquidity, return, and volatility characteristics also have extreme values but in smaller quantity. We were not able to identify a clear pattern. The problem is not exclusive to some tickers, characteristic or time period.

Unable to find a pattern in the errata, we adopted a quantitative method to filter out incorrect data. We analyzed each characteristic across the entire dataset, focusing on the central portion of our data distribution by setting specific quantile thresholds. From this subset, we calculated the mean and standard deviation, retaining only values within the mean ± 10 standard deviations. This method, using data inside 1-99% quantiles to compute the mean and standard deviation, excluded only 0.3% of our dataset, or 2,600 observations.

Following the exclusion of erratas, and as guided by Green et al. (2017), we applied winsorization and standardization to our dataset. Winsorization involved capping data at the 1st and 99th percentiles. After capping, we normalized the data to a mean of zero and a standard deviation of one, aiming to reduce the impact of outliers while preserving statistical properties.

We employ a two-step procedure: first excluding erratas and then winsorizing/standardizing. This is essential because incorrect data must be removed outright. For instance, we encountered an asset growth value incorrectly recorded as over 8 million percent, while the actual data showed a decrease in total assets. Winsorization alone would replace this erroneous value with a large positive number, which is misleading.

Finally, for missing values, we chose a practical approach by substituting them with zero. This allows us to utilize multiple characteristics in our models without omitting significant data.

3 Methodology

3.1 Fama-MacBeth

Our goal is to test which asset's characteristics are related with returns. For that, we begin with the Fama and MacBeth (1973) procedure. First, consider the case of single-characteristic, and then we extend for multiple characteristics. Consider a characteristic s for company i defined at each time period t as c_{sit} (for example book-to-market ratio). Fix a time period (i.e. one month) and check if the characteristic is relevant for explaining that particular cross-section of returns. The specification is:

$$R_i = \alpha + \lambda C_{si} + u_i \tag{1}$$

Where we omit the time-subscript to make it clear that the regression is in the cross-sectional dimension, returns R_i are in t and characteristics C_{si} are in $t-1$. This is calculated for each period of time to obtain vectors $\vec{\alpha} = \{\alpha_1, ..., \alpha_T\}$ and $\vec{\lambda} = \{\lambda_1, ..., \lambda_T\}$. We will check the stability of these parameters across our cross-sections by estimating:

$$\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T} \alpha_t \qquad \text{and} \qquad \hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \lambda_t \tag{2}$$

And, to make inference, we use the standard deviation of our estimates to compute the standard error of our estimated coefficients $\hat{\alpha}$ and $\hat{\lambda}$:

$$\sigma^2(\hat{\alpha}) = \frac{1}{T^2}\sum_{t=1}^{T}(\alpha_t - \hat{\alpha})^2 \qquad \text{and} \qquad \sigma^2(\hat{\lambda}) = \frac{1}{T^2}\sum_{t=1}^{T}(\lambda_t - \hat{\lambda})^2 \tag{3}$$

Which allows us to compute inference statistics[2]. If the characteristic is related with returns we expect evidence that $\hat{\lambda} \neq 0$.

This setting is easily extended for multiple characteristics. For a given period of time and a set of characteristics S, the multi-characteristic specification is:

$$R_i = \alpha + \sum_{s=1}^{S} \lambda_s C_{si} + u_i \tag{4}$$

Now, for each characteristic s we obtain its vector $\vec{\lambda_s}$ of estimates across cross-sections and compute the mean and standard error in the same way as described for the single-characteristic setting.

The multi-characteristic framework is implemented in five distinct specifications. Initially, we analyze returns by incorporating 22 characteristics simultaneously[3]. In the second specification, we examine each of the characteristics alongside market beta to assess the impact of the additional characteristic in a CAPM setting. In the next three specifications we use the characteristics employed by Fama and French for constructing factors in Fama and French (1993) and Fama and French (2015). Therefore, the third specification integrates market beta, book-to-market ratio, and size. In the fourth, we expand this to include asset growth and operating profitability. The fifth specification extends the 5-factor approach and adds 12-month momentum.

[2] For more information about this procedure, check Cochrane (2005), chapter 12.

[3] 6- and 12-month momentum have two different specifications: including or excluding the last preceding month. This makes both measures highly correlated which can cause issues in statistical analyses due to multicollinearity. To avoid that, in this section and whenever we consider all characteristics together, we only use 6- and 12-momentum measures that exclude the previous month.

As robustness checks, we analyze subsamples of our data. We make subsets on size, where we separate the sample by terciles to get small, medium and large firms from our data. We also use a filter to include only firms that are part of the Ibovespa index.

Two empirical considerations made us remove some cross-sections from our analysis. First, some cross-sections, specially in the beginning of the sample, have very few assets. To avoid weighting small cross-sections in our procedure, we only consider cross-sections with at least 20 assets. Second, recall from the data section that we replaced missing data by zero to be able to run models with multiple characteristics without omitting significant data. However, if a variable is missing for all companies for a given cross-section, this means we input zero to all companies (no variability). This second case is handled by a filtering process. We check if the cross-section has variability by calculating the variance. If the variance is zero for any of the model variables, we ignore that cross-section from the estimates.

3.2 Characteristic selection via machine learning

In addition to the conventional Fama and MacBeth (1973) methodology, we employ a suite of machine learning algorithms to identify the characteristics that account for asset returns. Specifically, we apply a LASSO (Least Absolute Shrinkage and Selection Operator) regression, which is the standard penalized regression technique proposed by Tibshirani (1996). We also apply the non-parametric group LASSO approach of Freyberger et al. (2020) to discern which asset characteristics maintain their explanatory power under non-linear regularization constraints. Additionally, we also apply Random Forest models and conduct a variable importance analysis to isolate the most influential predictors of returns.

3.2.1 Linear regularization

LASSO is a penalized regression technique that aims to improve the predictive performance and interpretability of regression models by adding regularization terms to the loss function. Its objective function is to minimize $||Y - X\beta||_2^2 + \varphi||\beta||_1$, so it adds an L_1 penalty term (given by $||\beta||_1$), proportional to the absolute value of the regression coefficients, to the ordinary least squares (OLS) loss function. The φ parameter controls the strength of the regularization, effectively driving some coefficients to zero as it increases. In our setting, we choose the φ that minimizes the Bayesian information criterion (BIC).

Unlike the Fama and MacBeth (1973) approach, which estimates separate cross-sectional regressions for each time period, our methodology employs the regularization technique on the entire panel of standardized returns and characteristics at once. Importantly, our model does not incorporate any provisions to account for time-specific effects or interactions between time and asset characteristics. While this approach forgoes the exploitation of time dependency inherent in the Fama and MacBeth (1973) method, it enables us to rigorously examine the importance of various asset characteristics in explaining returns by examining which characteristics remain important even in a penalized regression context.

3.2.2 Non-parametric regularization

In addition to the standard regularization procedures, we also implement the non-parametric group LASSO approach of Freyberger et al. (2020)[4]. They use the group LASSO procedure developed by Huang et al. (2010) for estimation and to select those characteristics that provide incremental information for expected returns, that is, for model selection. Below, we provide a brief explanation of their procedure.

For each characteristic s, define $\tilde{C}_{s,it-1}$ as the rank transformation of $C_{s,it-1}$, mapping the cross-sectional distribution of the characteristic to the unit interval, such that $\tilde{C}_{s,it-1} \in [0,1]$. It is demonstrated that a function $\tilde{m}_t$ exists and satisfies

$$\tilde{m}_t\left(\tilde{C}_{1,it-1},\ldots,\tilde{C}_{S,it-1}\right) = m_t\left(C_{1,it-1},\ldots,C_{S,it-1}\right). \tag{5}$$

Therefore, understanding the conditional mean function m_t is synonymous with knowing the transformed conditional mean function $\tilde{m}_t$, which we can estimate. Similar to portfolio sorting, interest does not lie in the actual value of a characteristic by itself, but in its rank within the cross-section.

The focus is on modeling returns as a function of characteristics, expressed as

$$R_{it} = \sum_{s=1}^{S} \tilde{m}_{ts}\left(\tilde{C}_{s,it-1}\right) + \epsilon_{it}, \tag{6}$$

where $\tilde{m}_{ts}(\cdot)$ are unknown functions and $\tilde{C}_{s,it-1}$ denotes the rank-transformed characteristic.

The method employs the group LASSO to non-parametrically estimate the functions $\tilde{m}_{ts}$, setting functions associated with non-predictive characteristics to zero. This process enables model selection, distinguishing between constant and non-constant functions $\tilde{m}_{ts}$.

The procedure views portfolio sorts as estimating $\tilde{m}_{ts}$ by a constant within each portfolio and partition the support of each characteristic into L intervals, setting the interval endpoints

[4] Our sincere appreciation goes to the authors for providing us with the replicable code.

to the quantiles of the transformed characteristic distribution. This approximates each function $\tilde{m}_{ts}$ with a quadratic function over each interval, ensuring the entire function's continuity and differentiability over $[0, 1]$, hence approximating $\tilde{m}_{ts}$ by quadratic splines. The estimator is thus seen as a refined version of portfolio sorts, where $\tilde{m}_{ts}$ is approximated by a linear combination of $L + 2$ basis functions:

$$\tilde{m}_{ts}(\tilde{c}) \approx \sum_{k=1}^{L+2} \beta_{tsk} p_k(\tilde{c}) \tag{7}$$

Here, $p_k(\tilde{c})$ are known functions, and β_{tsk} are parameters to be estimated, with L being a user-specified smoothing parameter.

The adaptive group LASSO they propose consists of two steps. Initially, estimate coefficients as

$$\tilde{\boldsymbol{\beta}}_t = \underset{b_{sk}:s=1,\dots,S;k=1,\dots,L+2}{\arg\min} \sum_{i=1}^{N} \left(R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k \left(\tilde{\boldsymbol{C}}_{s,it-1} \right) \right)^2 + \theta_1 \sum_{s=1}^{S} \left(\sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}, \tag{8}$$

where $\tilde{\boldsymbol{\beta}}_t$ is an $(L + 2) \times S$ vector of estimates and θ_1 is a penalty parameter. This equation includes the sum of squared residuals and the LASSO group penalty function, penalizing coefficients associated with a given characteristic. They set entire expansions of $\tilde{m}_t$ to zero when a characteristic is not incrementally informative for expected returns. They follow Yuan and Lin (2006) to select θ_1 based on minimizing the Bayesian information criterion (BIC).

However, the initial LASSO step may select too many characteristics, including those without predictive power. To rectify this, introduce characteristic-specific weights in the second step, adjusting the group LASSO penalty function based on first-step estimates.

We apply their approach to our full sample and for a range of values for L (knots) in order to select which characteristics provide incremental information for expected returns.

3.2.3 Random Forest

The Random Forest algorithm operates by constructing a multitude of decision trees at training time, and then outputting the mean prediction from each tree. The algorithm first randomly selects n samples from the dataset with replacement (where n is the size of the dataset), giving us many bootstrap samples. Then, for each bootstrap sample it grows a decision tree and, at each tree node, it randomly selects m features without replacement (where m is a hyperparameter smaller than the total number of features) and splits the node using the feature that provides the best split (e.g., maximum information gain or minimum impurity) among the m selected features. Finally, to make a prediction it takes the average prediction of all individual trees.

One of the advantages of Random Forests is that it allows us to compute variable importance, which works by evaluating the impact of each feature on the predictive accuracy of the model. Specifically, variable importance is assessed through a permutation-based procedure. During the training phase, each decision tree is built using a bootstrap sample, leaving out a subset of data known as ”Out-of-Bag” (OOB) samples. The baseline OOB error is initially computed by aggregating the prediction errors for each observation across the trees for which it serves as an OOB sample. To gauge the importance of a given feature X_i, its values are randomly permuted in the OOB samples, while other variables remain unchanged. The model's performance is then re-evaluated using this permuted data to obtain a new OOB error rate. The importance score for X_i is calculated as the difference between the permuted OOB error and the baseline OOB error, averaged across all the individual trees in the Random Forest ensemble. A higher importance score means that the feature is more critical for the model's predictive accuracy, indicating its relevance in explaining asset returns.

In line with standard machine learning practices, we partition our dataset into training, validation, and testing subsets. The training set spans the years 2000 to 2012, the validation set covers 2013 to 2015, and the testing set encompasses 2016 up to the end of the available data. Initially, we construct a Random Forest model using the training set, experimenting with various hyperparameters—specifically, the number of randomly selected candidate variables at each tree split, the total number of trees in the ensemble, and the minimum number of observations required for a node split. Performance is assessed by predicting outcomes in the validation set and calculating the Root-Mean-Squared-Error (RMSE) for each set of hyperparameters. The hyperparameter configuration yielding the lowest RMSE on the validation set is then employed to evaluate the model on the testing set and to compute variable importance. This selection of parameters using the validation set is an 'out-of-sample' way to mitigate risks of overfitting the data.

3.3 Alpha in single sorted portfolios

To further assess the importance of individual characteristics, we construct long-short portfolios using single-sort methods and evaluate their subsequent performance. Each month t, we sort stocks based on their characteristics from the preceding month $t-1$. We adopt a strategy that goes long on stocks in the top group, as determined by their $t-1$ characteristic, and short on those in the bottom group. Notably, for the characteristics illiquidity, idiosyncratic return volatility (i_ret_v), and return volatility (ret_v), we reverse the strategy by going short on stocks in the top group and long on those in the bottom group, so we adopt the same direction. Sorting thresholds are established using median, tercile, and quintile divisions, allowing us to construct both value-weighted and equal-weighted portfolios for comparison.

To test these portfolios, we obtain CAPM alphas by estimating the regression:

$$R_{it}^{e} = \alpha_i + \beta_i R_t^{e} + e_i \tag{9}$$

Where R_{it}^{e} and R_t^{e} are the excess return of the asset and the market, respectively. The market portfolio is computed as the value-weighted return of all companies available in our analysis minus monthly selic rate.

3.4 Translating expected return models into portfolios

Here, we propose a comprehensive way to translate the results from the characteristics analysis into investment portfolios in order to evaluate their performance. For each month, we have estimation, portfolio formation and evaluation steps. First, in the beginning of month t, we estimate Equation 10 based on a 60-month rolling window, where C^* represent sets of possible characteristics included in the predictive regression. This section considers three approaches in determining the members of C^* which are discussed in the following subsections 3.4.1 and 3.4.2. Figure 3 contains a diagram of the procedure implemented.

$$R_{i,t+1} = \alpha_{i,t} + \sum_{c \in C^*} b_{c,i,t} C_{i,t}^{*} + \varepsilon_{i,t}. \tag{10}$$

Second, after the estimation, Equation 10 allows us to compute next month expected returns for each asset, denoted as $\{E_t(R_{t+1}^{1}), ..., E_t(R_{t+1}^{N})\}$. We form portfolios by going long stocks in the top tercile of expected returns while shorting stocks in the bottom tercile. We hold such portfolio for 1 month. Third, at the end of the month, we compute realized returns from our portfolio.

Creating a universal strategy that accommodates the diversity in data and is effective across all models is a complex task. Our approach has its constraints, as it overlooks the covariances between strategies for portfolio construction, disregards transaction costs, and assumes cost-free shorting even for potentially illiquid firms. Despite these limitations, our aim is to provide clarity on the findings of our analysis and maintain a straightforward method that is applicable to various models and easy to comprehend.

