

Emerson Sousa Vieira

**Machine Learning Methods in Asset Pricing: An
Analysis of Cross-sectional Stock Returns with
Macroeconomic Factors in Brazil**

Sao Paulo - Brazil

2025

Emerson Sousa Vieira

**Machine Learning Methods in Asset Pricing: An Analysis
of Cross-sectional Stock Returns with Macroeconomic
Factors in Brazil**

Thesis presented to the program of Professional Master in Economics as a partial requirement to obtain the degree of Master in Economics

Inspere

Programa de Mestrado Profissional em Economia

Supervisor Professor PhD. Gustavo Barbosa Soares

Sao Paulo - Brazil

2025

Emerson Sousa Vieira

Machine Learning Methods in Asset Pricing: An Analysis of Cross-sectional Stock Returns with Macroeconomic Factors in Brazil/ Emerson Sousa Vieira. - Sao Paulo - Brazil: 2025

44 p.

Professional Master Thesis: Insper

Programa de Mestrado Profissional em Economia , 2025.

Supervisor Professor PhD. Gustavo Barbosa Soares

1. Empirical Asset Pricing. 2. Macroeconomic Factors. 3. Machine Learning.
I. Emerson Sousa Vieira. II. Machine Learning Methods in Asset Pricing: An Analysis of Cross-sectional Stock Returns with Macroeconomic Factors in Brazil.

Emerson Sousa Vieira

Machine Learning Methods in Asset Pricing: An Analysis of Cross-sectional Stock Returns with Macroeconomic Factors in Brazil

Thesis presented to the program of Professional Master in Economics as a partial requirement to obtain the degree of Master in Economics

Banca examinadora

**Professor PhD. Gustavo Barbosa
Soares**
Supervisor

Professor PhD. Ruy Monteiro Ribeiro
Examiner

**Professor PhD. Bernardo de Oliveira
Guerra Ricca**
Examiner

**Professor PhD. Fernando Tassinari
Moraes**
Examiner

Sao Paulo - Brazil
2025

Abstract

The literature on financial machine learning has growth rapidly with studies encompassing several asset classes, chiefly for the stock market. We apply machine learning methods to Brazilian stocks cross section of monthly excess returns by using Brazilian stock factors while adding a vast set macroeconomic ones as our research primary contribution, for which the literature on Brazilian equities is scarce. We confirm recent results that ML models drive a substantial improvement in out-of-sample R^2 predictive power over traditional OLS models. Running an out-of-sample variable importance analysis, we also found macroeconomic factors overweight firm-related ones, with a slight predominance of country risk (EMBI Brazil Index), followed by the expectations of economics conditions, and Brazil's commodities composite index, and credit-to-GDP ratio. Our findings suggest a high relevance of macroeconomic factors when predicting monthly excess returns for Brazilian stocks. (JEL C52, C55, C58, G0, G1, G17).

Keywords: Brazilian stocks, empirical asset pricing, macroeconomic factors, machine learning methods

List of Figures

Figure 1 – Regression Tree example Kelly e Xiu (2023)	20
Figure 2 – Evolution of the number of stocks across the cross-section	24
Figure 3 – Description of Brazil equity market risk and macroeconomic factors	27
Figure 4 – Description of Brazil equity market risk and macroeconomic factors	28
Figure 5 – Histogram of firm characteristics and factors	29
Figure 6 – Boxplot statistics of Brazil equity market risk and macroeconomic factors	30
Figure 7 – Factors correlation matrix	31
Figure 8 – Out-of-Sample R^2 for different subsets of stocks.	34
Figure 9 – Time-varying model complexity for various methods. For ENet we report the number of features selected to have nonzero coefficients; for PCR and PLS, we report the number of selected components; for RF, we report the average tree depth; and, for GBRT, we report the number of distinct characteristics entering into the trees. ENet, PLS and PCR used substantially more non-zero components (i.e. minimum 6 and maximum of 22) than did GBRT and RF, which topped at four.	35
Figure 10 – ElasticNet OOS Variable Importance	37
Figure 11 – PLS OOS Variable Importance	37
Figure 12 – PCR OOS Variable Importance	38
Figure 13 – Gradient Boosted Regression Tree OOS Variable Importance	38
Figure 14 – Random Forest OOS Variable Importance	39
Figure 15 – 3-hidden-layer Neural Network OOS Variable Importance	39
Figure 16 – All Models OOS Variable Importance	40

List of Tables

Table 1 – Hyperparameters tuning	32
Table 2 – R_{OOS}^2 performance metrics for different models.	33

Contents

1	INTRODUCTION	9
2	LITERATURE REVIEW	10
2.0.1	Price and Return	10
2.0.2	Modern Asset Pricing Theory	10
2.0.2.1	The Stochastic Discount Function (SDF)	11
2.0.3	Traditional Empirical Asset Pricing Models	12
2.0.3.1	Capital Asset Pricing Model (CAPM)	13
2.0.3.1.1	Empirical Evidence and Criticisms	13
2.0.3.2	Arbitrage Pricing Theory (APT)	14
2.0.3.2.1	Empirical Evidence and Criticisms	14
2.0.3.3	Multifactor Models - Fama-French three-and-five factor models	15
2.0.3.3.1	Empirical Evidence and Criticisms	15
2.0.4	Machine Learning Methods in Empirical Asset Pricing	16
2.0.4.1	Why Machine Learning (ML)?	16
2.0.4.2	ML Models	16
2.0.4.2.1	Penalized Linear Models: Elastic Net, Ridge and LASSO	16
2.0.4.2.2	Dimension Reduction: Principal Component Regression (PCR) and Partial Least Squares (PLS)	18
2.0.4.2.3	Regression Trees: Boosted regression trees and Random Forests	19
2.0.4.2.4	Boosting Regression Trees	20
2.0.4.2.5	Random Forests	21
2.0.4.3	Performance Evaluation through R^2_{OOS}	22
2.0.4.4	Variable importance from individual predictors	22
2.0.5	The Case for Brazilian Stocks	23
3	AN EMPIRICAL STUDY OF BRAZILIAN STOCKS	24
3.1	Data description	24
3.1.1	Data treatment	25
3.1.2	Training, validation and test sets	25
3.1.3	Predictor set	26
3.1.4	Hyperparameter tuning	32
3.1.5	Data Sampling	32
3.2	Results	32
3.2.1	The cross-section of individual stocks	32
3.2.2	The relevance of covariates	36

4	CONCLUSION	41
	REFERENCES	42

1 Introduction

Asset pricing is a fundamental concept in finance, focused on determining the value of financial assets. Traditional models, such as the Capital Asset Pricing Model (CAPM) and the Fama-French multifactor models, have provided significant insights but often fall short in capturing the intricate dynamics of financial markets. Machine learning offers powerful tools to address these limitations by handling large datasets and identifying non-linear patterns, while improving predictive performance.

The focus of our research is to conduct a comprehensive examination of a broad array of machine learning methods to investigate the behavior of expected stock returns, with particular emphasis on comparative analysis among these methods while introducing a broad set of Brazilian macroeconomic factors.

The literature in Financial Machine Learning (see a comprehensive overview [Bagnara \(2024\)](#) and [Kelly e Xiu \(2023\)](#)) is still nascent but has grown rapidly. Whilst research using US data sets is somehow abundant, research on ML methods for Brazilian stocks with a focus on macroeconomic factors is significantly scarcer. Recent work of [Ribeiro et al. \(2024\)](#) for instance investigate the impact of firm characteristics on stock returns in the Brazilian financial market, examining over 24 firm-level characteristics employing techniques such as Fama-MacBeth regressions, advanced machine learning techniques such as LASSO, non-parametric LASSO and Random Forest analysis. [Arismendi-Zambrano, Genaro e Alexandre \(2023\)](#) study intraday stock returns using Ridge Regression, LASSO, elastic net, PCR and PLS. [Carosia, Silva e Coelho \(2024\)](#) combine the use of historical stock prices, financial technical indicators, and financial news into Deep Learning models for select Brazilian stocks. [Silva e Sartiro \(2024\)](#) study the predictive power of CatBoost, gradient boosting, AdaBoost, LightGBM and XGBoost models of stock return. [Ferreira, Gandomi e Cardoso \(2020\)](#) use Decision Trees and Convolutional Networks in analyzing the trend of future financial asset price movements.

The research done so far lack the application of a wider range of ML models to Brazilian stocks as seen in [Gu, Kelly e Xiu \(2020\)](#) for the US market and do not incorporate Brazilian macroeconomic factors in their models. We intend to incorporate such innovations in our work as the primary contributions.

2 Literature Review

In this section we start by providing the basic definitions for measuring returns and a quick review on Stochastic Discount Factor model under Modern Asset Pricing Theory, which provides a general framework for pricing assets. Then, we conduct a literature review exploring both traditional (econometrics-based) and nonstandard statistical tools (Machine Learning-based) techniques in analyzing cross-section returns for Brazilian equities. We also review recent results of machine learning (ML) methods in empirical asset pricing carried out for developed markets data and for Brazilian equities. We also intend to show that the span of ML models used so far Brazilian equities data set is scarce and fail to use a larger set or predictors that include Brazilian macroeconomic factors, which is the main innovation proposed in this work.

2.0.1 Price and Return

One stream of literature tries to identify predictors for the cross-section of stock returns. Starting from a general additive prediction error model:

$$r_{i,t+1} = \mathbb{E}_t[r_{i,t+1}] + \epsilon_{i,t+1} \quad (2.0.1)$$

where $r_{i,t+1}$ is the return of stock i in excess of the risk-free rate between time t and $t + 1$. The conditional mean $\mathbb{E}_t[r_{i,t+1}]$ is often modeled as an unknown function $g * (\cdot)$ that tries to capture and maximize the explanatory power the predictor set for the observed realized returns. $z_{i,t}$ is a M -dimensional vector that contains the predictor set. Note that $g * (\cdot)$ has not predefined functional form, which shall be implicitly obtained from the tested machine learning models. The model can be expressed in Equation 2.0.2 as:

$$r_{i,t+1} = g * (z_{i,t}; \theta) + \epsilon_{i,t+1} \quad (2.0.2)$$

As we shall see below, the function for $g * (z_{i,t}; \theta)$ can assume virtually any form. Under traditional empirical asset pricing models, it has usually consisted of a single or a range of predictors in θ expressed in a linear manner (i.e. CAPM, APT and other factor models), while ML methods have allowed for the exploration of non-linearity in the dataset, usage a large set of predictors and the inclusion of penalty functions.

2.0.2 Modern Asset Pricing Theory

Over the past four decades, the greatest endeavor in Asset Pricing Theory has been documenting the properties of the stochastic discount factor (SDF), which allows to price

any asset with unknown future payoff, with the goal of understanding the determinants of asset returns. We lay out the core below, while a textbook treatment can be seen in [Cochrane \(2005\)](#) and [Back \(2017\)](#).

2.0.2.1 The Stochastic Discount Function (SDF)

Given the importance of such concept in Asset Pricing Theory, we derive the SDF equation below.

Consider an investor who seeks to maximize their expected utility of consumption over multiple periods. The utility function $u(C_t)$ represents the investor's preferences, where C_t denotes consumption at time t . The objective of the investor is to maximize the expected utility of lifetime consumption:

$$\max \mathbb{E}_0 \left[\sum_{t=0}^T \beta^t u(C_t) \right] \quad (2.0.3)$$

where \mathbb{E}_0 is the expectation based on information available at time 0, β is a discount factor, $0 < \beta < 1$, T is the terminal period.

The investor's consumption choices are subject to a budget constraint, which can be expressed as:

$$W_{t+1} = (W_t - C_t)(1 + R_{t+1}) + Y_{t+1} \quad (2.0.4)$$

where W_t is the wealth at time t , C_t is the consumption at time t , R_{t+1} is the return on the investment between t and $t + 1$, Y_{t+1} is the income received at time $t + 1$.

To derive the SDF, we start with the first-order condition for utility maximization, which leads to the Euler equation. The investor chooses consumption C_t to maximize expected utility, subject to the budget constraint. The Lagrangian for this problem is:

$$\mathcal{L} = \mathbb{E}_0 \left[\sum_{t=0}^T \beta^t u(C_t) \right] + \sum_{t=0}^T \lambda_t [W_{t+1} - (W_t - C_t)(1 + R_{t+1}) - Y_{t+1}] \quad (2.0.5)$$

Taking the derivative with respect to C_t and setting it to zero, we obtain the Euler equation:

$$\frac{\partial \mathcal{L}}{\partial C_t} = \beta^t u'(C_t) - \lambda_t(1 + R_{t+1}) = 0 \quad (2.0.6)$$

Solving for λ_t , we get:

$$\lambda_t = \frac{\beta^t u'(C_t)}{1 + R_{t+1}} \quad (2.0.7)$$

At time $t + 1$, the Euler equation is:

$$\lambda_{t+1} = \frac{\beta^{t+1} u'(C_{t+1})}{1 + R_{t+2}} \quad (2.0.8)$$

Using the budget constraint and the definition of λ_t and λ_{t+1} , we can derive the relationship between them. Since $\lambda_t = \mathbb{E}_t[\lambda_{t+1}]$:

$$\lambda_t = \mathbb{E}_t \left[\frac{\beta^{t+1} u'(C_{t+1})}{1 + R_{t+2}} \right] \quad (2.0.9)$$

By definition, the Stochastic Discount Factor M_{t+1} is the ratio of the marginal utility of consumption at time $t + 1$ to the marginal utility of consumption at time t , adjusted by the subjective discount factor β :

$$M_{t+1} = \beta \frac{u'(C_{t+1})}{u'(C_t)} \quad (2.0.10)$$

This expression represents the SDF, which discounts future cash flows to their present value, accounting for time preferences (β) and changes in marginal utility ($\frac{u'(C_{t+1})}{u'(C_t)}$).

Most existing pricing methods, as the ones shown in sections 2.0.3 and 2.0.4, are particular versions of the SDF.

2.0.3 Traditional Empirical Asset Pricing Models

Traditional asset pricing models have been central to financial economics, providing foundational frameworks for understanding the relationship between risk and return. These models include for instance the Capital Asset Pricing Model of [Sharpe \(1964\)](#), the Arbitrage Pricing Theory of [Ross \(1976\)](#), and multifactor models like the Fama-French three-factor [Fama e French \(1992\)](#) and five-factor models [Fama e French \(2015\)](#). Nonetheless, such models have been designed under a traditional econometric approach to financial market research that i) first specifies a functional form for the return forecasting model motivated by a theoretical economic model, then ii) estimates parameters to understand how candidate information sources associate with observed market prices within the confines of the chosen model. As we shown in more detail in the [Modern Asset Pricing Theory](#) section , the flexibility of the pricing Kernel (or alternatively Stochastic Discount Function, SDF) on modern financial analysis allows for a wide variety of structural economic assumptions, which implies that there is no consensus about which specific structural formulations are viable (i.e. what is the optimal and most statistically significant set of predictors? Should we embed a linear or non-linear structure to the observed data? Should we consider interactions among predictors? Are they are static or time-variant?) as opposed to the more rigid specifications under traditional models. Indeed, it does come as no surprise

that traditional models have failed to match market price data as shown in Mehra e Prescott (1985) describing the *equity premium puzzle*. Albeit the shortcomings faced by the traditional empirical asset pricing models on the predictive power of asset returns, it is key to laid them out to better grasp the advantages brought by ML methods.