3.4.1 Traditional models and selection of characteristics

Here, the set of selected characteristics is composed by a base set of characteristics plus individually selected characteristics, corresponding to the Fama-Macbeth framework outlined in subsection 3.1. The base set of characteristics can be those related to the single-characteristic, CAPM, Fama-French 3, Fama-French 5 and Fama-French 5 + momentum models.

Particularly, in the case of the single-characteristic framework, each characteristic is individually tested for significance in the Fama-MacBeth regression, and those significant at the 5% level are then the single element of C^*. For each C^*, we then compute the predictive regression, form the long-short portfolio and compute returns, giving us the single characteristic portfolio. Then, the final portfolio is an equal-weighted average of all single characteristic portfolios.

For the other models, we keep the model characteristic's fixed. In the case of a traditional CAPM model, for example, the market beta is the base characteristic and each characteristic is evaluated with regards to the additional information relative to market beta in the Fama-MacBeth regressions. Then, the characteristics that show statistical significance at 5% are added to C^*, together with market beta. Finally, in the predictive regressions, the base characteristics are kept fixed, and a single individual characteristic is cycled through, conditional on being statistically significant in the characteristic analysis, giving us multiple predictive regressions. Then, the final portfolio is an equal-weighted average of all the base + individually selected characteristic portfolios.

3.4.2 Joint selection of characteristics

Finally, this case corresponds to the multi-characteristic framework where all characteristics are tested jointly in the Fama-MacBeth regression and the machine learning approaches, outlined in subsections 3.1 and 3.2 respectively. Here, C^* is composed by all the significant characteristics, which all enter the predictive regression and end up forming a single long-short portfolio, without a need for averaging.

4 Empirical analysis

4.1 Single-characteristic Fama-MacBeth

In this analysis, we dissect the predictive power of individual financial characteristics across various segments of the Brazilian equity market. Our investigation is detailed in two panels within the accompanying Figure 4: Panel A presents a heatmap indicating the significance levels of characteristics across a spectrum of t-value thresholds for subsets of all firms, as well as segmented by company size (big, medium, small) and Ibovespa firms. The significance is mapped against t-value thresholds ranging from > 1.96 to > 4.00, illustrating the robustness of each characteristic's predictive ability. This approach highlights the characteristics that are most influential at varying levels of statistical confidence. Panel B displays the coefficient values. The coefficient values presented in Panel B should be interpreted as standard deviations since our dataset is standardized. This normalization allows for an equitable comparison across all characteristics. Notably, there is a degree of parameter stability observed, especially among significant parameters, indicating a consistency in the direction of influence across different market segments.

When considering the totality of firms, ten characteristics stand out as significant at the 5% level, predominantly related to price and volume rather than accounting metrics. These include earning profitability, idiosyncratic return volatility, illiquidity, 12-month momentum (with and without excluding the most recent month), 6-month momentum (with and without excluding the most recent month), return volatility, sales-to-price ratio and size. Tightening our t-value threshold to > 3.00, eight characteristics remain significant, excluding sales-to-price and size, indicating their strong relevance in explaining the cross-section of returns individually.

Among these characteristics, illiquidity and 12-month momentum (both with and without lag) are particularly prominent. Illiquidity achieves significant coefficients in all subsamples analysis, in some case with t-values exceeding 4.00. The 12-month momentum is significant for samples that include all assets with, again with t-values exceeding 4.00. These results underscore the critical role of illiquidity and momentum in influencing returns, reaffirming the importance of this market dynamic.

The significance of characteristics varies when analyzed by size terciles and Ibovespa firms, with an average of 3.25 characteristics being significant at the 5% level. For big and medium companies results are the same: size, volume and illiquidity not only remain significant but exhibit strong statistical significance with t-values above 4.00. Smaller firms show a narrower set of significant characteristics, with size and illiquidity being notable. Interestingly, for Ibovespa companies, 6-month momentum characteristics and illiquidity are exclusively significant, reflecting perhaps the market's emphasis on trend-following behaviors for these firms.

Despite the parameter stability for significant coefficients, size and illiquidity measures display sign changes across different subsamples. These variations necessitate a nuanced approach when considering these characteristics as part of investment strategies.

It is important to acknowledge the limitations within our findings. Notably, the beta is consistently non-significant with negative coefficients across all subsets, challenging the conventional capital asset pricing model (CAPM) perspective within the Brazilian market context. This absence of significance indicates that beta may not be a sufficient measure of risk in this market or that its pricing differs across market segments.

In sum, our single-characteristic analysis elucidates the differential impact of financial characteristics on equity returns within the Brazilian market. The results obtained provide a better understanding of market behavior, which can be useful for both investors and academics in developing asset pricing models and investment strategies.

4.2 Multi characteristic Fama-MacBeth

The main motivation behind the multicharacteristic version of the Fama-Mcbeth analysis is to check if a characteristic has explanatory power over stock returns, even when controlling for the influence of other characteristics. This is an important notion since characteristics might appear important only because they are related with other characteristics that are actually the true drivers of returns. Independent determinants, however, are those that retain their statistical significance and influence even when this interdependence is accounted for. For every specification there is a table with panels A and B presenting the same information as in the single-characteristic analysis.

The CAPM model, using market beta as a benchmark characteristic, supports the findings related to single specific characteristic. The results in Figure 6 shows that accounting characteristics like the earnings-to-price ratio and operating profitability typically have a positive relationship with the returns across all stocks sub-samples. 6- and 12-month momentum indicators are linked to higher returns in many groups. Idiosyncratic and return volatility, volume and volatility of liquidity generally have a negative relation with returns. When we set a stricter criterion for significance (t-value $>$ 3.5), illiquidity, 6- and 12-month momentum, size and volume remain important. This model, by including systematic risk, highlights certain characteristics that were not significant when only one characteristic was analyzed.

When we set market beta, book-to-market and size as benchmark variables, the results shown in Figure 7 are very similar to the CAPM specification, with previous liquidity and volatility signals playing an important role. Tightening the t-value threshold to $>$ 3.5, only illiquidity (for small and big stocks) and 12-month momentum (for all stocks) remain significant.

Further extending our model to include asset growth and operating profitability as benchmark variables, shown in Figures 8 and 9, we observed a consistent pattern from the previous specification. Moreover, when we add the 12-month momentum in the model (Figure 9), momentum characteristics are not significant anymore (as expected) and only illiquidity remains significant for t-values > 3.5, for the small stocks subsample.

In an additional exercise, we analyzed all 22 characteristics simultaneously, as illustrated in Figure 5. In this analysis, size emerges as the most influential factor, surpassing even the 12-month momentum in significance. While idiosyncratic and return volatility, illiquidity and volume continue to play significant roles, their impact varies across different subsamples compared to previous findings. By imposing a stricter t-value threshold of >3.5, size is the only characteristic that retains its significance. Overall, incorporating all characteristics in this specification diminishes the individual statistical significance of many traits.

Finally, single- and multi-characteristic analyses of the Brazilian equity market reveal key similarities and differences. Both studies highlight liquidity and momentum, especially the 12-month momentum and illiquidity as significant influencers of stock returns, while challenging the explanatory power of market beta in this context. Differences emerge in the scope of the accounting as book-to-market and asset growth that start to play a role only in the multi characteristic analysis when we consider the benchmark models.

4.3 Variable selection via machine learning

In this section, our focus shifts towards variable selection rather than estimating risk premiums associated with certain characteristics. We apply the three methodologies as outlined in Section 3.2: LASSO, non-parametric group LASSO and Random Forest.

The LASSO methodology reveals that even under stringent parameterization, certain economic variables remain significant, suggesting their substantial impact on asset pricing, in line with finance theories. This conclusion is drawn from an analysis depicted in Figure 10, which shows the estimated coefficients for the parametrization the minimizes the BIC. Out of 21 coefficients, only nine are set to values different to zero, indicating a stringent penalization. Only asset growth, book-to-market, leverage, sales-to-price, idiosyncratic return volatility (*i_ret_v*), illiquidity, 1-month momentum, 12-month momentum, volatility of liquidity (*std_vlm*) are set to values different to zero. This restricted model underscores the importance of these variables, highlighting a significant influence of liquidity and trading volume, as well as short-term and intermediate-term momentum effects (1-month momentum and 12-month momentum), and a value and asset effect. These aspects are consistent with established finance principles, particularly regarding the roles of liquidity, momentum, and value in influencing asset prices.

The non-parametric group LASSO approach highlights that volatility, and momentum are the predominant factors in explaining stock returns. This conclusion is drawn from the analysis presented in Table 3, where we employ the methodology of Freyberger et al. (2020) with varying numbers of knots (L). Overall, five different variables are selected in this approach, which are idiosyncratic return volatility (i_ret_v), 1-month momentum, 12-month momentum, 60-month momentum, and price delay, corresponding to volatility and momentum. For higher values of knots in the spline (L), the non-parametric group LASSO does not select any variable, indicating a more complex selection process within the non-parametric group LASSO framework when compared to the standard LASSO. In general, we can see that 12-month momentum is the one that is consistently selected up to 10 knots, giving us an indication that there is a momentum effect in the Brazilian market.

The variable importance analysis from our Random Forest model reaffirms that characteristics associated with momentum, volatility, and volume are crucial in explaining stock returns, with size also emerging as a significant factor. In Table 4, we outline the Random Forest model parameters selected during the validation phase, along with training and test statistics, and Figure 11 graphically represents the variable importance from the trained model. It is noteworthy that the variables prioritized in the Random Forest analysis closely align with those identified by linear regularization methods. Specifically, the variables that exhibited larger coefficients in linear models, similarly demonstrate high importance in the Random Forest context. However, we see a disparity between variable importance in the Random Forest model and the characteristics selected in the non-parametric group LASSO approach. In both cases, 1-month momentum and 12-month momentum are relevant, but 60-month momentum, idiosyncratic return volatility (i_ret_v) and price delay, which are selected in the non-parametric group LASSO, do not have high importance in our Random Forest model, even though both approaches take into account non-linearities between characteristics. This difference could be because of the different ways in which each model operates. Non-parametric group LASSO applies a penalty to coefficients based on their significance, which can lead to a different subset of variables being selected compared to Random Forests. Random Forests do not explicitly penalize model complexity; rather, they reduce overfitting through ensemble learning. This difference in approach to regularization and variable selection might result in different variables being highlighted as important.

4.4 Alpha analysis for single-sorted portfolios

Even though the focus of this paper is on characteristics, and not factors, we build characteristic based long-short portfolios to evaluate the performance of the characteristics over time. As described previously, we build portfolios based on median, terciles and quintiles sorts, and weight them by market-value or equally. Portfolios go long on stocks in the top group, as determined by their $t-1$ characteristic, and short on those in the bottom group, and the opposite for characteristics illiquidity, idiosyncratic return volatility (i_ret_v), and return volatility (ret_v).

Tables 5 and 6 display some descriptive statistics of the monthly return of value-weighted and equal weighted portfolios, respectively. In the value-weighted portfolios, as we move from median to tercile, and then to quintile sorts, there is a noticeable increase in both average annual returns and volatility. This trend suggests that portfolios based on more extreme characteristic values tend to exhibit amplified performance metrics. In contrast, the equal-weighted portfolios show a similar trend but with generally greater volatility, indicating a small stock effect inherent in equal weighting. Notably, in both weighting schemes, characteristics like leverage and 12-month momentum consistently show strong performance across all sorts.

Max drawdown tends to increase with the shift from median to tercile and quintile sorts, for both value and equal-weighted portfolios. For instance, leverage and 12-month momentum in both value and equal-weighted portfolios increase their max drawdown when moving from median sorts to more extreme sorts. Besides that, we can see that equal weighed portfolios in general have a more severe max drawdown, again suggesting higher risk due to small stock effect. Regarding cumulative returns, displayed in Figures 12 and 13, portfolios like leverage and 12-month momentum in value-weighted schemes and earnings-to-price, leverage, sales-to-price, 6-month momentum and 12-month momentum in equal-weighted schemes show impressive cumulative returns, particularly in tercile and quintile sorts. This highlights the potential for higher gains in portfolios sorted based on more extreme characteristic values.

Figures 14 and 15 display the distribution of monthly portfolio returns, and along with Tables 5 and 6 we can see that skewness and kurtosis metrics in both value-weighted and equal-weighted portfolios reveal significant insights into the distribution of returns. In value-weighted portfolios, we observe a range of skewness values across different sorts, indicating varied asymmetry in return distributions. For instance, portfolios like sales growth and operating profitability exhibit high positive skewness, suggesting a longer right tail with potential for occasional large positive returns. On the other hand, characteristics such as volatility of liquidity (*std_vlm*) and idiosyncratic return volatility (*i_ret_v*) display negative skewness, indicating a propensity for negative returns. Kurtosis values in these portfolios also vary, with some portfolios showing high kurtosis (e.g., volatility of liquidity (*std_vlm*) and sales growth in value-weighted portfolios), indicating a higher likelihood of extreme returns compared to a normal distribution. This implies a higher risk of encountering significant positive or negative returns. In equal-weighted portfolios, the trend in skewness and kurtosis remains somewhat similar, however, it is clear that there are more positively skewed portfolios than in the value-weighed case, indicating some benefits from diversification.

Lastly, in Table 7, the CAPM alphas for the various portfolios provide essential insights into the risk-adjusted performance of these portfolios relative to the market. A notable trend across most portfolios is the presence of negative alphas, indicating underperformance against the CAPM benchmark. This suggests that, after accounting for market risk, the portfolios generally do not yield returns exceeding the market's, as per CAPM predictions. The statistical significance of these alphas, as evidenced by the t-statistics, further bolsters the reliability of this observation.

However, among the sea of negative alphas, a few portfolios stand out with positive alphas. Notably, 12-month momentum shows positive alphas in some sorts, indicating that this portfolio might be generating excess returns over the market when adjusted for risk. This exception is particularly interesting as it suggests the presence of certain stock characteristics that can potentially outperform the market, as per the CAPM model.

The variation in alphas across portfolios sorted by median, terciles, and quintiles highlights the influence of stock characteristic grouping on performance relative to the CAPM. For example, portfolios like those built on asset growth and volume consistently exhibit negative alphas across all sorting categories, indicating a regular pattern of underperformance. Moreover, a comparison between value-weighted and equal-weighted portfolios reveals a subtle but noticeable difference in the magnitude of negative alphas. The value-weighted portfolios tend to show slightly more pronounced negative alphas, suggesting a higher degree of underperformance compared to their equal-weighted counterparts.

4.5 Assessing performance of predictive portfolios

In assessing the effectiveness of various predictive strategies for portfolio formation, we observe a hierarchy in performance, shown in Table 9 and Figure 16. The Random Forest strategy emerges as the most superior, with a mean annualized monthly return of 20% and annualized Sharpe ratio of 1.30, indicating robust risk-adjusted returns. This model benefits significantly from the inclusion of momentum characteristics, specifically 1-month momentum, 6-month momentum, 12-month momentum, return volatility (*ret_v*), and change in returns (*ch_mom*). These features, primarily reflecting momentum, appear to be substantial drivers of the predictive success.

Closely following the Random Forest model is the Fama-French 5-factor model supplemented by 12-month momentum (FF5 + Mom), which shows the strength of incorporating momentum into traditional asset pricing models. With a mean annualized return of 18% and annualized Sharpe ratio of 1.02, it underscores the enduring relevance of momentum factors in market prediction.

The Non Parametric Lasso and Lasso models also highlight the prominence of momentum in their characteristic selections, delivering commendable performance metrics. Specifically, Non Parametric Lasso incorporates a blend of short and long-term momentum signals alongside price delay, while Lasso selects a mix of asset growth, book-to-market ratio, and various momentum factors.

It is noteworthy that strategies involving individual characteristic selection consistently underperform in comparison to models that jointly consider multiple characteristics, with FF5 + momentum being an exception. The linear progression in mean returns across models from Single-characteristic to FF5 + Mom indicates the added value of a multifaceted approach over a single-factor analysis.

The findings suggest that momentum is a predominant force across models, with its inclusion seeming to elevate the predictive power significantly. The monotonic increase in returns with the addition of more variables to the individual selection models substantiates this assertion. However, the simplicity of models like CAPM (Beta) and the single-characteristic framework, despite their lower performance metrics, serves as a reminder of the trade-off between complexity and interpretability in predictive modeling.

5 Conclusion

In this study, we embarked on a comprehensive examination of the Brazilian financial market to unravel the influence of firm characteristics on stock returns, employing a robust methodological framework that combines Fama-MacBeth regressions, advanced machine learning techniques such as LASSO, non-parametric LASSO and Random Forest analysis, and portfolio assessment strategies. Drawing from an extensive dataset derived from the Eikon Refinitiv database, our analysis filtered for liquidity and employed rigorous data cleaning processes to ensure the reliability of our findings.

Our findings reveal that price-related metrics (mainly 12-month momentum), liquidity measures, size, and volatility metrics, exhibit a more significant association with stock returns compared to accounting variables, albeit the latter demonstrate relevance in specific contexts. This underscores the peculiarities of the Brazilian market and its departure from patterns observed in more developed markets like the United States. By applying a variety of analytical techniques, we were able to isolate the characteristics that persist in their explanatory power, highlighting the roles of liquidity, momentum, and volatility.

Moreover, the portfolio analysis provided insights into the practical application of these characteristics, with value-weighted portfolios showing enhanced performance metrics through sorting and equally-weighted portfolios not providing diversification benefits. Despite a general trend of negative CAPM alphas, portfolios sorted on 12-month momentum exhibited positive alphas, suggesting avenues for excess market returns when adjusted for risk. These results are further corroborated by the long-short portfolios, where models that jointly selected multiple characteristics generally chose momentum and outperformed models with less weight on price-related metrics.

The contribution of this paper is manifold. Firstly, it enriches the understanding of the Brazilian financial market, a relatively underexplored terrain, by establishing a foundational dataset that is poised to facilitate further academic inquiry. Secondly, by testing the applicability of characteristics identified in the U.S. market within the Brazilian context, our study not only highlights market-specific idiosyncrasies but also validates the robustness of established financial theories across divergent market environments.

In conclusion, this research delineates a nuanced landscape of stock return predictors in Brazil, emphasizing the importance of price-related characteristics. While acknowledging the influence of accounting metrics, the study suggests that liquidity, momentum, and volatility related characteristics serve as more reliable indicators of stock performance in this emerging market. Looking forward, the dataset and findings presented herein offer possibilities for future research, encouraging a deeper exploration of country-specific risk factors.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5:31–56.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, 61:259–299.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9:3–18.

Barbee, J. W. C., Mukherji, S., and Raines, G. A. (1996). Do sales-price and debt-equity explain stock returns better than book-market and firm size? *Financial Analysts Journal*, 52:56–60.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, 32:663–682.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, 43:507–528.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Chordia, T., Subrahmanyam, A., and Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, 59:3–32.

Cochrane, J. (2005). *Asset Pricing*, volume 1. Princeton University Press, 2nd edition edition.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66:1047–1108.

Cooper, M. J., Gulen, H., Schill, M. J., Cliff, M., Daniel, K., Daniel, N., Denis, D., Easterwood, J., Ferson, W., Griffin, J., Gutierrez, R., Lemmon, M., Liu, L. X., Loutskina, E., Maxwell, B., McConnell, J., McQueen, G., Rau, R., Sagi, J., Simko, P., and Wurgler, J. (2008). Asset growth and the cross-section of stock returns.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116:1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81:607–636.

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

Gettleman, E. and Marks, J. M. (2006). Acceleration strategies. *SSRN Electronic Journal.*

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

Hou, K. and Moskowitz, T. J. (2005). Market frictions, price delay, and the cross-section of expected returns. *Review of Financial Studies*, 18:981–1020.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, 38.

Hwang, L.-S. and Lee, W.-J. (2013). Stock return predictability of residual-income-based valuation: Risk or mispricing? *Abacus*, 49:219–241.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, 45:881–898.

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

Lakonishok, J., Shleifer, A., and Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The Journal of Finance*, 49:1541–1578.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108:1–28.

Rosenberg, B., Reid, K., and Lanstein, R. (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management*, 11:9–16.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68:49–67.

Tables and figures

Table 1: List of firm characteristics

Acronym	Firm characteristic	Author(s)	Definition of the characteristic-based anomaly variable
asset_gr	Asset Growth	Cooper et al. (2008)	Annual percent change in total assets
btm	Book-to-market ratio	Rosenberg et al. (1985)	Book value of equity divided by end of fiscal year-end market capitalization
beta	Market beta		Estimated market beta from weekly returns and equal weighted market returns for 3 years ending month $t-1$ with at least 52 weeks of returns
ch_mom	Change in returns	Gettleman and Marks (2006)	Cumulative returns from months $t-6$ to $t-1$ minus months $t-12$ to $t-7$
vlm	Trading volume	Chordia et al. (2001)	Natural log of trading volume times price per share from month $t-2$
earn_pr	Earnings-to-price-ratio	Basu (1977)	Annual income before extraordinary items divided by end of fiscal year market cap
gt_pft	Growth	Novy-Marx (2013)	Revenues minus cost of goods sold divided by lagged total assets
i_ret_v	Idiosyncratic return volatility	Hwang and Lee (2013)	Standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end
ill	Illiquidity	Amihud (2002)	Average of daily (absolute return / volume)
levg	Leverage	Bhandari (1988)	Total liabilities divided by fiscal year-end market capitalization
mom1m	1 month momentum	Jegadeesh and Titman (1993)	1-month cumulative return
mom6m	6 month momentum	Jegadeesh and Titman (1993)	5-month cumulative returns ending one month before month end
mom12m	12 month momentum	Jegadeesh (1990)	11-month cumulative returns ending one month before month end
mom36m	3 years momentum	Jegadeesh and Titman (1993)	Cumulative returns from months t-36 to t-13
mom60m	5 years momentum	Jegadeesh and Titman (1993)	Cumulative returns from months-60 to t-13
sz	Size	Banz (1981)	Natural log of market capitalization at end of month $t-1$
op_pft	Operating profitability	Fama and French (2015)	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity
pr_delay	Price delay	Hou and Moskowitz (2005)	The proportion of variation in weekly returns for 36 months ending in month t explained by 4 lags of weekly market returns incremental to contemporaneous market return
ret_v	Return volatility	Ang et al. (2006)	Standard deviation of daily returns from month $t-1$
gr_sl	Sales growth	Lakonishok et al. (1994)	Annual percent change in sales
sl_pr	Sales-to-price	Barbee et al. (1996)	Annual revenue divided by fiscal year-end market capitalization
std_vlm	Volatility of liquidity	Chordia et al. (2001)	Monthly standard deviation of daily trading volume

Table 2: Characteristics descriptive statistics

Variable	Mean	Std	n.NA	pct.NA	Min	p25	Median	p75	Max	Skewness	Kurtosis
monthly_return	0.008	0.131	259	0.009	-1.078	-0.059	0.009	0.078	1.046	-0.146	5.124
asset_gr	0.219	0.455	817	0.030	-0.996	0.032	0.114	0.245	4.119	4.106	23.485
btm	0.700	0.840	1769	0.065	-5.636	0.283	0.529	0.910	7.023	1.382	15.258
sz	22.379	1.612	778	0.028	15.387	21.354	22.375	23.407	27.230	-0.010	0.368
op_pft	0.217	0.525	947	0.035	-3.873	0.056	0.180	0.317	4.250	0.809	18.682
earn_pr	0.029	0.247	1860	0.068	-2.244	0.019	0.053	0.099	2.390	-3.926	34.802
gt_pft	0.266	0.223	5483	0.201	-0.236	0.118	0.214	0.350	2.015	2.218	8.700
levg	1.929	3.477	1773	0.065	0.000	0.339	0.819	1.798	34.562	4.554	27.006
gr_sl	0.187	0.399	2425	0.089	-2.708	0.025	0.125	0.272	3.742	2.810	20.499
sl_pr	1.117	1.525	3119	0.114	-1.460	0.294	0.630	1.236	14.315	3.493	16.285
i_ret_v	0.049	0.022	5559	0.204	0.020	0.035	0.043	0.055	0.239	2.463	8.951
ret_v	0.026	0.014	81	0.003	0.000	0.017	0.023	0.030	0.144	2.568	10.890
beta	0.988	0.409	5522	0.202	-0.783	0.692	0.964	1.228	2.982	0.587	0.839
mom1m	0.007	0.132	316	0.012	-0.832	-0.064	0.003	0.074	1.543	0.647	6.796
mom6m	0.050	0.339	311	0.011	-0.999	-0.139	0.024	0.200	8.020	2.287	23.158
mom12m	0.151	0.633	538	0.020	-0.999	-0.195	0.058	0.362	12.566	4.208	44.049
mom36m	0.560	1.635	4833	0.177	-0.991	-0.145	0.239	0.803	50.730	10.008	185.545
mom60m	1.438	5.577	7953	0.291	-0.996	-0.071	0.574	1.580	332.309	28.480	1249.178
ch_mom	-0.035	0.535	741	0.027	-10.188	-0.273	-0.029	0.215	8.975	-0.777	26.124
pr_delay	0.009	0.161	5176	0.190	-2.910	-0.031	0.002	0.042	4.773	0.434	63.542
ill	0.000	0.000	647	0.024	0.000	0.000	0.000	0.000	0.000	4.471	24.950
std_vlm	21.174	40.902	679	0.025	0.000	2.334	6.593	20.957	425.098	4.386	25.162
vlm	16.452	1.857	363	0.013	6.894	15.133	16.453	17.755	23.682	0.030	0.065