2.0.3.1 Capital Asset Pricing Model (CAPM)

Developed independently by Sharpe (1964), Lintner (1965), and Mossin (1966) the CAPM extends the principles of portfolio theory introduced by Markowitz (1952), providing a model that establishes a linear relationship between expected returns of assets and their systematic risk, as measured by beta.

The CAPM posits that the expected return on an asset is linearly related to its systematic risk, which is non-diversifiable. The fundamental equation of CAPM is expressed as:

$$E(R_i) = R_f + \beta_i (E(R_m) - R_f) \quad (2.0.11)$$

where $E(R_i)$ is the expected return on asset i , R_f is the risk-free rate, β_i is the asset's beta, and $E(R_m)$ is the expected return of the market portfolio.

2.0.3.1.1 Empirical Evidence and Criticisms

The stability of beta over time is a critical assumption of CAPM. However, empirical research has shown that beta values can change significantly over time, affecting the reliability of the model's predictions as seen in Blume (1971). Moreover, CAPM relies on a single market factor to explain asset returns. Multifactor models, such as the Fama-French three-and-five factor models, discussed below, have been developed to address the limitations of CAPM by incorporating additional factors like size and value. Furthermore, empirical studies have identified several anomalies that CAPM fails to explain:

- **Size Effect:** Smaller firms tend to have higher risk-adjusted returns than larger firms
- **Value Effect:** Firms with high book-to-market ratios outperform those with low book-to-market ratios
- **Momentum Effect:** Stocks that have performed well in the past continue to perform well in the short term

2.0.3.2 Arbitrage Pricing Theory (APT)

APT and CAPM share the goal of explaining the expected returns of assets, but they differ significantly in their approach. Introduced by [Ross \(1976\)](#), APT provides an alternative model specification by relaxing some of its restrictive assumptions and allowing for multiple sources of systematic risk based on the notion that asset returns can be modeled as a linear function of various macroeconomic factors. Empirical studies have used various macroeconomic variables such as inflation, industrial production, interest rates, and exchange rates as factors. [Chen et al. \(1986\)](#) conducted a seminal study that identified these macroeconomic factors as significant in explaining stock returns. The key assumptions of APT include:

1. **No Arbitrage:** In equilibrium, there are no arbitrage opportunities; otherwise, they would be exploited by investors until eliminated.
2. **Linear Factor Structure:** Asset returns are generated by a linear combination of several risk factors.
3. **Idiosyncratic Risk:** Each asset has a unique component of risk (idiosyncratic risk) that is uncorrelated with the factors and can be diversified away in a large portfolio.

The APT model is specified as:

$$E(R_i) = R_f + \beta_{i1}F_1 + \beta_{i2}F_2 + \cdots + \beta_{ik}F_k \quad (2.0.12)$$

where $E(R_i)$ is the expected return on asset i , R_f is the risk-free rate, β_{ij} is the sensitivity of asset i to factor j , and F_j is the risk premium associated with factor j .

2.0.3.2.1 Empirical Evidence and Criticisms

Empirical tests of APT involve estimating the sensitivities (betas in equation 2.0.12) of asset returns to the identified factors and examining whether the estimated model fits the observed returns. While some studies have shown that APT provides a better fit than CAPM, others highlight difficulties in factor selection and model estimation as shown in [Shanken \(1982\)](#). Among the main issues with APT, one can mention:

1. **Factor Selection** - Identifying the correct set of factors is a significant limitation of APT. The model does not specify which factors should be included, leading to potential overfitting or underfitting depending on the factors chosen.
2. **Empirical Implementation** - Moreover, different studies often identify different sets of factors, leading to inconsistent results.

3. **Market Anomalies** - Like other asset pricing models, APT struggles to explain certain market anomalies, such as momentum and the low volatility anomaly. These anomalies suggest that additional factors or alternative models might be needed to fully capture the dynamics of asset returns.

2.0.3.3 Multifactor Models - Fama-French three-and-five factor models

Eugene F. Fama and Kenneth R. French have significantly contributed to asset pricing theory through their influential research. Their papers [Fama e French \(1992\)](#) and [Fama e French \(2015\)](#) represent key milestones in the development of multifactor asset pricing models challenging the traditional CAPM. They utilize regression analysis to assess the explanatory power of various factors, including market factor (excess return on the market portfolio), firm size (captures the return spread between small and large firms), and value factor (captures the return spread between high book-to-market and low book-to-market firms) in the three-factor model. In the five-factor model, they extend the three-factor counterpart by including profitability and investment factors, thus providing a more comprehensive explanation of stock returns. The data set used NYSE, AMEX, and NASDAQ stocks from 1963 to 1990 and 1963 to 2013. The model specifications follow as [2.0.13](#) for the 3-factor model and [2.0.14](#) for the 5-factor.

$$E(R_i) - R_f = \beta_{iM} (E(R_m) - R_f) + \beta_{iSMB} \cdot SMB + \beta_{iHML} \cdot HML \quad (2.0.13)$$

$$E(R_i) - R_f = \beta_{iM} (E(R_m) - R_f) + \beta_{iSMB} \cdot SMB + \beta_{iHML} \cdot HML + \beta_{iRMW} \cdot RMW + \beta_{iCMA} \cdot CMA \quad (2.0.14)$$

2.0.3.3.1 Empirical Evidence and Criticisms

While the three-factor and five-factor models have significantly advanced the understanding of asset pricing, they are not without their limitations. Both models face challenges related to omitted factors, empirical anomalies, complexity, and international applicability. Additionally, the potential redundancy of factors and the assumption of linear relationships pose further issues. Those issues are well-documented in [N. e Titman \(2015\)](#), [Ang et al. \(2006\)](#), [Sloan \(1996\)](#), [Novy-Marx \(2013\)](#), and [Asness, Moskowitz e Pedersen \(2013\)](#).

2.0.4 Machine Learning Methods in Empirical Asset Pricing

2.0.4.1 Why Machine Learning (ML)?

The asset pricing literature has over time documented a wide variety of factors purportedly predicting cross-sectional variation in expected returns in [Harvey, Liu e Zhu \(2016\)](#) and [Hou, Xue e Zhang \(2015\)](#). This extensive collection of predictors, often defined as the *factor zoo* has faced criticism due to doubts about the actual usefulness of the proposed factors and the challenges associated with incorporating numerous sources of risk into models. While this abundance of predictors promises enhanced return predictability, it also raises several important questions on which variables are truly significant, how do they interact, how can they be effectively integrated into models, to name a few.

At an intuitive level, machine learning is a way to pursue statistical analysis when the analyst is unsure which specific structure their statistical model should take. ([KELLY; XIU, 2023](#))

This growing body of research has garnered considerable attention, as the results achieved in several cases have been outstanding, providing new insights into the structure of the Stochastic Discount Factor (SDF). Thus ML has been recognized as a powerful alternative to traditional approaches, with relevant potential for future applications. A thorough understanding of these models is crucial for those wishing to stay at the forefront of financial research.

Thus, we follow [Gu, Kelly e Xiu \(2020\)](#) and use the term *machine learning* to describe a collection of high-dimensional models for statistical prediction, combined with regularization methods for model selection and mitigation of overfit, and efficient algorithms for searching among myriad number of potential model specifications. Put it another way, machine learning offers a way to perform statistical analysis when the exact structure of the statistical model is not clear.

2.0.4.2 ML Models

Following [Bagnara \(2024\)](#) , one could group recent studies into five categories according to the main ML approach they adopt, namely: i) regularization, ii) dimension reduction, iii) regression trees, iv) random forest (RF), v) and neural networks. We include a subsection to each ML area we use in this work with a brief introduction to its mechanics.