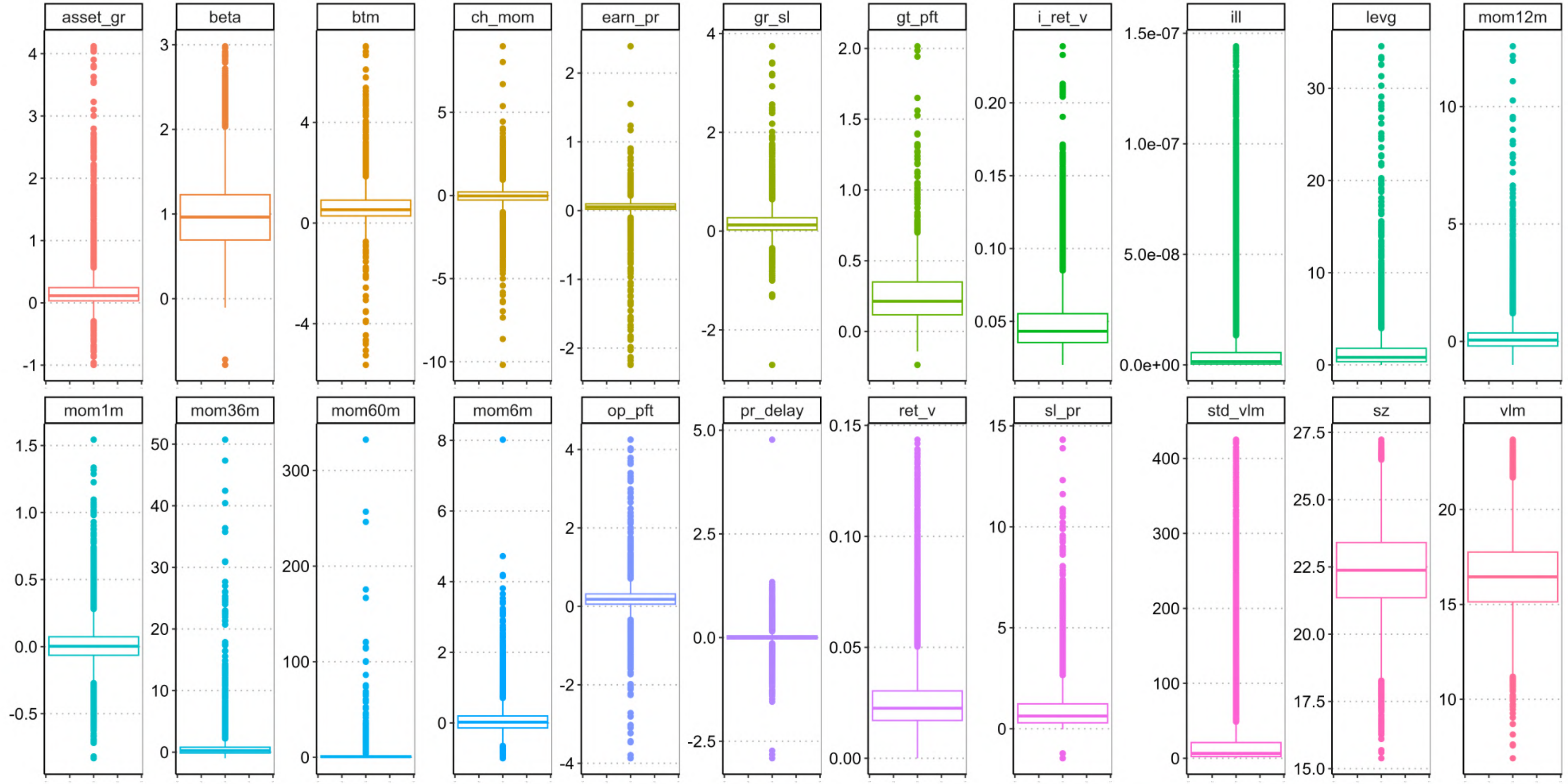


Figure 1: Boxplots of characteristics

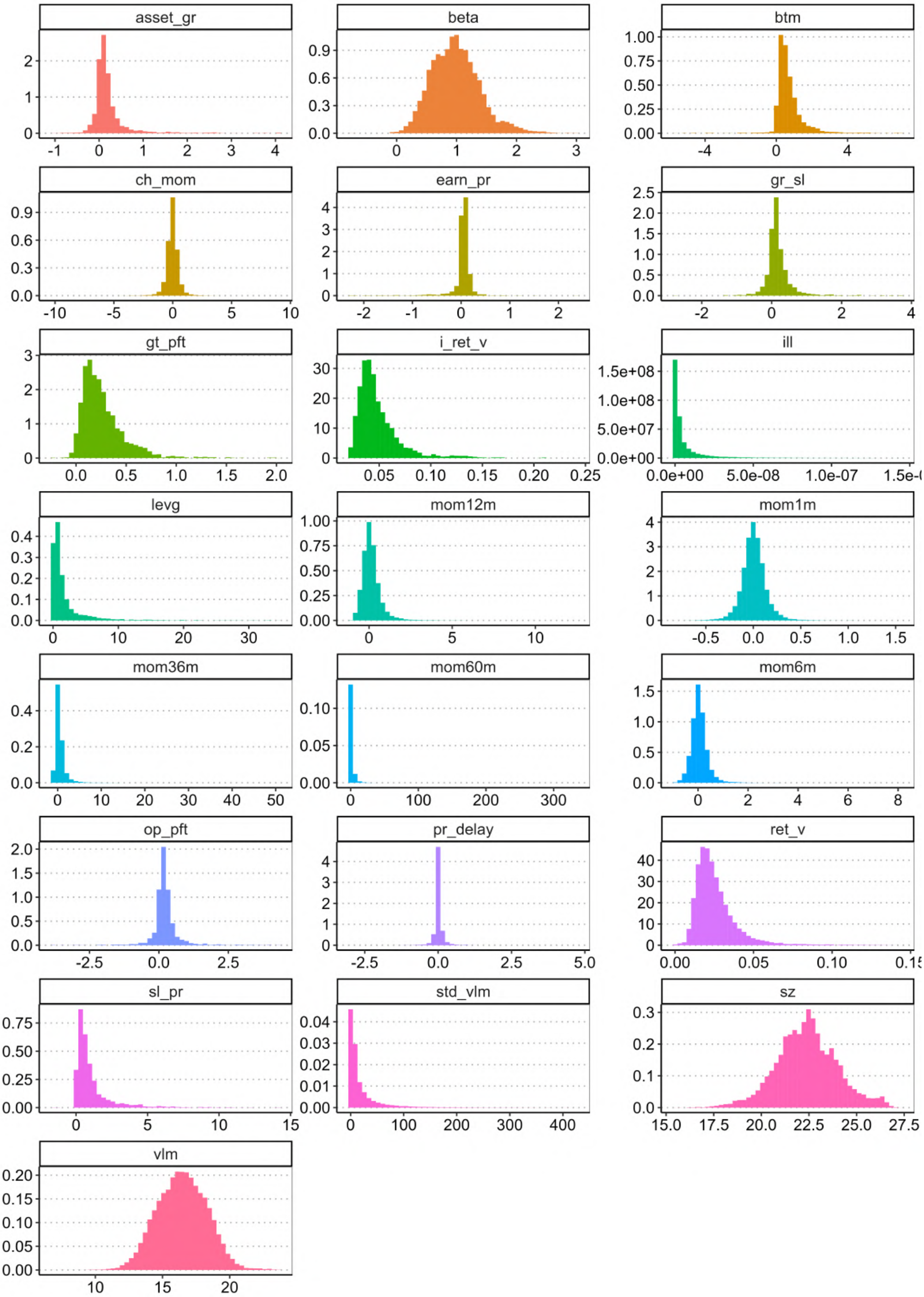


Figure 2: Histogram of characteristics

Diagram - Predictive Portfolio Modeling: Estimating Asset Returns for Long-Short Strategies

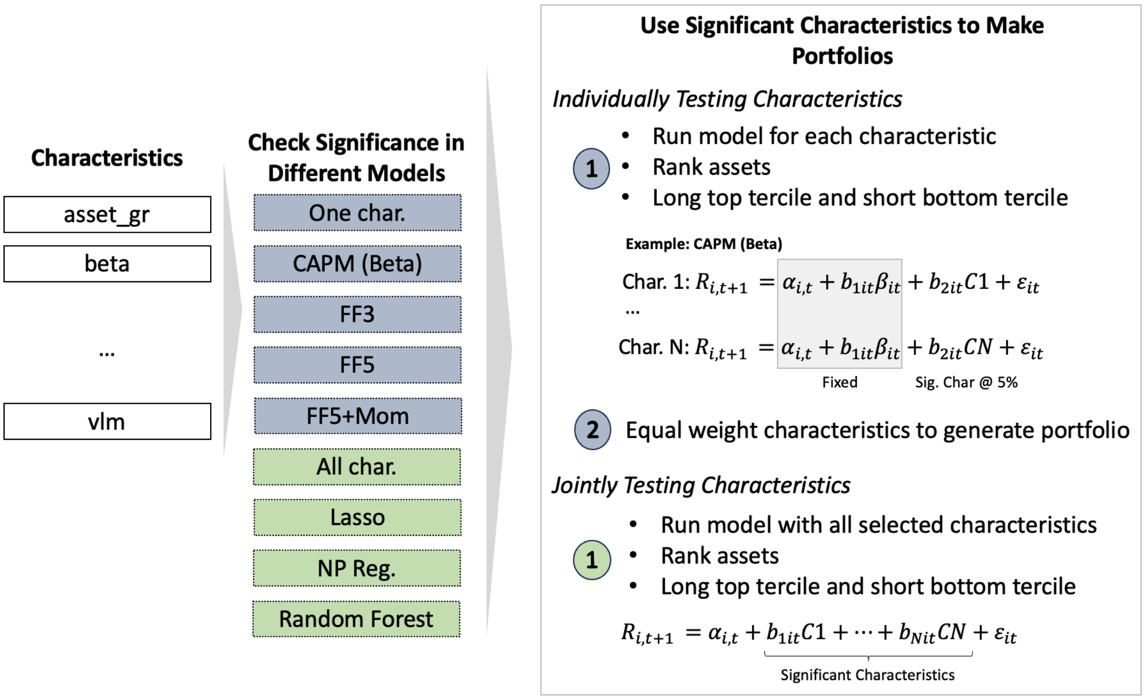
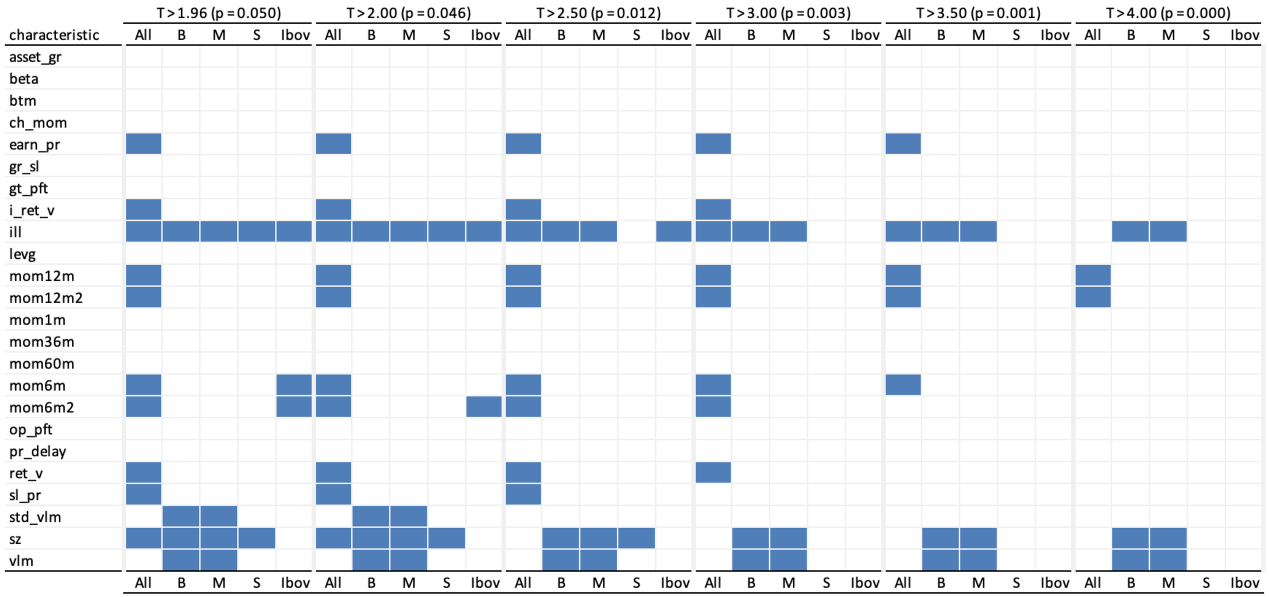

Characteristics
asset_gr
beta
...
vlm
Check Significance in Different Models
One char.
CAPM (Beta)
FF3
FF5
FF5+Mom
All char.
Lasso
NP Reg.
Random Forest
Use Significant Characteristics to Make Portfolios
Individually Testing Characteristics
1
Run model for each characteristic
Rank assets
Long top tercile and short bottom tercile
Example: CAPM (Beta)
Char. 1: $R_{i,t+1} = \alpha_{i,t} + b_{1it}\beta_{it} + b_{2it}C1 + \varepsilon_{it}$
...
Char. N: $R_{i,t+1} = \alpha_{i,t} + b_{1it}\beta_{it} + b_{2it}CN + \varepsilon_{it}$
Fixed
Sig. Char @ 5%
2
Equal weight characteristics to generate portfolio
Jointly Testing Characteristics
1
Run model with all selected characteristics
Rank assets
Long top tercile and short bottom tercile
$R_{i,t+1} = \alpha_{i,t} + \underbrace{b_{1it}C1 + \cdots + b_{Nit}CN}_{\text{Significant Characteristics}} + \varepsilon_{it}$

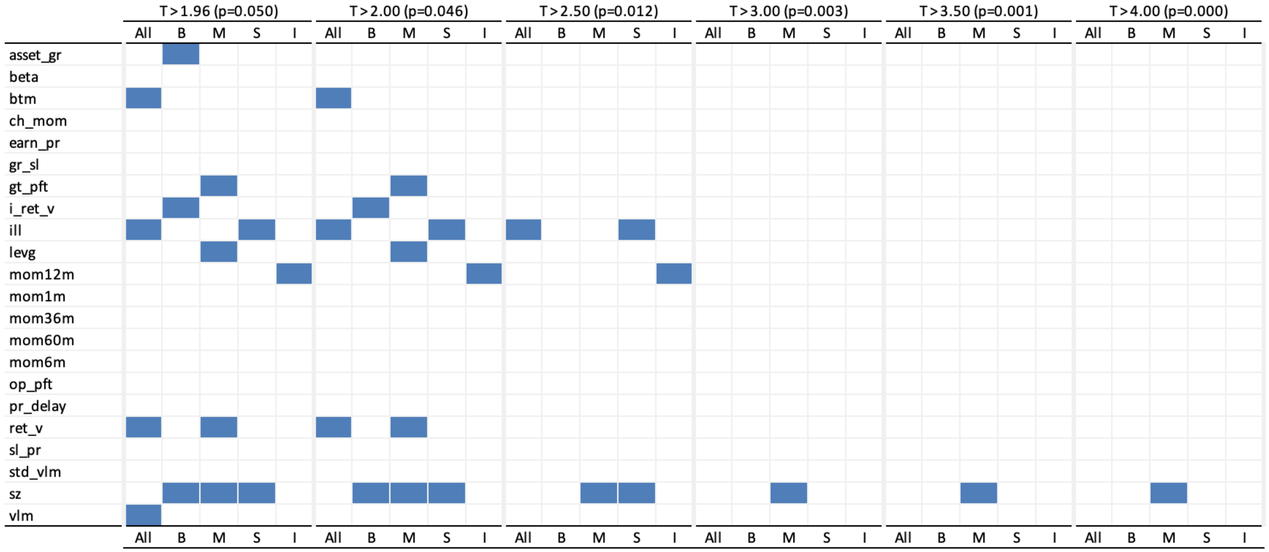
Figure 3: Method to generate portfolios with significant characteristics. In blue, models that individually test characteristics. In green, models that jointly test characteristics.



characteristic	T>1.96 (p=0.050)					T>2.00 (p=0.046)					T>2.50 (p=0.012)					T>3.00 (p=0.003)					T>3.50 (p=0.001)					T>4.00 (p=0.000)				
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov
asset_gr																														
beta																														
btm																														
ch_mom																														
earn_pr	🟦					🟦					🟦					🟦					🟦									
gr_sl																														
gt_pft																														
i_ret_v	🟦					🟦					🟦					🟦														
ill	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦			🟦	🟦	🟦				🟦	🟦		
levg																														
mom12m	🟦					🟦					🟦					🟦					🟦					🟦				
mom12m2	🟦					🟦					🟦					🟦					🟦					🟦				
mom1m																														
mom36m																														
mom60m																														
mom6m	🟦				🟦	🟦					🟦					🟦					🟦									
mom6m2	🟦				🟦	🟦				🟦	🟦					🟦														
op_pft																														
pr_delay																														
ret_v	🟦					🟦					🟦					🟦														
sl_pr	🟦					🟦					🟦																			
std_vlm		🟦	🟦				🟦	🟦																						
sz	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦			🟦	🟦	🟦			🟦	🟦				🟦	🟦				🟦	🟦		
vlm		🟦	🟦				🟦	🟦				🟦	🟦				🟦	🟦				🟦	🟦				🟦	🟦		
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov

All	Big	Medium	Small	Ibov	Mean	(+)	(-)
-0.030	-0.011	-0.021	-0.015	-0.032	-0.023	0	5
-0.026	0.000	-0.015	-0.056	-0.013	-0.014	0	5
0.016	0.026	0.038	0.031	0.031	0.030	5	0
0.006	0.005	0.026	0.006	0.009	0.014	5	0
0.072	-0.005	0.001	0.037	0.040	0.022	4	1
0.008	-0.001	0.008	-0.036	-0.004	0.004	2	3
0.001	-0.001	-0.012	0.033	0.002	-0.004	3	2
-0.043	0.023	-0.019	0.007	-0.018	-0.015	2	3
-0.041	**0.292**	**0.136**	**-0.046**	**0.850**	0.275	3	2
0.012	0.007	-0.012	0.014	0.012	0.001	4	1
0.080	0.038	0.018	0.058	0.048	0.040	5	0
0.077	0.036	0.010	0.069	0.056	0.038	5	0
0.017	0.016	0.005	-0.011	0.029	0.014	4	1
0.011	-0.022	-0.024	0.001	-0.006	-0.013	2	3
0.012	-0.012	-0.047	0.021	0.008	-0.017	3	2
0.061	0.030	0.037	0.012	**0.049**	0.043	5	0
0.061	0.030	0.032	0.017	**0.054**	0.042	5	0
0.015	-0.012	0.004	0.002	-0.007	0.001	3	2
-0.007	-0.024	0.001	0.012	-0.021	-0.010	2	3
-0.051	-0.017	-0.012	-0.023	-0.026	-0.024	0	5
0.038	0.017	-0.010	0.035	0.012	0.009	4	1
0.020	**-0.039**	**-0.306**	-0.182	-0.014	-0.129	1	4
0.025	**-0.077**	**-0.508**	**-0.212**	-0.001	-0.214	1	4
0.011	**-0.066**	**-0.102**	-0.030	-0.028	-0.057	1	4
All	Big	Medium	Small	Ibov	Mean	(+)	(-)

Figure 4: Single-characteristic Analysis: Only one characteristic is used in the model. Left panel shows a heatmap of statistical significance with blue cells highlighting T-stats beyond significance thresholds. Right panel presents Fama-MacBeth procedure point estimates for each subsample, including the mean of coefficients and their signs. Coefficients significant at the 5% level are emphasized in bold with gray background..