2.0.4.2.1 Penalized Linear Models: Elastic Net, Ridge and LASSO

The simple linear model imposes that conditional expectations $g^*(\cdot)$ can be approximated by a linear function of the raw predictor variables and the parameter vector

θ ,

$$g(z_{i,t}; \theta) = z'_{i,t} \theta. \quad (2.0.15)$$

This model imposes a simple regression specification and does not allow for nonlinear effects or interactions between predictors. The loss function for the ordinary OLS is given by 2.0.16

$$\mathcal{L}(\theta; z_{i,t}) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (r_{i,t+1} - z'_{i,t} \theta)^2 \quad (2.0.16)$$

OLS models, while simple and parsimonious, tend to show poor out-of-sample performance, becoming inefficient or inconsistent when there are many predictors used and they surpass the number of observations in the dataset. Moreover, beyond the issues with the model structure, realized stock returns typically carry low signal-to-noise ratios, so that OLS tend to over fit noise. A key strategy for preventing over fitting is to minimize the number of parameters being estimated. The most prevalent machine learning method for enforcing parameter simplicity is to incorporate a penalty into the objective function, promoting more parsimonious models.

The statistical model for the penalized linear model is the same as the simple linear model in 2.0.16, but it incorporates a penalty $\phi(\theta; \cdot)$ to the objective loss function defined in 2.0.18

$$\mathcal{L}_{penalized}(\theta; z_{i,t}) = \mathcal{L}(\theta; z_{i,t}) + \phi(\theta, \lambda, \rho) \quad (2.0.17)$$

$$\phi(\theta, \lambda, \rho) = \lambda(1 - \rho) \sum_{j=1}^P |(\theta_j)| + \frac{1}{2} \lambda \rho \sum_{j=1}^P (\theta_j)^2 \quad (2.0.18)$$

Equation 2.0.18 is defined as elastic net, introduced by Zou e Hastie (2015), allows for two types of models depending on the value of ρ . When $\rho = 1$ the model is named the Ridge Regression Hoerl e Kennard (1970); when $\rho = 0$ we have the *Least Absolute Shrinkage and Selection Operator (LASSO)* model Tibshirani (1996) ¹

By definition, the variables λ and ρ are the so-called *hyperparameters* of the model, which shall be determined through a cross-validation procedure².

Diving a bit deeper into Ridge regression, LASSO and λ :

¹ According to Hastie Robert Tibshirani (2013) LASSO models show equivalence to Support Vector Machine, which we do not include in our study.

² Cross-validation textbook information can be found in Hastie Robert Tibshirani (2013) on pages 241-247

- Ridge regression ($\rho = 1$): it penalizes the estimator for large values of *theta*, performing *shrinkage*, i.e., draws all estimates coefficients in the OLS model **towards** zero, but retaining them.
- LASSO ($\rho = 0$): it *forces* some of the estimated coefficients in the OLS model to *zero*, instead of shrinking them.
- λ : this hyperparameter controls for the amount of shrinkage. When $\lambda = 0$ there is not imposition of penalization, but as it increases, the model flexibility decreases, leading to higher bias but lower variance.

2.0.4.2.2 Dimension Reduction: Principal Component Regression (PCR) and Partial Least Squares (PLS)

The penalized linear models utilize shrinkage and variable selection to handle high dimensionality by imposing predictors X_p close (Ridge) to or exactly to zero (LASSO).

However, [Gu, Kelly e Xiu \(2020\)](#) calls the attention that this approach can result in less effective forecasts when the predictors are highly correlated.

The basic idea behind *dimension reduction* methods relies in the transformation of the original predictor set of size p into a $M < p$ linear combination of those p predictors and then fit a least squares model using the transformed predictor set. Thus, instead of estimating $p + 1$ coefficient, the model estimates $M + 1$. According to [Bagnara \(2024\)](#), the PCR method is the one applied the most in asset pricing, which we kick-off with.

The *principal components regression (PCR)*³ method is based on the construction of the first M principal components, Z_M , and then using these components as the predictors in a linear regression model that is fitted using least squares. The central concept is that a limited number of principal components are frequently enough to capture the majority of the variability within the data and adequately explain the relationship with the response variable. Under this approach, the implicit assumption is that the directions in which the original predictors X_p present the most variation are the ones associated with Y .⁴ The main advantage of this approach is that it theoretically mitigates over fitting as all relevant information is contained within Z_M . An example of such principal component regression (PCR) can be found in [Gu, Kelly e Xiu \(2020\)](#).

According to [Gu, Kelly e Xiu \(2020\)](#) and [Hastie Robert Tibshirani \(2013\)](#), a limitation of PCR is that, because it is based on a unsupervised method (i.e. the response

³ The PCR is based in the Principal Components Analysis model, which is the per se method for computing the principal components in the predictor set. The first principal component direction of the data is that along which the observations vary the most.

⁴ Note that such assumption is not always true, but it often is a good approximation according to [Hastie Robert Tibshirani \(2013\)](#)

Y is not used to help determine the principal component directions), one cannot assure that the directions that best explain the predictors will also best explain the response Y. In other words, it does not integrate the primary statistical goal of forecasting returns during the dimension reduction phase⁵

The *Partial Least Squares method (PLS)* in contrast to PCR, PLS follows a supervised methodology, achieving dimension reduction by directly leveraging the covariation between predictors and the response Y.

For each predictor p , one estimates its univariate return prediction coefficient using conventional OLS. The estimated coefficient ϕ_p , indicates the partial sensitivity of returns to each predictor p . Subsequently, one averages all predictors into a single aggregate component, assigning weights proportional to ϕ_p . Thus, the absolute largest univariate predictors receive the highest weights, while the weakest receive the lowest. Accordingly, PLS performs dimension reduction with the primary objective of forecasting returns. [Kelly e Pruitt \(2013\)](#) and [Kelly e Pruitt \(2015\)](#) show, for instance, for asymptotic theory of PLS regression and its application to forecasting risk premiums in financial markets.

2.0.4.2.3 Regression Trees: Boosted regression trees and Random Forests

Modern asset pricing models exhibit a high degree of state dependence in financial market behaviors. This suggests that including interaction effects in empirical models is potentially important as evidenced in [Hong, Lim e Stein \(2000\)](#), which found that momentum factor, for instance, interacts with firm size.

While one could in a straightforward manner incorporate such interactions in a OLS model, computational costs would increase dramatically becoming infeasible because the multi-way interactions increase the number of parameters combinatorially, further the previously discussed factor zoo issue. Note that penalization discussed above does not solve the difficulty of estimating linear models when the number of predictors is larger than observations ($p \gg n$).

Regression trees come as an alternative as it allows the incorporation of myriad predictor interactions at significantly lower computational costs. The basic idea of regression trees is that, as they partition data observations into groups that share common feature interactions, we can use past data for a given group to forecast the behavior of a new observation that arrives in the group. Overall, the tree stratifies or segments the predictors into regions of predictor space.

Intuitively, we follow the example of [Kelly e Xiu \(2023\)](#) to show in [Figure 1](#) how regression trees work for a set a two predictors, namely *size* and *value or book-to-market*

⁵ Principal Component Analysis (PCA) compacts data into components based on the covariation among the predictors. This process occurs before the forecasting step and does not take into account the relationship between the predictors and future returns.

(b/m) factors.

1. observations are sorted on size.
2. those above the breakpoint of 0.5 are assigned to Category 3.
3. those with small size are then further sorted by b/m
4. observations with small size and b/m below 0.3 are assigned to Category 1
5. those with b/m above 0.3 go into Category 2.
6. forecasts for observations in each partition are defined as the simple average of the outcome variable's value among observations in that partition.

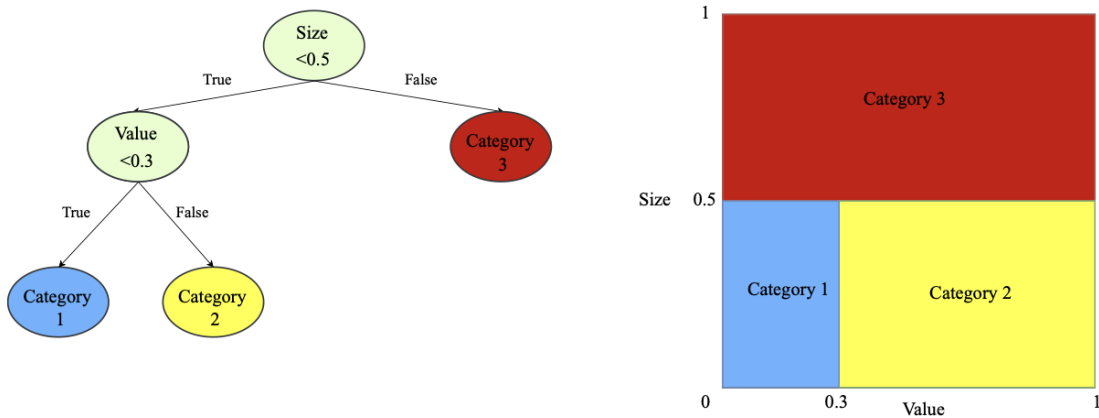


Figure 1 – Regression Tree example Kelly e Xiu (2023)

More formally, the prediction of a tree, T , with K “leaves” (terminal nodes), and depth L , can be written as 2.0.19

$$g(z_{i,t}; \theta, K, L) = \theta_k \mathbf{1}\{z_{i,t} \in C_k(L)\} \quad (2.0.19)$$

where $C_k(L)$ is one of the K partitions of the data. Each partition is a product of up to L indicator functions of the predictors. The constant associated with partition k (denoted θ_k) is defined to be the sample average of outcomes within the partition.

2.0.4.2.4 Boosting Regression Trees

While regression trees are quite flexible, they are among the prediction methods most prone to overfit, and therefore must be regularized.

To address this issue, trees are typically used in regularized ensembles. A common ensembling method is boosting [E.Schapire \(1990\)](#) and [Friedman \(2001\)](#), specifically Gradient Boosted Regression Trees (GBRT). This technique recursively combines forecasts from many individual trees, which are shallow and weak predictors on their own, but together form a single strong predictor. In other words, instead of "*fitting the data hard*", the boosting approach "*learns slowly*", according to [James et al. \(2023\)](#).

The boosting procedure starts by fitting a shallow tree (depth of $L = 1$). Next, a second shallow tree is used to fit the prediction *residuals* from the first tree. Forecasts from these two trees are added together to form an ensemble prediction of the outcome, but the forecast component from the second tree is shrunk by a factor $\nu \in (0, 1)$ to help prevent the model from overfitting the residuals. This process is repeated iteratively to form an additive ensemble of B shallow trees. As a result, the boosted ensemble has L , ν , and B as hyperparameters.

2.0.4.2.5 Random Forests

To understand Random Forests we first need to have a look at the concepts of *Bagging or Bootstrap aggregation*, introduced by [Breiman \(2001\)](#). The decision trees discussed above suffer from high variance (i.e. meaningful difference in predictions for randomly different training sets). In contrast, a low variance method will yield similar results if applied repeatedly to distinct data sets. The bagging approach is based on the trivial concept that averaging a set of n observations reduces variance from σ^2 for each observation to $\frac{\sigma^2}{n}$ for the sample mean. Bagging is performed by sampling with replacement (i.e. bootstrapping) from the single training data set to generate B different bootstrapped training data sets. Then, we then train the method on each of the b -th bootstrapped training set and finally average all the predictions to obtain the final prediction with lower variance as it decorrelates the trees. Note that *bagging* can be applied to many regression methods, but it is particularly useful for decision trees according to [James et al. \(2023\)](#).

Random forests, similarly to *Bagging*, we construct several decision trees using bootstrapped training samples. In addition to that, when constructing these trees, each time a split is considered, a random sample of m predictors is selected from the full set of p predictors as split candidates. A new sample of $m \approx \sqrt{p}$ ⁶ predictors is taken at each split (i.e. this process, named *dropout* does not allow for the majority of predictors to be considered during the splits). This further decorrelates decision trees. A simple situation explains why this works on reducing variance and improving Bagging's predictive power: suppose there is one predictor which is the strongest among all predictors, while others have only moderate explanatory power. In this case, during the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Consequently, all

⁶ Note that if we choose $m = p$, then we fall in the Bagging case.

of the bagged trees will look quite similar to each other and highly correlated⁷. In a nutshell, process decorrelates the trees, thereby making the average of the resulting trees less variable and hence more reliable. The depth L of the individual trees, the number of bootstrap samples B , and the dropout rate are tuning parameters of the Random Forest model.

2.0.4.3 Performance Evaluation through R_{OOS}^2

We assess predictive performance for individual stock return forecasts, by calculating the out-of-sample R_{OOS}^2 , following [Gu, Kelly e Xiu \(2020\)](#).

$$R_{\text{OOS}}^2 = 1 - \frac{\sum_{(i,t) \in \mathcal{T}_3} (r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t) \in \mathcal{T}_3} r_{i,t+1}^2}, \quad (2.0.20)$$

In equation 2.0.20 the term τ_3 refers the testing samples, which has never been used for neither trained nor validation. R_{OOS}^2 pools forecasting errors across stocks and over time into a large panel-level assessment of each ML model.

An important feature calculation approach shown in Equation 2.0.20 is that is calculated without demeaning. In a traditional calculation, the denominator is often based on the variance or sum of squared deviations of the dependent variable from its mean (i.e., demeaning the data). In this case, the denominator is the sum of squared excess returns without demeaning, which means that instead of subtracting the mean of the returns before squaring, the raw excess return values are squared and summed directly. According to [Gu, Kelly e Xiu \(2020\)](#), predicting future excess stock returns with historical averages typically underperforms a naive forecast of zero by a relevant margin. This is avoided by benchmarking the R^2 against a forecast value of zero.

2.0.4.4 Variable importance from individual predictors

We seek to identify predictors that have relevant effect on the cross-section of stock return while controlling for the remaining ones. We do this by ranking them according to a notion of variable importance, which is obtained by the reduction in stock panel predictive R_{OOS}^2 from setting all values of predictor k to zero, while holding the remaining model estimates fixed.

⁷ Averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities, which entails that bagging might not lead to a substantial reduction in variance over a single tree in this setting.

2.0.5 The Case for Brazilian Stocks

The literature in Financial Machine Learning (see a comprehensive overview [Bag-nara \(2024\)](#) and [Kelly e Xiu \(2023\)](#)) is still nascent, but have grown rapidly for US data sets. Whilst research using US data sets is somehow abundant, research on ML methods for Brazilian stocks is significantly scarcer. Recent work of [Ribeiro et al. \(2024\)](#) for instance investigate the impact of firm characteristics on stock returns in the Brazilian financial market, examining over 24 firm-level characteristics employing techniques such as Fama-MacBeth regressions, advanced machine learning techniques such as LASSO, non-parametric LASSO and Random Forest analysis. [Arismendi-Zambrano, Genaro e Alexandre \(2023\)](#) study intraday stock returns using Ridge Regression, LASSO, elastic net, PCR and PLS. [Carosia, Silva e Coelho \(2024\)](#) combine the use of historical stock prices, financial technical indicators, and financial news into Deep Learning models for select Brazilian stocks. [Silva e Sartiro \(2024\)](#) study the predictive power of CatBoost, gradient boosting, AdaBoost, LightGBM and XGBoost models of stock return. [Ferreira, Gandomi e Cardoso \(2020\)](#) use Decision Trees and Convolutional Networks in analyzing the trend of future financial asset price movements.

The works done so far lack the application of a wider range of ML models to Brazilian stocks as seen in [Gu, Kelly e Xiu \(2020\)](#) for the US market and do not incorporate Brazilian macroeconomic factors in their models. We intend to incorporate such innovations in our work as the primary contributions.