	T>1.96 (p=0.050)					T>2.00 (p=0.046)					T>2.50 (p=0.012)					T>3.00 (p=0.003)					T>3.50 (p=0.001)					T>4.00 (p=0.000)				
	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I
asset_gr		■																												
beta																														
btm	■					■																								
ch_mom																														
earn_pr																														
gr_sl																														
gt_pft			■					■																						
i_ret_v		■					■																							
ill	■			■		■			■		■			■																
levg			■					■																						
mom12m					■					■					■															
mom1m																														
mom36m																														
mom60m																														
mom6m																														
op_pft																														
pr_delay																														
ret_v	■		■			■		■																						
sl_pr																														
std_vlm																														
sz		■	■	■			■	■	■				■	■				■					■					■		
vlm	■																													
	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I	All	B	M	S	I

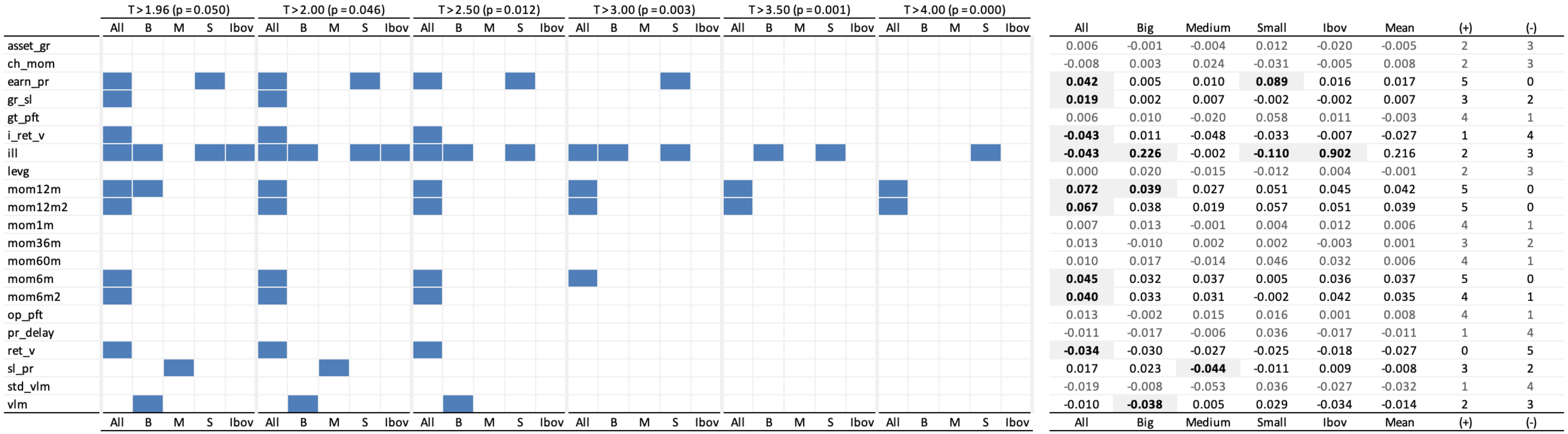
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)
asset_gr	-0.005	**-0.034**	0.004	0.031	-0.084	-0.018	2	3
beta	-0.007	-0.026	-0.103	122.282	-0.036	24.422	1	4
btm	**0.022**	0.023	0.022	-0.009	0.027	0.017	4	1
ch_mom	-0.018	-0.015	0.034	-0.104	-0.006	-0.022	1	4
earn_pr	0.022	0.015	-0.005	0.079	-0.033	0.016	3	2
gr_sl	0.011	0.021	0.000	0.061	0.061	0.031	4	1
gt_pft	0.007	0.013	**-0.041**	0.087	0.002	0.014	4	1
i_ret_v	-0.021	**0.049**	0.070	-5.479	0.059	-1.064	3	2
ill	**-0.035**	0.093	0.091	**-0.120**	-0.078	-0.010	2	3
levg	0.012	0.017	**-0.160**	-0.676	-0.026	-0.167	2	3
mom12m	0.020	0.060	-0.004	-0.239	**0.129**	-0.007	3	2
mom1m	0.002	0.003	-0.034	0.070	0.058	0.020	4	1
mom36m	-0.003	-0.017	-0.025	52.551	0.017	10.505	2	3
mom60m	-0.022	0.001	-0.038	-47.913	0.021	-9.590	2	3
mom6m	0.027	-0.028	-0.026	0.391	-0.082	0.056	2	3
op_pft	0.007	-0.005	-0.016	0.071	-0.006	0.010	2	3
pr_delay	-0.011	-0.017	-0.026	NA	0.243	0.047	2	3
ret_v	**-0.051**	-0.047	**-0.058**	-0.060	-0.005	-0.044	0	5
sl_pr	-0.035	-0.008	-0.040	0.080	0.120	0.023	2	3
std_vlm	0.016	0.014	-0.151	2.157	-0.037	0.400	3	2
sz	0.011	**-0.047**	**-0.315**	**-0.498**	0.097	-0.150	2	3
vlm	**-0.028**	-0.015	0.013	-0.151	-0.052	-0.047	1	4
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)

Figure 5: All Characteristics Analysis: twenty-two characteristics are used together in the model. Left panel shows a heatmap of statistical significance with blue cells highlighting T-stats beyond significance thresholds. Right panel presents Fama-MacBeth procedure point estimates for each subsample, including the mean of coefficients and their signs. Coefficients significant at the 5% level are emphasized in bold with gray background.

	T > 1.96 (p = 0.050)					T > 2.00 (p = 0.046)					T > 2.50 (p = 0.012)					T > 3.00 (p = 0.003)					T > 3.50 (p = 0.001)					T > 4.00 (p = 0.000)				
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov
asset_gr																														
btm																														
ch_mom																														
earn_pr	■					■					■					■														
gr_sl																														
gt_pft																														
i_ret_v	■					■					■					■														
ill	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■			■	■	■				■	■		
levg																														
mom12m	■					■					■					■					■					■				
mom12m2	■				■	■				■	■					■					■					■				
mom1m																														
mom36m																														
mom60m																														
mom6m	■	■			■	■	■			■	■					■					■									
mom6m2	■				■	■				■	■					■														
op_pft	■					■																								
pr_delay																														
ret_v	■					■					■					■														
sl_pr																														
std_vlm		■	■				■	■				■	■																	
sz	■	■	■	■		■	■	■	■			■	■	■			■	■	■			■	■				■	■		
vlm		■	■				■	■				■	■				■	■				■	■				■	■		
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov

	All	Big	Medium	Small	Ibov	Mean	(+)	(-)
asset_gr	0.005	-0.001	-0.008	-0.015	-0.023	-0.007	1	4
btm	0.012	0.018	0.036	0.049	0.017	0.024	5	0
ch_mom	-0.007	0.008	0.028	0.003	0.006	0.013	4	1
earn_pr	**0.053**	0.001	0.001	0.035	0.035	0.018	5	0
gr_sl	0.019	0.001	0.011	-0.037	-0.002	0.008	3	2
gt_pft	0.003	0.006	-0.015	0.029	0.005	-0.003	4	1
i_ret_v	**-0.046**	0.024	0.006	-0.001	-0.014	-0.005	2	3
ill	**-0.048**	**0.299**	**0.131**	**-0.047**	**0.748**	0.252	3	2
levg	0.006	0.008	-0.006	0.026	0.009	0.002	4	1
mom12m	**0.074**	0.037	0.013	0.051	0.049	0.037	5	0
mom12m2	**0.073**	0.035	0.003	0.063	**0.054**	0.034	5	0
mom1m	0.016	0.009	0.001	-0.003	0.018	0.009	4	1
mom36m	0.015	-0.018	-0.020	-0.012	-0.001	-0.009	1	4
mom60m	0.020	0.003	-0.023	0.008	0.026	0.001	4	1
mom6m	**0.052**	**0.037**	0.036	0.006	**0.045**	0.041	5	0
mom6m2	**0.049**	0.033	0.029	0.013	**0.049**	0.038	5	0
op_pft	**0.018**	-0.005	0.006	0.002	0.000	0.005	3	2
pr_delay	-0.010	-0.021	-0.001	0.019	-0.008	-0.008	1	4
ret_v	**-0.044**	-0.020	-0.009	-0.011	-0.023	-0.021	0	5
sl_pr	0.018	0.019	-0.007	0.027	0.010	0.007	4	1
std_vlm	0.017	**-0.039**	**-0.306**	-0.159	-0.009	-0.129	1	4
sz	**0.024**	**-0.075**	**-0.502**	**-0.236**	0.001	-0.211	2	3
vlm	0.015	**-0.069**	**-0.105**	-0.018	-0.018	-0.056	1	4
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)

Figure 6: Beta + Characteristic Analysis: Characteristic and market beta are used in the model. Left panel shows a heatmap of statistical significance with blue cells highlighting T-stats beyond significance thresholds. Right panel presents Fama-MacBeth procedure point estimates for each subsample, including the mean of coefficients and their signs. Coefficients significant at the 5% level are emphasized in bold with gray background.



	T>1.96 (p=0.050)					T>2.00 (p=0.046)					T>2.50 (p=0.012)					T>3.00 (p=0.003)					T>3.50 (p=0.001)					T>4.00 (p=0.000)				
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov
asset_gr																														
ch_mom																														
earn_pr	🟦			🟦		🟦			🟦		🟦			🟦					🟦											
gr_sl	🟦					🟦																								
gt_pft																														
i_ret_v	🟦					🟦					🟦																			
ill	🟦	🟦		🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦		🟦		🟦	🟦		🟦			🟦		🟦					🟦	
levg																														
mom12m	🟦	🟦				🟦					🟦					🟦					🟦					🟦				
mom12m2	🟦					🟦					🟦					🟦					🟦					🟦				
mom1m																														
mom36m																														
mom60m																														
mom6m	🟦					🟦					🟦					🟦														
mom6m2	🟦					🟦					🟦																			
op_pft																														
pr_delay																														
ret_v	🟦					🟦					🟦																			
sl_pr			🟦					🟦																						
std_vlm																														
vlm		🟦					🟦					🟦																		
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov

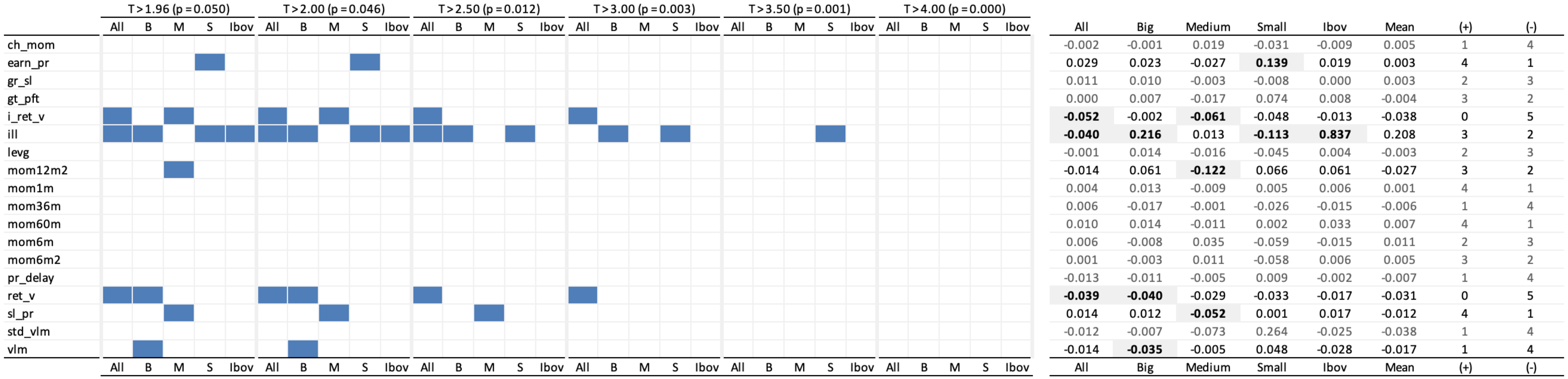
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)
asset_gr	0.006	-0.001	-0.004	0.012	-0.020	-0.005	2	3
ch_mom	-0.008	0.003	0.024	-0.031	-0.005	0.008	2	3
earn_pr	**0.042**	0.005	0.010	**0.089**	0.016	0.017	5	0
gr_sl	**0.019**	0.002	0.007	-0.002	-0.002	0.007	3	2
gt_pft	0.006	0.010	-0.020	0.058	0.011	-0.003	4	1
i_ret_v	**-0.043**	0.011	-0.048	-0.033	-0.007	-0.027	1	4
ill	**-0.043**	**0.226**	-0.002	**-0.110**	**0.902**	0.216	2	3
levg	0.000	0.020	-0.015	-0.012	0.004	-0.001	2	3
mom12m	**0.072**	**0.039**	0.027	0.051	0.045	0.042	5	0
mom12m2	**0.067**	0.038	0.019	0.057	0.051	0.039	5	0
mom1m	0.007	0.013	-0.001	0.004	0.012	0.006	4	1
mom36m	0.013	-0.010	0.002	0.002	-0.003	0.001	3	2
mom60m	0.010	0.017	-0.014	0.046	0.032	0.006	4	1
mom6m	**0.045**	0.032	0.037	0.005	0.036	0.037	5	0
mom6m2	**0.040**	0.033	0.031	-0.002	0.042	0.035	4	1
op_pft	0.013	-0.002	0.015	0.016	0.001	0.008	4	1
pr_delay	-0.011	-0.017	-0.006	0.036	-0.017	-0.011	1	4
ret_v	**-0.034**	-0.030	-0.027	-0.025	-0.018	-0.027	0	5
sl_pr	0.017	0.023	**-0.044**	-0.011	0.009	-0.008	3	2
std_vlm	-0.019	-0.008	-0.053	0.036	-0.027	-0.032	1	4
vlm	-0.010	**-0.038**	0.005	0.029	-0.034	-0.014	2	3
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)

Figure 7: FF3 + Characteristic Analysis: Characteristic, market beta, size and book-to-market are used in the model. Left panel shows a heatmap of statistical significance with blue cells highlighting T-stats beyond significance thresholds. Right panel presents Fama-MacBeth procedure point estimates for each subsample, including the mean of coefficients and their signs. Coefficients significant at the 5% level are emphasized in bold with gray background.

	T > 1.96 (p = 0.050)					T > 2.00 (p = 0.046)					T > 2.50 (p = 0.012)					T > 3.00 (p = 0.003)					T > 3.50 (p = 0.001)					T > 4.00 (p = 0.000)				
characteristic	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov
ch_mom																														
earn_pr	🟦			🟦		🟦			🟦					🟦																
gr_sl																														
gt_pft																														
i_ret_v	🟦		🟦			🟦		🟦			🟦																			
ill	🟦	🟦		🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦		🟦		🟦	🟦		🟦			🟦		🟦					🟦	
levg																														
mom12m	🟦	🟦				🟦	🟦				🟦					🟦					🟦					🟦				
mom12m2	🟦	🟦				🟦	🟦				🟦					🟦					🟦					🟦				
mom1m																														
mom36m																														
mom60m																														
mom6m	🟦					🟦					🟦					🟦														
mom6m2	🟦					🟦					🟦																			
pr_delay																														
ret_v	🟦					🟦					🟦																			
sl_pr			🟦					🟦					🟦																	
std_vlm																														
vlm		🟦					🟦					🟦																		
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov

	All	Big	Medium	Small	Ibov	Mean	(+)	(-)
ch_mom	-0.005	-0.006	0.026	-0.029	-0.013	0.006	1	4
earn_pr	**0.034**	0.005	-0.016	**0.087**	0.016	0.005	4	1
gr_sl	0.020	0.015	0.004	-0.010	0.009	0.010	4	1
gt_pft	0.000	0.010	-0.020	0.058	0.004	-0.005	3	2
i_ret_v	**-0.044**	0.012	**-0.062**	-0.035	-0.010	-0.033	1	4
ill	**-0.044**	**0.239**	0.006	**-0.110**	**0.881**	0.218	3	2
levg	0.001	0.018	-0.019	-0.029	0.001	-0.004	3	2
mom12m	**0.072**	**0.043**	0.029	0.052	0.045	0.044	5	0
mom12m2	**0.065**	**0.041**	0.015	0.056	0.050	0.037	5	0
mom1m	0.004	0.012	-0.006	0.003	0.011	0.003	4	1
mom36m	0.009	-0.008	0.009	-0.002	-0.011	0.002	2	3
mom60m	0.008	0.018	-0.006	-0.002	0.032	0.009	3	2
mom6m	**0.046**	0.030	0.038	0.008	0.030	0.036	5	0
mom6m2	**0.038**	0.030	0.027	0.004	0.035	0.031	5	0
pr_delay	-0.011	-0.015	-0.002	0.037	-0.012	-0.008	1	4
ret_v	**-0.036**	-0.035	-0.027	-0.029	-0.018	-0.029	0	5
sl_pr	0.021	0.017	**-0.055**	-0.009	0.015	-0.011	3	2
std_vlm	-0.015	-0.009	-0.073	0.237	-0.025	-0.039	1	4
vlm	-0.011	**-0.039**	-0.007	0.038	-0.032	-0.019	1	4
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)

Figure 8: FF5 + Characteristic Analysis: Characteristic, market beta, book-to-market ratio and size are used in the model. Left panel shows a heatmap of statistical significance with blue cells highlighting T-stats beyond significance thresholds. Right panel presents Fama-MacBeth procedure point estimates for each subsample, including the mean of coefficients and their signs. Coefficients significant at the 5% level are emphasized in bold with gray background.