3 An Empirical Study of Brazilian Stocks

3.1 Data description

We construct a panel consisting of i) monthly stocks returns from 312 stocks listed in the Brazilian stock exchange from January 2011 to April 2024, amounting to 134 months of data and 32 thousand observations for the 312 unique stocks. We also obtain the Brazil SELIC rate to proxy for the risk-free rate from which we calculate individual excess returns. Note that we follow [Ribeiro et al. \(2024\)](#) and [Gu, Kelly e Xiu \(2020\)](#), but taking a factor-based approach instead of firm-characteristics. We also introduce a wide set of macroeconomics factors from the Brazilian economy in our models in addition to US fed funds rate, given the widely known correlation of Brazilian equities with the US economy as seen in [Wongswan \(2009\)](#), [Tabak, Laiz e Cajueiro \(2017\)](#) and [Robitaille e Roush \(2006\)](#).

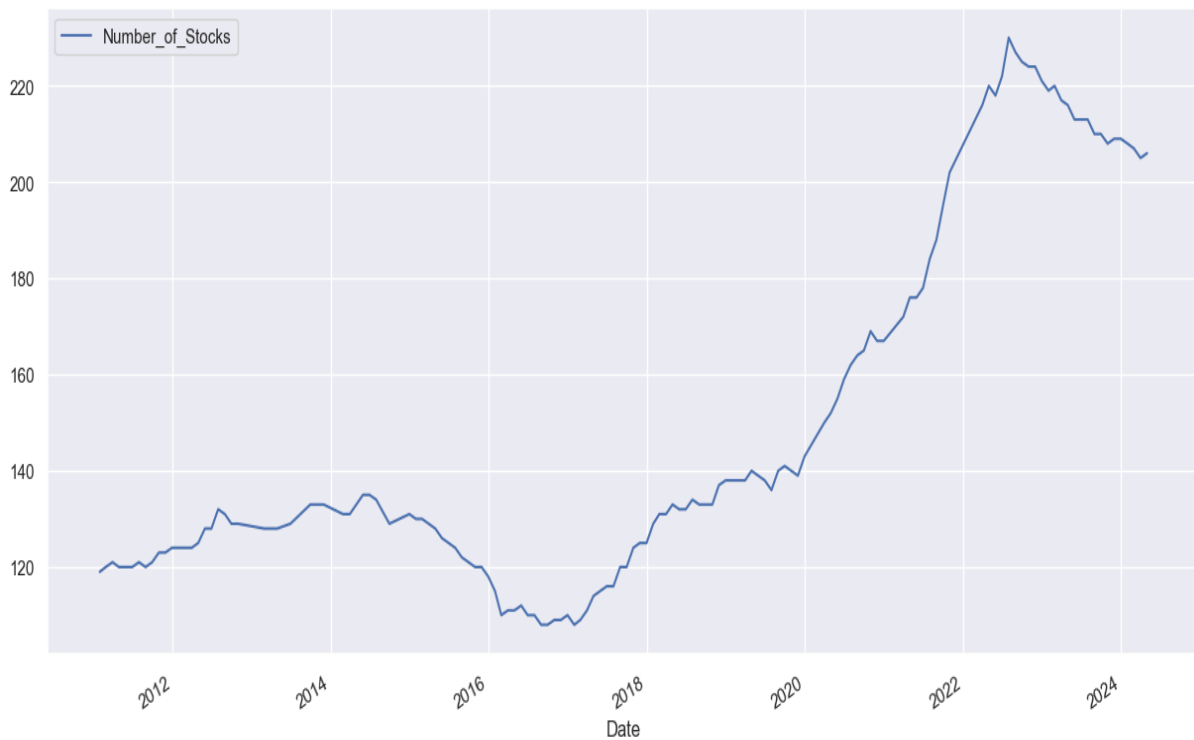


Figure 2 – Evolution of the number of stocks across the cross-section

For Brazilian stock returns, we use Eikon Refinitiv database. We used Brazil stock market risk factors following a mean-sorted equal-weighted portfolio scheme, available at [CEFIM \(2024\)](#) and calculated a wide of Brazilian macroeconomic factors based on innovations of the raw variables, which were obtained by several sources, including the Brazilian Central Bank Time Series Management System and the Brazilian Institute of Geography and Statistics (IBGE). When implementing our ML models (ElasticNet,

Principal Component Analysis, Partial Least Squares, Gradient Boosted Regression Trees and Random Forest), we closely follow [Gu, Kelly e Xiu \(2020\)](#) in terms of range of tested *hyperparameters*.

3.1.1 Data treatment

The stock return dataset used in [Ribeiro et al. \(2024\)](#) implements a liquidity filtering in which a company was considered eligible for inclusion if over the trailing 12-month period the stock i) traded on more than 80% of the trading days, ii) the average trading volume surpassed the R\$ 1 million mark, iii) the median daily volume trading volume surpassed R\$ 1 million, iv) the stock must have been listed for at least six months, v) the stock must have been available for trading in the month preceding the assessment, since return prediction exercise use 1-month lagged data. This methodology closely follows [Ribeiro et al. \(2024\)](#) for comparability and assures stocks includes in our analysis dataset show relevant market activity to enable empirical investigation.

When performing the analysis we apply a winsorization to the whole dataset excluding sample outliers, namely the top and bottom one-percentile. In contrast to [Gu, Kelly e Xiu \(2020\)](#) the dependent variable is also winsorized.

For handling missing data, only independent variables with less than 25% missing values are retained. The remaining missing values are imputed using the cross-sectional median for each stock within each month, following the approach outlined by [Gu, Kelly e Xiu \(2020\)](#).

3.1.2 Training, validation and test sets

In machine learning, splitting the data into train, validation, and test sets is a fundamental practice for building, evaluating, and fine-tuning predictive models. Each subset serves a distinct purpose which we detail. The training set is used to fit the machine learning model. It provides the model with the data it needs to learn patterns and relationships between the input features and the target variable. The validation set is used to evaluate the model during training and guide the selection of hyperparameters (e.g., learning rate, regularization strength, or the number of layers in a neural network). The test set is used only once, at the end of model development, to assess the model's performance on unseen data.

We estimate our models over a 12-month rolling window that contains 60 months for training, 24 months for validation, and for 12 months for testing. Our model coefficients are calculated in each month for a range of potential hyperparameters in the training set. We then use the validation set to compute predictions and calculate the mean squared error (MSE) for the different models test (each with a specific value for the hyperparameter).

In the test set, we use the optimal hyperparameters (i.e. those that minimize MSE in the validation set) to calculate predictions and calculate R_{OOS}^2 for the current testing rolling window. Finally, we aggregate all windows to obtain the model performance over the full test set period, following [Gu, Kelly e Xiu \(2020\)](#).

3.1.3 Predictor set

We selected a total of 53 factors, being 29 Brazil market risk factors from [CEFIM \(2024\)](#) following a median-sorted weighted average scheme and 24 Brazilian macroeconomic ones, calculated as monthly innovations over the raw variables. Sources for the macroeconomic variables were the Brazilian Central Bank (BCB), the Brazilian Institute of Geography and Statistics (IBGE) and collected data from data aggregator consultancy firm MCM Consultores Associados, which we describe in [Figure 3](#). Statistical description of the predictor set used is shown in [Figures 4, 5, 6, and 7](#).