	T > 1.96 (p = 0.050)					T > 2.00 (p = 0.046)					T > 2.50 (p = 0.012)					T > 3.00 (p = 0.003)					T > 3.50 (p = 0.001)					T > 4.00 (p = 0.000)				
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov
ch_mom																														
earn_pr				🟦					🟦																					
gr_sl																														
gt_pft																														
i_ret_v	🟦		🟦			🟦		🟦			🟦					🟦														
ill	🟦	🟦		🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦		🟦			🟦		🟦					🟦						
levg																														
mom12m2			🟦																											
mom1m																														
mom36m																														
mom60m																														
mom6m																														
mom6m2																														
pr_delay																														
ret_v	🟦	🟦				🟦	🟦				🟦					🟦														
sl_pr			🟦					🟦					🟦																	
std_vlm																														
vlm		🟦					🟦																							
	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov	All	B	M	S	Ibov

	All	Big	Medium	Small	Ibov	Mean	(+)	(-)
ch_mom	-0.002	-0.001	0.019	-0.031	-0.009	0.005	1	4
earn_pr	0.029	0.023	-0.027	**0.139**	0.019	0.003	4	1
gr_sl	0.011	0.010	-0.003	-0.008	0.000	0.003	2	3
gt_pft	0.000	0.007	-0.017	0.074	0.008	-0.004	3	2
i_ret_v	**-0.052**	-0.002	**-0.061**	-0.048	-0.013	-0.038	0	5
ill	**-0.040**	**0.216**	0.013	**-0.113**	**0.837**	0.208	3	2
levg	-0.001	0.014	-0.016	-0.045	0.004	-0.003	2	3
mom12m2	-0.014	0.061	**-0.122**	0.066	0.061	-0.027	3	2
mom1m	0.004	0.013	-0.009	0.005	0.006	0.001	4	1
mom36m	0.006	-0.017	-0.001	-0.026	-0.015	-0.006	1	4
mom60m	0.010	0.014	-0.011	0.002	0.033	0.007	4	1
mom6m	0.006	-0.008	0.035	-0.059	-0.015	0.011	2	3
mom6m2	0.001	-0.003	0.011	-0.058	0.006	0.005	3	2
pr_delay	-0.013	-0.011	-0.005	0.009	-0.002	-0.007	1	4
ret_v	**-0.039**	**-0.040**	-0.029	-0.033	-0.017	-0.031	0	5
sl_pr	0.014	0.012	**-0.052**	0.001	0.017	-0.012	4	1
std_vlm	-0.012	-0.007	-0.073	0.264	-0.025	-0.038	1	4
vlm	-0.014	**-0.035**	-0.005	0.048	-0.028	-0.017	1	4
	All	Big	Medium	Small	Ibov	Mean	(+)	(-)

Figure 9: FF5 + mom12m + Characteristic Analysis: Characteristic, market beta, book-to-market ratio, size and 12-month momentum (mom12m) are used in the model. Left panel shows a heatmap of statistical significance with blue cells highlighting T-stats beyond significance thresholds. Right panel presents Fama-MacBeth procedure point estimates for each subsample, including the mean of coefficients and their signs. Coefficients significant at the 5% level are emphasized in bold with gray background.

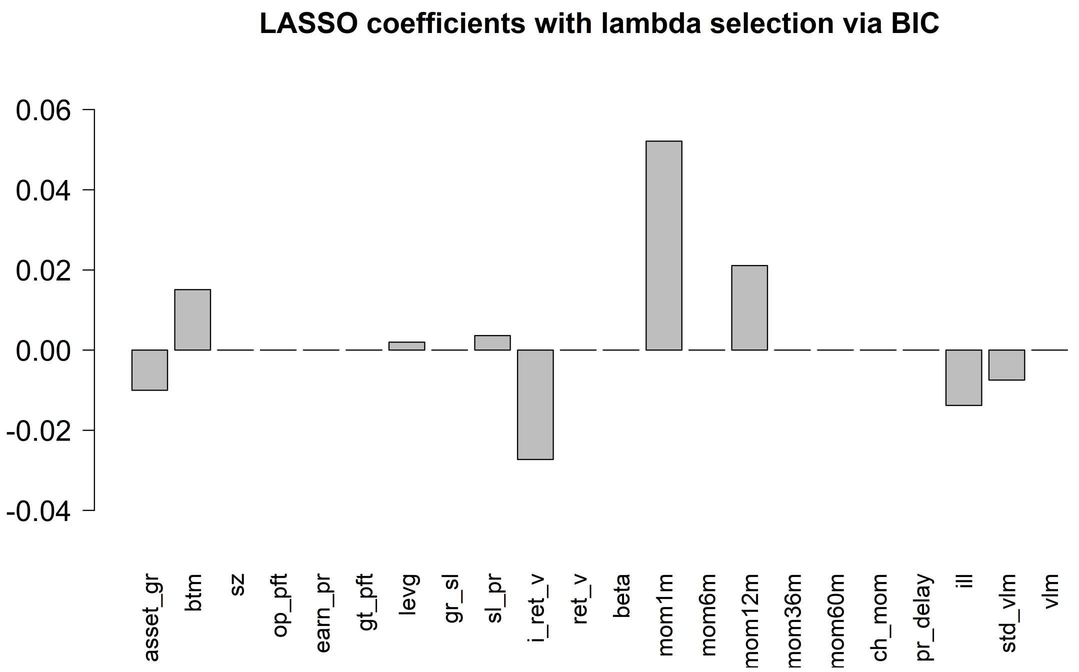


Figure 10: LASSO results

Table 3: Selected characteristics in nonparametric adaptive group LASSO model

Knots		1	2-6	7	8	9	10	> 10
# Selected		2	1	3	3	2	4	0
Characteristics	# Selected	(1)	(2)	(3)	(4)	(5)	(6)	(7)
i_ret_v	3			i_ret_v			i_ret_v	
mom1m	1	mom1m						
mom12m	10	mom12m	mom12m	mom12m	mom12m	mom12m	mom12m	
mom60m	4			mom60m	mom60m	mom60m	mom60m	
pr_delay	2				pr_delay		pr_delay	

Table 4: Random forest parameters and measures

Measure	Value
mtry	5
trees	500
min_n	1
Train R^2	0.0513
Test R^2	-0.0316
Validation RMSE	0.8672
Test RMSE	1.0630

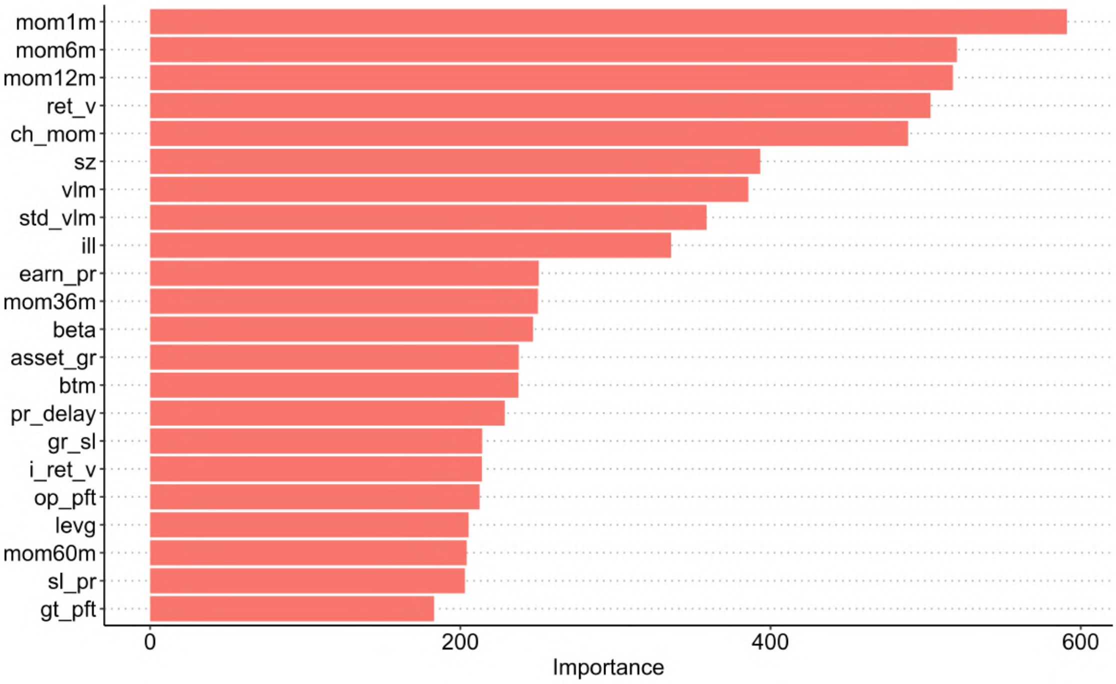


Figure 11: Random Forest variable importance

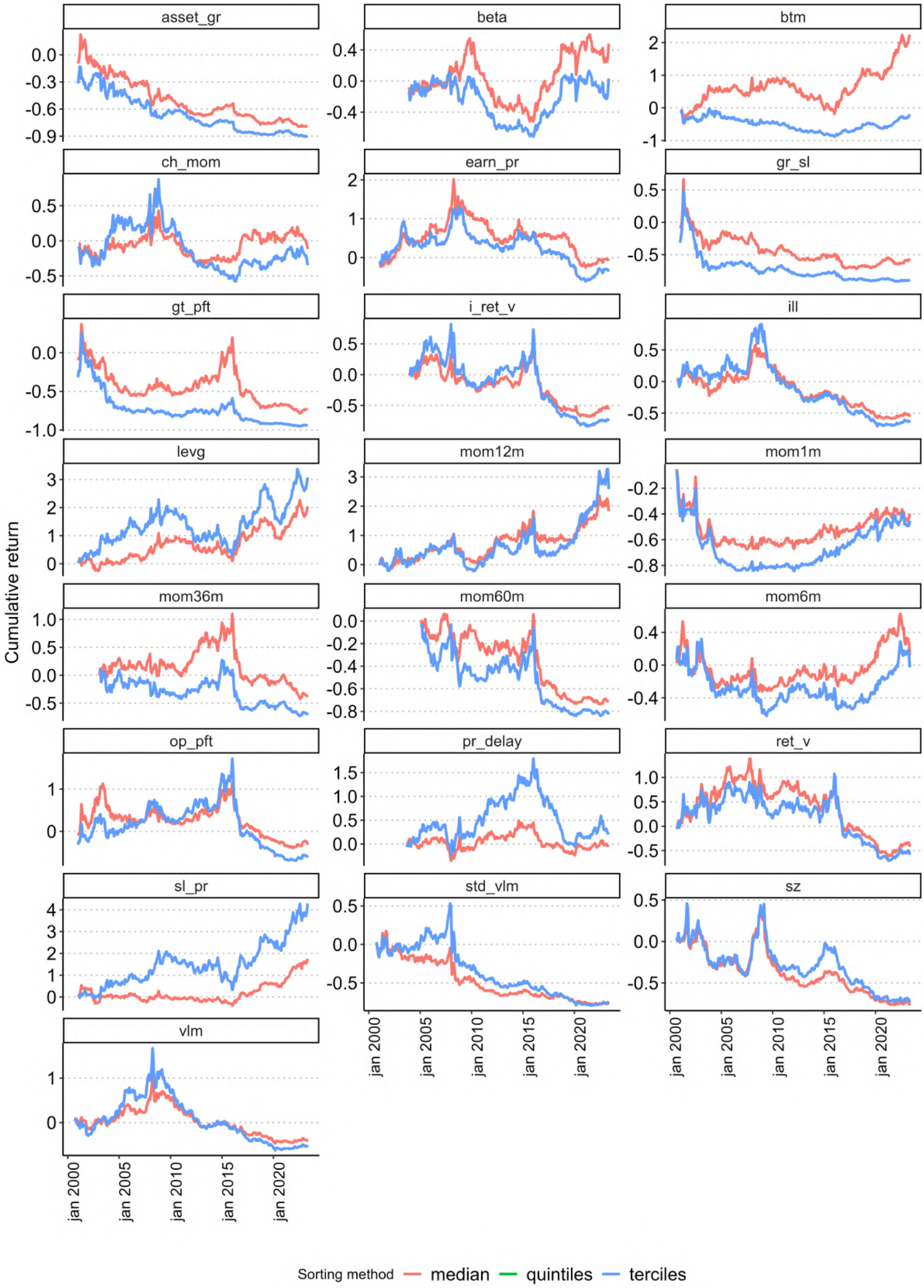


Figure 12: Cumulative return of value weighted long-short portfolios sorted on characteristics

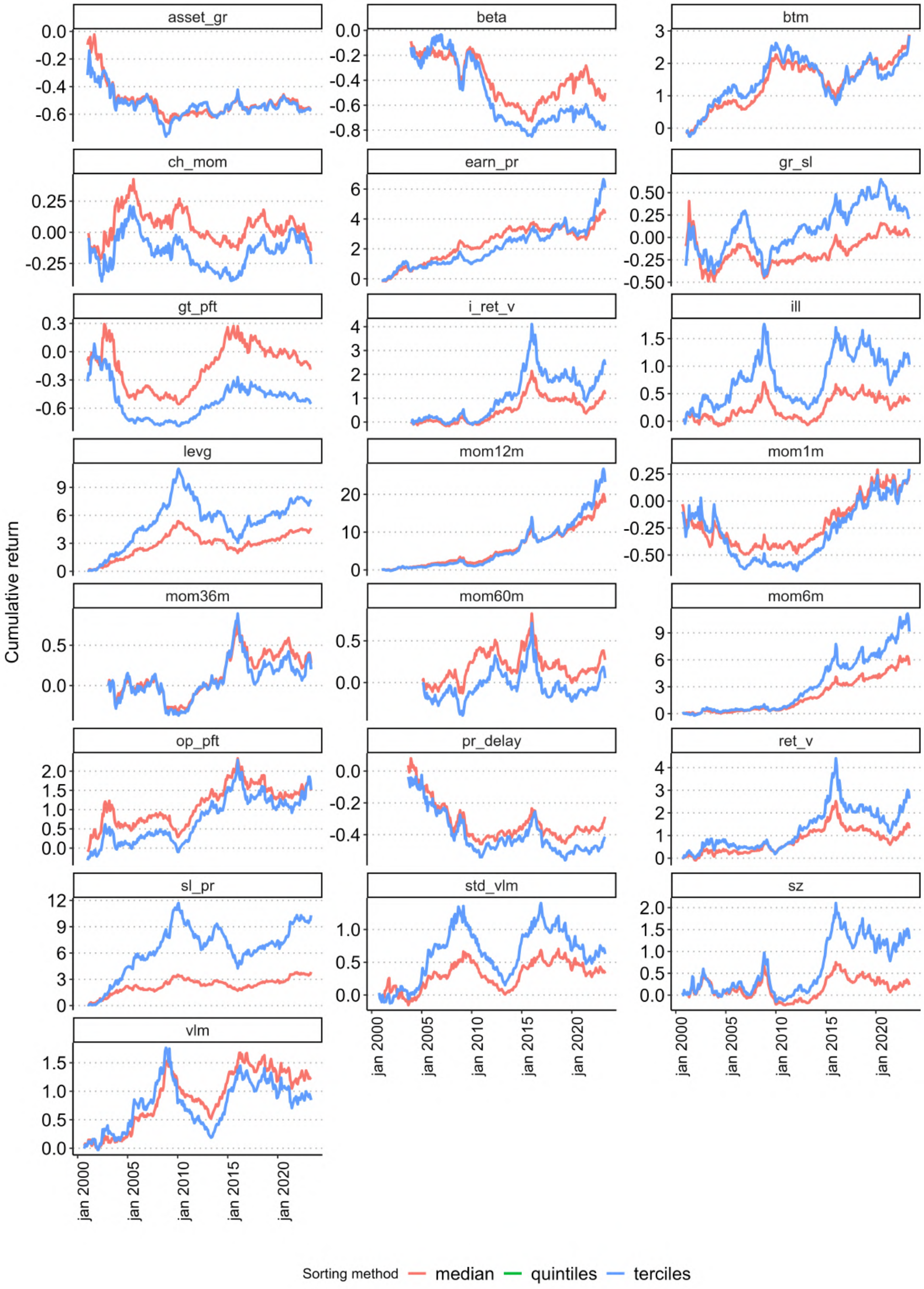


Figure 13: Cumulative return of equal weighted long-short portfolios sorted on characteristics

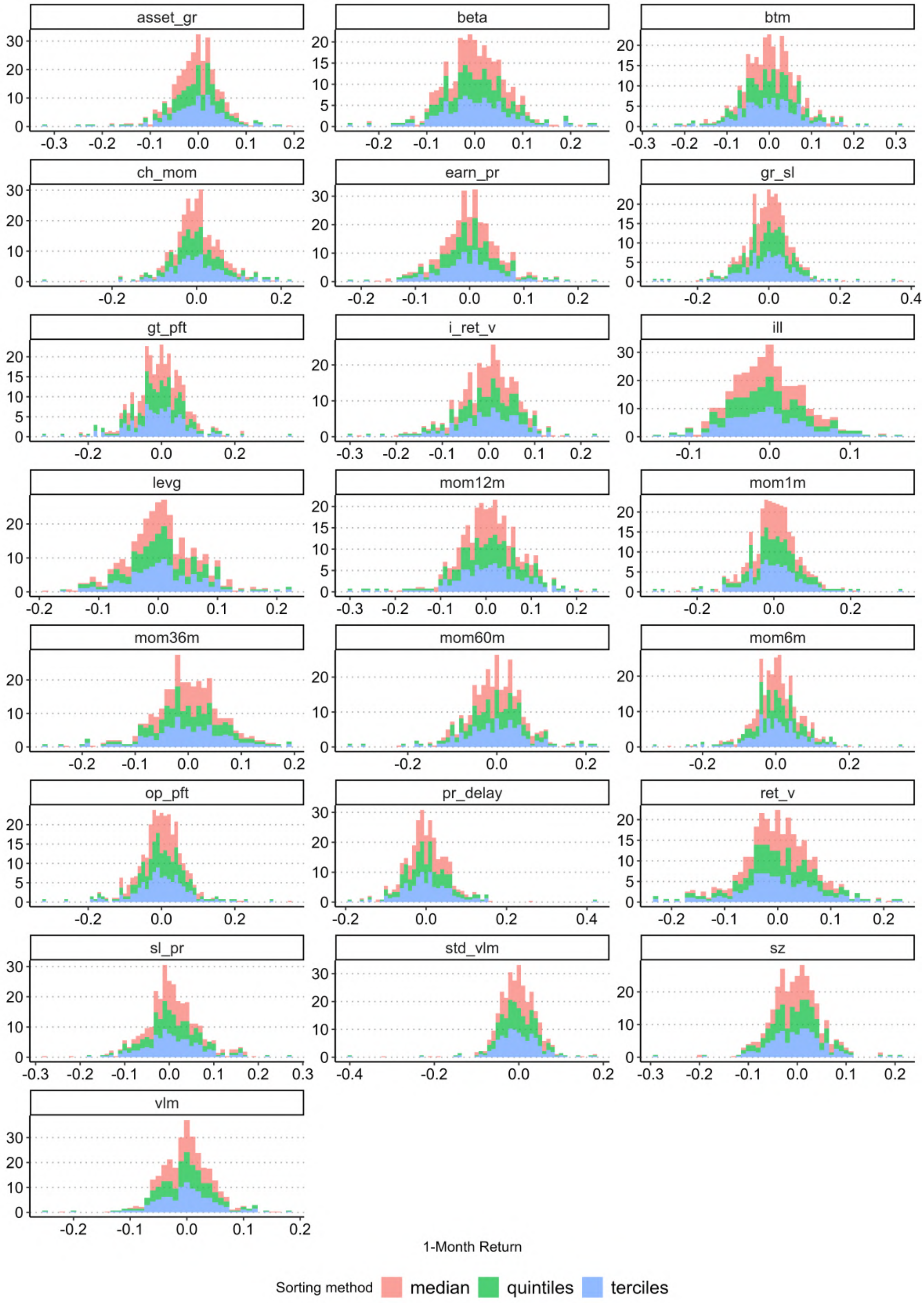


Figure 14: Distribution of monthly return of value weighted long-short portfolios sorted on characteristics

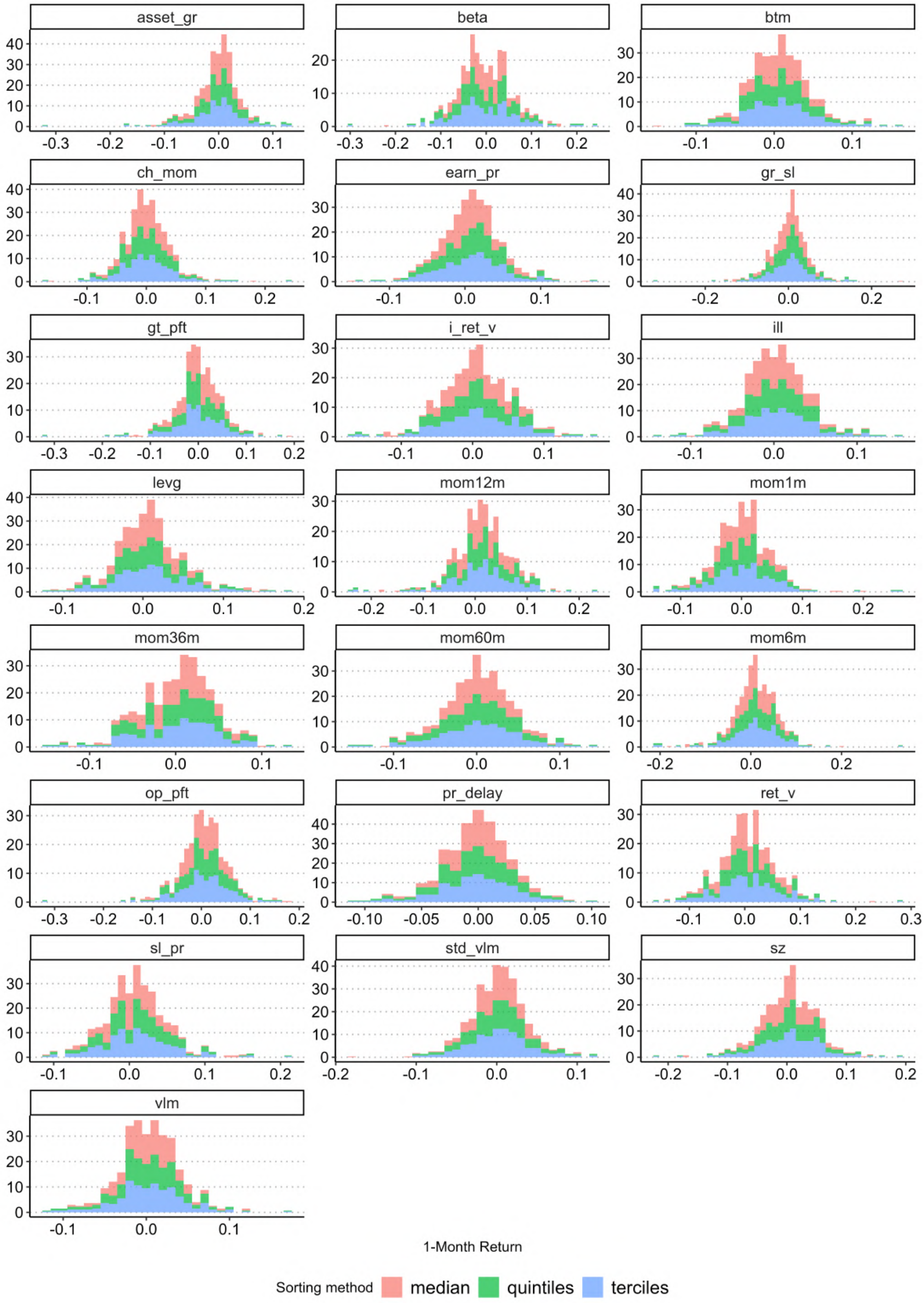


Figure 15: Distribution of monthly return of equal weighted long-short portfolios sorted on characteristics

Table 5: Descriptive statistics of monthly value weighted long-short portfolio returns

Portfolio	Avg. ann. return	Ann. vol.	Skewness	Kurtosis	Max drawdown	Cum return
Median sorted						
asset_gr	-0.057	0.153	-0.098	2.310	-1.016	-0.787
beta	0.040	0.202	0.112	1.060	-1.066	0.480
btm	0.070	0.187	0.002	1.248	-1.106	2.248
ch_mom	0.010	0.176	-0.494	3.427	-0.760	-0.118
earn_pr	0.011	0.171	-0.210	1.013	-2.268	-0.085
gr_sl	-0.018	0.194	0.837	7.383	-1.382	-0.560
gt_pft	-0.034	0.216	0.084	1.169	-1.155	-0.723
i_ret_v	-0.022	0.200	-0.717	2.529	-1.187	-0.566
ill	-0.027	0.134	0.183	0.145	-1.169	-0.559
levg	0.070	0.199	-0.050	0.737	-0.990	2.050
mom12m	0.066	0.197	-0.144	1.058	-1.073	1.827
mom1m	-0.003	0.192	-0.337	2.638	-0.625	-0.397
mom36m	-0.006	0.187	-0.334	1.842	-1.520	-0.390
mom60m	-0.051	0.183	-0.214	1.521	-0.799	-0.715
mom6m	0.025	0.196	-0.487	3.432	-0.854	0.122
op_pft	0.002	0.200	0.915	5.684	-1.548	-0.320
pr_delay	0.011	0.166	0.540	4.023	-0.725	-0.056
ret_v	-0.004	0.198	-0.063	0.846	-1.992	-0.422
sl_pr	0.066	0.204	-0.072	3.135	-0.927	1.749
std_vlm	-0.054	0.159	-1.329	10.039	-0.957	-0.783
sz	-0.052	0.158	-0.163	3.031	-1.139	-0.774
vlm	-0.014	0.144	-0.208	3.518	-1.564	-0.425
Tercile sorted						
asset_gr	-0.078	0.207	-1.027	4.298	-0.772	-0.896
beta	0.033	0.251	0.151	1.432	-0.825	0.032
btm	0.024	0.259	0.074	1.741	-0.853	-0.200
ch_mom	0.009	0.233	-0.195	3.387	-1.456	-0.348
earn_pr	0.001	0.204	0.135	1.585	-1.917	-0.360
gr_sl	-0.061	0.261	-0.211	3.575	-1.400	-0.885
gt_pft	-0.092	0.263	-0.076	2.851	-1.200	-0.943
i_ret_v	-0.039	0.249	-0.707	1.732	-1.653	-0.752
ill	-0.032	0.168	0.259	0.293	-1.614	-0.652
levg	0.085	0.214	0.384	0.880	-2.001	3.077
mom12m	0.093	0.265	-0.560	1.793	-1.324	2.579
mom1m	0.007	0.256	-0.168	3.111	-0.781	-0.451
mom36m	-0.029	0.250	-0.337	1.024	-1.002	-0.713
mom60m	-0.061	0.259	-0.417	2.472	-0.811	-0.829
mom6m	0.033	0.259	-0.132	3.149	-0.940	-0.026
op_pft	-0.012	0.249	-0.247	2.989	-2.415	-0.622
pr_delay	0.033	0.217	1.251	7.987	-1.842	0.243
ret_v	-0.005	0.259	-0.145	0.775	-1.778	-0.585
sl_pr	0.098	0.219	0.401	1.188	-1.788	4.287
std_vlm	-0.045	0.185	-1.495	11.868	-1.326	-0.766
sz	-0.040	0.191	-0.217	3.340	-1.194	-0.737
vlm	-0.020	0.174	-0.297	2.956	-2.291	-0.552
Quintile sorted						
asset_gr	-0.078	0.207	-1.027	4.298	-0.772	-0.896
beta	0.033	0.251	0.151	1.432	-0.825	0.032
btm	0.024	0.259	0.074	1.741	-0.853	-0.200
ch_mom	0.009	0.233	-0.195	3.387	-1.456	-0.348
earn_pr	0.001	0.204	0.135	1.585	-1.917	-0.360
gr_sl	-0.061	0.261	-0.211	3.575	-1.400	-0.885
gt_pft	-0.092	0.263	-0.076	2.851	-1.200	-0.943
i_ret_v	-0.039	0.249	-0.707	1.732	-1.653	-0.752
ill	-0.032	0.168	0.259	0.293	-1.614	-0.652
levg	0.085	0.214	0.384	0.880	-2.001	3.077
mom12m	0.093	0.265	-0.560	1.793	-1.324	2.579
mom1m	0.007	0.256	-0.168	3.111	-0.781	-0.451
mom36m	-0.029	0.250	-0.337	1.024	-1.002	-0.713
mom60m	-0.061	0.259	-0.417	2.472	-0.811	-0.829
mom6m	0.033	0.259	-0.132	3.149	-0.940	-0.026
op_pft	-0.012	0.249	-0.247	2.989	-2.415	-0.622
pr_delay	0.033	0.217	1.251	7.987	-1.842	0.243
ret_v	-0.005	0.259	-0.145	0.775	-1.778	-0.585
sl_pr	0.098	0.219	0.401	1.188	-1.788	4.287
std_vlm	-0.045	0.185	-1.495	11.868	-1.326	-0.766
sz	-0.040	0.191	-0.217	3.340	-1.194	-0.737
vlm	-0.020	0.174	-0.297	2.956	-2.291	-0.552

Table 6: Descriptive statistics of monthly equal weighted long-short portfolio returns