Brazil Equity Market Risk Factors - Median Sorted Long-Short Portfolios		Anomaly	Key original references	Implementation details
Abbr.	Value vs Growth			
earn_pr	Earnings-to-price		Basu (1977)	Trailing 12-month net profit at t-5 to the market value of equity at t-1
b/m	Book-to-market ratio		Rosenberg et al. (1985)	Book value of equity at t-5 to market value of equity at t-1
sl_pr	Sales-to-price		Barbee et al. (1986)	Trailing 12-month sales at t-5 to the market value of equity at t-1
ebit_reve	Ebit-to-revenue		n.a.	Operating profit divided by revenue in t-12
earn_rev	Earnings-to-revenue		n.a.	Net earnings divided by revenue at t-12
sz	Market Value		Banz (1981)	Natural log of market capitalization at end of month t - 1
ill	Illiquidity		Amihud (2002)	Average of daily (absolute return / volume)
lev2	Leverage		Bhandari (1988)	Total liabilities divided by fiscal year-end market capitalization
vim	Trading volume		Chordia et al. (2001)	Natural log of trading volume times price per share from month t - 2
sid_vim	Volatility of liquidity		Chordia et al. (2001)	Monthly standard deviation of daily trading volume
netdebt	Net debt		n.a.	Gross debt minus cash & equivalents at t-12
ev_eb	Enterprise value		n.a.	Enterprise value at t-1
cr_return	Revenue		n.a.	Revenue at t-5
beta	Market beta		n.a.	Estimated market beta from weekly returns and equal weighted market
l_rel_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
rel_v	Return volatility		Ang et al. (2006)	Standard deviation of daily returns from month t - 1
Momentum				
ch_mom	Change in returns		Geltman and Marks (2006)	Cumulative returns from months t - 6 to t - 1 minus months t - 12 to t - 7
mom1m	1 month momentum		Jegadeesh and Titman (1993)	1-month cumulative return
mom2m	2 month momentum		Jegadeesh and Titman (1993)	2-month cumulative returns ending one month before month end
mom3m	3 month momentum		Jegadeesh and Titman (1993)	3-month cumulative returns ending one month before month end
mom6m	6 month momentum		Jegadeesh and Titman (1993)	6-month cumulative returns from months t-36 to t-13
pr_delay	Price delay		Hour and Moskowitz (2005)	Cumulative returns from months t-60 to t-13
Investment and Issuance				
asset_gr	Asset Growth		Cooper et al. (2008)	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
Profitability				
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
gr_pft	Growth		Novy-Marx (2013)	Revenues minus cost of goods sold divided by lagged total assets
gr_sl	Sales growth		Lakonishok et al. (1994)	Annual percentage in sales at t-1
Macroeconomic factors				
household_indebt_pct_chg	Household indebtedness		Brazilian Central Bank (BCB)	mom change in the percentage at t-1
ile_br_chg	Economic uncertainty index		Fundacao Getulio Vargas (FGV)	mom percentage change in index at t-1
icg_non_sa_toda_chg	Non-seasonally adjusted total economic condition index		Fundacao Getulio Vargas (FGV)	mom percentage change in index at t-1
icg_non_sa_spec_chg	Non-seasonally adjusted expectations of economic conditions index		Fundacao Getulio Vargas (FGV)	mom percentage change in index at t-1
bcb_economic_activity_index_br_chg	Brazilian economic activity		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
services_revenue_index_chg	Services subsector revenue		Brazilian Institute of Geography and Statistics (IBGE)	mom percentage change in index at t-1
industry_production	Industrial production		Brazilian Institute of Geography and Statistics (IBGE)	mom percentage change in index at t-1
nominal_retail_sales_index_chg	Retail sales		Brazilian Institute of Geography and Statistics (IBGE)	mom percentage change in index at t-1
consumer_confidence_index_chg	Consumer confidence index		Fundacao Getulio Vargas (FGV)	mom percentage change in index at t-1
retailer_confidence_index_chg	Retailers confidence		Fundacao Getulio Vargas (FGV)	mom percentage change in index at t-1
expectations_of_retailers_confidence	Expectations of retailers confidence		Fundacao Getulio Vargas (FGV)	mom percentage change in index at t-1
global_pmi_composite_index	Global PMI composite index		Bloomberg	mom percentage change in index at t-1
brazil_pmi_composite_index	Brazil PMI composite index		Bloomberg	mom percentage change in index at t-1
embi_brazil_index_chg	Emerging Market Bond Index		JP Morgan	mom percentage change in index at t-1
outstanding_credit_financial_system_bcb_br_bn_chg	Outstanding credit in the financial system		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
credit_pct_gdp_br_chg	Total credit relative to GDP		Brazilian Central Bank (BCB)	mom change in the percentage at t-1
free_credit_pct_gdp_br_chg	Total credit relative to GDP		Brazilian Central Bank (BCB)	mom change in the percentage at t-1
sterilized_credit_pct_gdp_br_chg	Non-sterilized credit relative to GDP		Brazilian Central Bank (BCB)	mom change in the percentage at t-1
ipca_avg	Earmarked credit relative to GDP		Brazilian Central Bank (BCB)	mom change in the percentage at t-1
five_inflation_core_avg_chg	Total consumer inflation index		Brazilian Institute of Geography and Statistics (IBGE)	mom change in the percentage at t-1
bcb_brazil_commodities_composite_index_chg	Inflation, avg of five core measures		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
bcb_brazil_commodities_agro_index_chg	Brazil commodities composite index		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
bcb_brazil_commodities_metal_index_chg	Brazil commodities agricultural index		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
bcb_brazil_commodities_energy_index_chg	Brazil commodities metals index		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
bcb_brazil_trade_balance	Brazil trade balance		Brazilian Central Bank (BCB)	mom percentage change in index at t-1
unemp_rate_chg	Unemployment rate		Brazilian Central Bank (BCB)	mom change in the percentage at t-1