Portfolio	Avg. ann. return	Ann. vol.	Skewness	Kurtosis	Max drawdown	Cum return
Median sorted						
asset_gr	-0.031	0.109	-0.829	1.245	-0.641	-0.567
beta	-0.021	0.172	0.107	1.641	-0.642	-0.501
btm	0.068	0.118	0.049	2.161	-1.343	2.888
ch_mom	-0.000	0.119	0.387	4.307	-0.576	-0.150
earn_pr	0.082	0.119	0.110	3.044	-1.153	4.357
gr_sl	0.012	0.147	0.379	6.435	-0.894	0.041
gt_pft	0.002	0.146	0.045	2.008	-0.851	-0.172
i_ret_v	0.050	0.147	-0.106	0.682	-1.781	1.168
ill	0.020	0.112	0.059	0.580	-0.778	0.353
levg	0.086	0.129	0.482	2.001	-3.420	4.643
mom12m	0.144	0.159	-0.576	2.556	-4.234	17.810
mom1m	0.019	0.139	0.657	2.974	-0.465	0.244
mom36m	0.020	0.125	-0.294	0.611	-0.573	0.277
mom60m	0.021	0.124	-0.104	0.702	-0.830	0.272
mom6m	0.093	0.146	-0.494	3.792	-1.020	5.393
op_pft	0.051	0.146	0.332	2.935	-1.064	1.499
pr_delay	-0.013	0.089	-0.146	1.069	-0.539	-0.287
ret_v	0.048	0.152	0.312	2.435	-1.920	1.308
sl_pr	0.079	0.129	0.541	2.196	-1.810	3.838
std_vlm	0.019	0.112	-0.629	3.632	-0.655	0.328
sz	0.020	0.144	-0.325	1.860	-0.888	0.234
vlm	0.041	0.110	0.109	1.142	-1.066	1.209
Tercile sorted						
asset_gr	-0.023	0.171	-1.212	6.583	-0.622	-0.575
beta	-0.044	0.236	0.105	1.865	-0.817	-0.754
btm	0.071	0.150	0.264	0.787	-1.907	2.854
ch_mom	-0.001	0.156	0.583	3.772	-0.600	-0.254
earn_pr	0.099	0.149	0.004	0.871	-1.200	6.083
gr_sl	0.026	0.179	-0.802	5.844	-0.728	0.233
gt_pft	-0.018	0.186	-1.010	4.906	-0.874	-0.553
i_ret_v	0.081	0.188	-0.223	0.690	-3.240	2.396
ill	0.042	0.150	0.105	0.925	-1.540	1.028
levg	0.109	0.159	0.381	1.028	-7.924	7.698
mom12m	0.167	0.214	-0.671	2.861	-6.578	23.279
mom1m	0.028	0.181	0.703	3.815	-0.676	0.298
mom36m	0.022	0.161	-0.404	0.298	-0.874	0.202
mom60m	0.016	0.164	-0.212	0.489	-0.939	0.053
mom6m	0.122	0.200	-0.202	4.974	-2.786	9.144
op_pft	0.059	0.183	-0.861	4.885	-1.372	1.517
pr_delay	-0.020	0.115	-0.273	0.706	-0.536	-0.413
ret_v	0.075	0.194	0.401	1.646	-3.305	2.631
sl_pr	0.121	0.160	0.512	1.502	-7.448	10.303
std_vlm	0.030	0.131	0.118	0.548	-1.207	0.625
sz	0.053	0.184	-0.292	1.640	-1.323	1.284
vlm	0.037	0.140	0.034	1.213	-1.574	0.847
Quintile sorted						
asset_gr	-0.023	0.171	-1.212	6.583	-0.622	-0.575
beta	-0.044	0.236	0.105	1.865	-0.817	-0.754
btm	0.071	0.150	0.264	0.787	-1.907	2.854
ch_mom	-0.001	0.156	0.583	3.772	-0.600	-0.254
earn_pr	0.099	0.149	0.004	0.871	-1.200	6.083
gr_sl	0.026	0.179	-0.802	5.844	-0.728	0.233
gt_pft	-0.018	0.186	-1.010	4.906	-0.874	-0.553
i_ret_v	0.081	0.188	-0.223	0.690	-3.240	2.396
ill	0.042	0.150	0.105	0.925	-1.540	1.028
levg	0.109	0.159	0.381	1.028	-7.924	7.698
mom12m	0.167	0.214	-0.671	2.861	-6.578	23.279
mom1m	0.028	0.181	0.703	3.815	-0.676	0.298
mom36m	0.022	0.161	-0.404	0.298	-0.874	0.202
mom60m	0.016	0.164	-0.212	0.489	-0.939	0.053
mom6m	0.122	0.200	-0.202	4.974	-2.786	9.144
op_pft	0.059	0.183	-0.861	4.885	-1.372	1.517
pr_delay	-0.020	0.115	-0.273	0.706	-0.536	-0.413
ret_v	0.075	0.194	0.401	1.646	-3.305	2.631
sl_pr	0.121	0.160	0.512	1.502	-7.448	10.303
std_vlm	0.030	0.131	0.118	0.548	-1.207	0.625
sz	0.053	0.184	-0.292	1.640	-1.323	1.284
vlm	0.037	0.140	0.034	1.213	-1.574	0.847

Table 7: Factor CAPM alphas

Portfolio	(1)	(2)	(3)	(4)	(5)	(6)
asset_gr	-0.014	-0.015	-0.015	-0.012	-0.011	-0.011
	(-5.077)	(-4.080)	(-4.080)	(-6.048)	(-3.495)	(-3.495)
btm	-0.004	-0.010	-0.010	-0.005	-0.005	-0.005
	(-1.167)	(-2.137)	(-2.137)	(-2.185)	(-1.791)	(-1.791)
sz	-0.013	-0.012	-0.012	-0.006	-0.003	-0.003
	(-4.675)	(-3.495)	(-3.495)	(-2.485)	(-0.991)	(-0.991)
op_pft	-0.008	-0.008	-0.008	-0.003	-0.002	-0.002
	(-2.282)	(-1.922)	(-1.922)	(-1.277)	(-0.696)	(-0.696)
earn_pr	-0.010	-0.010	-0.010	-0.002	0.000	0.000
	(-3.196)	(-2.698)	(-2.698)	(-0.736)	(0.021)	(0.021)
gt_pft	-0.010	-0.014	-0.014	-0.007	-0.008	-0.008
	(-2.673)	(-3.046)	(-3.046)	(-2.882)	(-2.604)	(-2.604)
levg	-0.007	-0.006	-0.006	-0.004	-0.003	-0.003
	(-2.111)	(-1.812)	(-1.812)	(-2.024)	(-1.129)	(-1.129)
gr_sl	-0.010	-0.014	-0.014	-0.009	-0.007	-0.007
	(-2.977)	(-3.017)	(-3.017)	(-3.322)	(-2.288)	(-2.288)
sl_pr	-0.006	-0.004	-0.004	-0.004	-0.001	-0.001
	(-1.741)	(-1.148)	(-1.148)	(-1.788)	(-0.364)	(-0.364)
i_ret_v	-0.007	-0.009	-0.009	-0.001	0.002	0.002
	(-1.987)	(-1.915)	(-1.915)	(-0.387)	(0.627)	(0.627)
ret_v	-0.007	-0.006	-0.006	-0.003	0.001	0.001
	(-2.099)	(-1.373)	(-1.373)	(-1.091)	(0.263)	(0.263)
beta	-0.010	-0.013	-0.013	-0.015	-0.019	-0.019
	(-3.077)	(-3.516)	(-3.516)	(-5.962)	(-5.767)	(-5.767)
mom1m	-0.009	-0.008	-0.008	-0.007	-0.006	-0.006
	(-2.631)	(-1.751)	(-1.751)	(-2.776)	(-1.835)	(-1.835)
mom6m	-0.006	-0.005	-0.005	-0.000	0.003	0.003
	(-1.810)	(-1.041)	(-1.041)	(-0.125)	(0.814)	(0.814)
mom12m	-0.002	0.001	0.001	0.005	0.007	0.007
	(-0.683)	(0.156)	(0.156)	(1.760)	(2.005)	(2.005)
mom36m	-0.008	-0.010	-0.010	-0.007	-0.006	-0.006
	(-2.262)	(-2.060)	(-2.060)	(-2.885)	(-2.132)	(-2.132)
mom60m	-0.011	-0.011	-0.011	-0.006	-0.006	-0.006
	(-3.025)	(-2.266)	(-2.266)	(-2.465)	(-1.865)	(-1.865)
ch_mom	-0.008	-0.008	-0.008	-0.010	-0.010	-0.010
	(-2.663)	(-2.014)	(-2.014)	(-4.490)	(-3.418)	(-3.418)
pr_delay	-0.007	-0.005	-0.005	-0.010	-0.010	-0.010
	(-2.300)	(-1.237)	(-1.237)	(-5.597)	(-4.615)	(-4.615)
ill	-0.012	-0.013	-0.013	-0.007	-0.005	-0.005
	(-5.135)	(-4.272)	(-4.272)	(-3.555)	(-1.829)	(-1.829)
std_vlm	-0.015	-0.014	-0.014	-0.008	-0.007	-0.007
	(-5.314)	(-4.316)	(-4.316)	(-4.233)	(-3.197)	(-3.197)
vlm	-0.011	-0.012	-0.012	-0.006	-0.006	-0.006
	(-4.582)	(-4.005)	(-4.005)	(-2.951)	(-2.514)	(-2.514)

Note: Intercept of a regression of each factors monthly excess return on the monthly equal weighted excess market return. Risk free rate given by the monthly selic rate. Columns (1) to (3) are the alphas of value weighted factors built using stocks sorted on the median, terciles and quintiles, respectively, and columns (4) to (6) are the alphas of equal weighted factors built using stocks sorted on the median, terciles and quintiles, respectively. Between parenthesis are the t-stats.

Table 8: Variables used to construct long-short portfolios.

Model	Fixed Chacteristics	Selected characteristics	Method
One-characteristic	-	earn_pr, i_ret_v, ill, mom12m, mom12m2, mom6m, mom6m2, ret_v, sl_pr, sz	Cycle each and make ew port.
CAPM	beta	earn_pr, i_ret_v, ill, mom12m, mom12m2, mom6m, mom6m2, op_pft, ret_v, sz	Cycle each and make ew port.
FF3	beta, btm, sz	earn_pr, gr_sl, i_ret_v, ill, mom12m, mom12m2, mom6m, mom6m2, ret_v	Cycle each and make ew port.
FF5	beta, btm, sz, asset_gr, op_pft	earn_pr, i_ret_v, ill, mom12m, mom12m2, mom6m, mom6m2, ret_v	Cycle each and make ew port.
FF5 + Mom	beta, btm, sz, asset_gr, op_pft, mom12m	i_ret_v, ill, ret_v	Cycle each and make ew port.
LASSO	-	asset_gr, btm, i_ret_v, mom1m, mom12m, ill, std_vlm	Compute directly w/ all
Non Parametric LASSO	-	i_ret_v, mom1m, mom12m, mom60m, pr_delay	Compute directly w/ all
Random Forest	-	mom1m, mom6m, mom12m, ret_v, ch_mo	Compute directly w/ all
All characteristics	-	btm, ill, ret_v, vlm	Compute directly w/ all

Table 9: Summary Statistics for Investment Strategies.

This table presents annualized Mean Return, Standard Deviation and Sharpe Ratio and is sorted by Sharpe Ratio. Mean strategy is the mean of all nine strategies.

Strategy	Mean	SD	SR
Random Forest	0.20	0.15	1.30
Mean	0.14	0.14	1.03
FF5 + Mom	0.18	0.18	1.02
Non Parametric Lasso (Freyberger)	0.17	0.17	1.01
FF5	0.16	0.17	0.96
Lasso	0.15	0.17	0.91
One Characteristic	0.09	0.10	0.83
FF3	0.14	0.17	0.80
CAPM (Beta)	0.11	0.17	0.64
All Characteristics	0.06	0.17	0.37

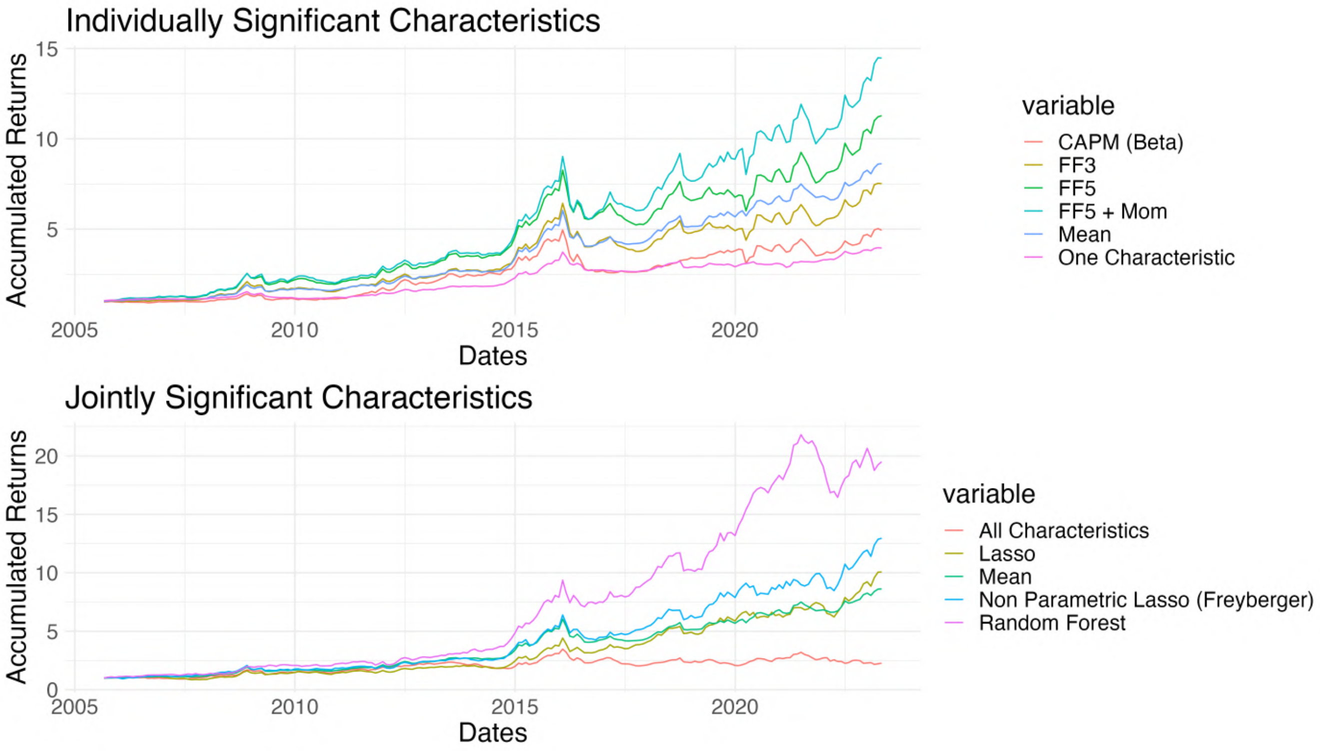


Figure 16: Return of long short strategy per model

In this exercise, we consider significant characteristics to build long-short portfolios. In top figure, we run 60-month rolling OLS regressions with each significant characteristic and the fixed model characteristics and then compute an equal-weight portfolio of all. In the bottom figure, the models jointly select characteristics and we run 60-month rolling OLS regressions and compute the returns. Both figure contains the same mean, which is the mean of all 9 methods.