Figure 3 – Description of Brazil equity market risk and macroeconomic factors

	mean	std	min	25%	50%	75%	max
Revenue	0.0042	0.0283	-0.0601	-0.0159	0.0038	0.0246	0.0730
asset_gr	-0.0016	0.0232	-0.0691	-0.0153	0.0009	0.0141	0.0460
beta	-0.0035	0.0542	-0.1115	-0.0446	-0.0061	0.0293	0.1615
btm	0.0044	0.0289	-0.0683	-0.0145	0.0052	0.0236	0.0895
ch_mom	-0.0030	0.0249	-0.0684	-0.0178	-0.0041	0.0108	0.0624
eam_pr	0.0076	0.0301	-0.0674	-0.0109	0.0088	0.0265	0.0799
eam_rev	0.0081	0.0342	-0.0701	-0.0105	0.0066	0.0304	0.0861
ebit_reve	0.0083	0.0307	-0.0573	-0.0099	0.0074	0.0267	0.0860
ev	0.0030	0.0330	-0.0835	-0.0194	0.0039	0.0240	0.0875
ev_eb	0.0004	0.0280	-0.0660	-0.0185	0.0010	0.0177	0.0681
gr_sl	0.0025	0.0237	-0.0635	-0.0123	0.0049	0.0179	0.0537
gt_pft	0.0025	0.0290	-0.0940	-0.0144	0.0006	0.0213	0.0807
i_ret_v	0.0076	0.0421	-0.1177	-0.0209	0.0070	0.0357	0.1223
ill	0.0020	0.0281	-0.0734	-0.0148	0.0027	0.0245	0.0614
lev2	0.0080	0.0368	-0.1036	-0.0136	0.0077	0.0321	0.0996
levg	0.0007	0.0283	-0.0663	-0.0172	0.0003	0.0187	0.0687
mom12m	0.0132	0.0400	-0.1119	-0.0091	0.0141	0.0367	0.1089
mom1m	0.0073	0.0275	-0.0602	-0.0116	0.0091	0.0227	0.0620
mom36m	0.0034	0.0315	-0.0705	-0.0157	0.0070	0.0245	0.0757
mom60m	0.0016	0.0343	-0.1057	-0.0172	0.0053	0.0237	0.0784
mom6m	0.0083	0.0371	-0.1423	-0.0095	0.0098	0.0307	0.0895
netdebt	0.0053	0.0314	-0.0732	-0.0152	0.0046	0.0247	0.0788
op_pft	0.0048	0.0337	-0.0905	-0.0156	0.0071	0.0253	0.0957
pr_delay	0.0015	0.0190	-0.0490	-0.0081	-0.0007	0.0139	0.0478
ret_v	0.0048	0.0416	-0.1080	-0.0203	0.0053	0.0340	0.0914
sl_pr	0.0017	0.0256	-0.0644	-0.0166	-0.0001	0.0199	0.0684
std_vlm	-0.0010	0.0245	-0.0642	-0.0153	-0.0002	0.0171	0.0438
sz	0.0032	0.0351	-0.0949	-0.0180	0.0049	0.0258	0.0966
vlm	0.0005	0.0261	-0.0743	-0.0162	0.0015	0.0189	0.0494
household_indebt_pct_chg	0.0772	0.3207	-0.7000	-0.1000	0.0500	0.2000	1.0300
ie_br_chg	0.1254	7.7083	-16.7000	-4.1000	-0.4000	3.3000	43.4000
ice_non_sa_total_chg	0.2355	2.6399	-8.7000	-1.0000	0.4000	1.7000	8.9000
ice_non_sa_expec_chg	0.0903	3.8109	-13.1000	-2.4000	0.6000	2.2000	9.4000
bcbr_economic_activity_index_br_chg	0.0027	0.0362	-0.0626	-0.0189	-0.0022	0.0197	0.1164
industry_production_index_chg	0.0025	0.0660	-0.1380	-0.0433	0.0028	0.0488	0.1492
nominal_retail_sales_index_chg	0.0131	0.0951	-0.2779	-0.0280	0.0170	0.0613	0.3056
consumer_confidence_index_chg	0.0018	0.0394	-0.1286	-0.0221	0.0061	0.0252	0.0923
consumer_confidence_expectations_index_chg	0.0033	0.0485	-0.1474	-0.0214	0.0040	0.0311	0.1289
retailer_confidence_index_chg	0.0020	0.0509	-0.2138	-0.0229	0.0031	0.0304	0.1414
retailer_confidence_expectations_index_chg	-0.0012	0.0679	-0.2340	-0.0395	0.0055	0.0397	0.1767
EMBI_brazil_index_chg	-0.0018	0.1012	-0.2141	-0.0605	-0.0083	0.0490	0.3489
effec_us_fed_funds_avg_chg	0.0567	0.1747	-0.6000	-0.0021	0.0048	0.0891	0.7000
outstanding_credit_financial_system_bcb_brl_bn_chg	0.0083	0.0074	-0.0081	0.0027	0.0082	0.0142	0.0228
credit_pct_gdp_br_chg	0.0758	0.3517	-0.7566	-0.1743	0.0521	0.3243	0.9108
free_credit_pct_gdp_br_chg	0.0427	0.2387	-0.4480	-0.1241	0.0438	0.1991	0.6563
stereed_credit_pct_gdp_br_chg	0.0312	0.1835	-0.3201	-0.0877	0.0079	0.1503	0.5378
IPCA_chg	0.0068	0.3521	-1.3500	-0.1500	0.0300	0.2100	0.8800
five_inflation_core_avg_chg	-0.0008	0.1517	-0.3653	-0.0889	-0.0031	0.0966	0.3393
bcbr_brazil_commodities_composite_index_chg	0.0067	0.0354	-0.0585	-0.0176	0.0044	0.0308	0.1069
bcbr_brazil_commodities_agro_index_chg	0.0068	0.0349	-0.0697	-0.0207	0.0048	0.0286	0.0952
bcbr_brazil_commodities_metals_index_chg	0.0044	0.0443	-0.0979	-0.0203	0.0032	0.0305	0.1188
bcbr_brazil_commodities_energy_index_chg	0.0080	0.0718	-0.1939	-0.0286	0.0066	0.0542	0.2312
bcbr_brazil_trade_balance_chg	0.0098	0.9959	-4.2929	-0.3013	-0.0437	0.3277	4.9712
lead_log_er_monthly_return	-0.0058	0.1227	-0.4163	-0.0686	0.0014	0.0676	0.3013

Figure 4 – Description of Brazil equity market risk and macroeconomic factors

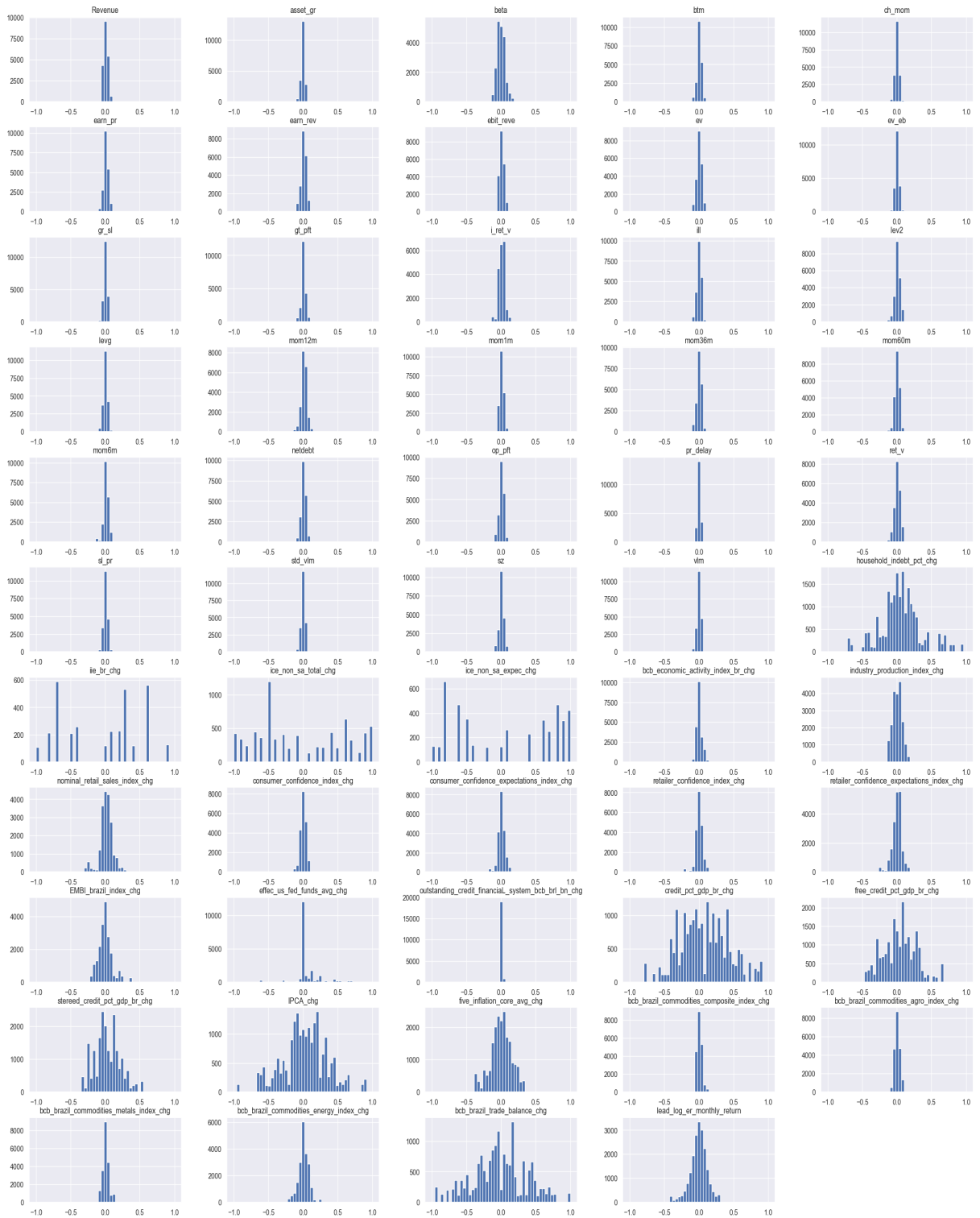


Figure 5 – Histogram of firm characteristics and factors

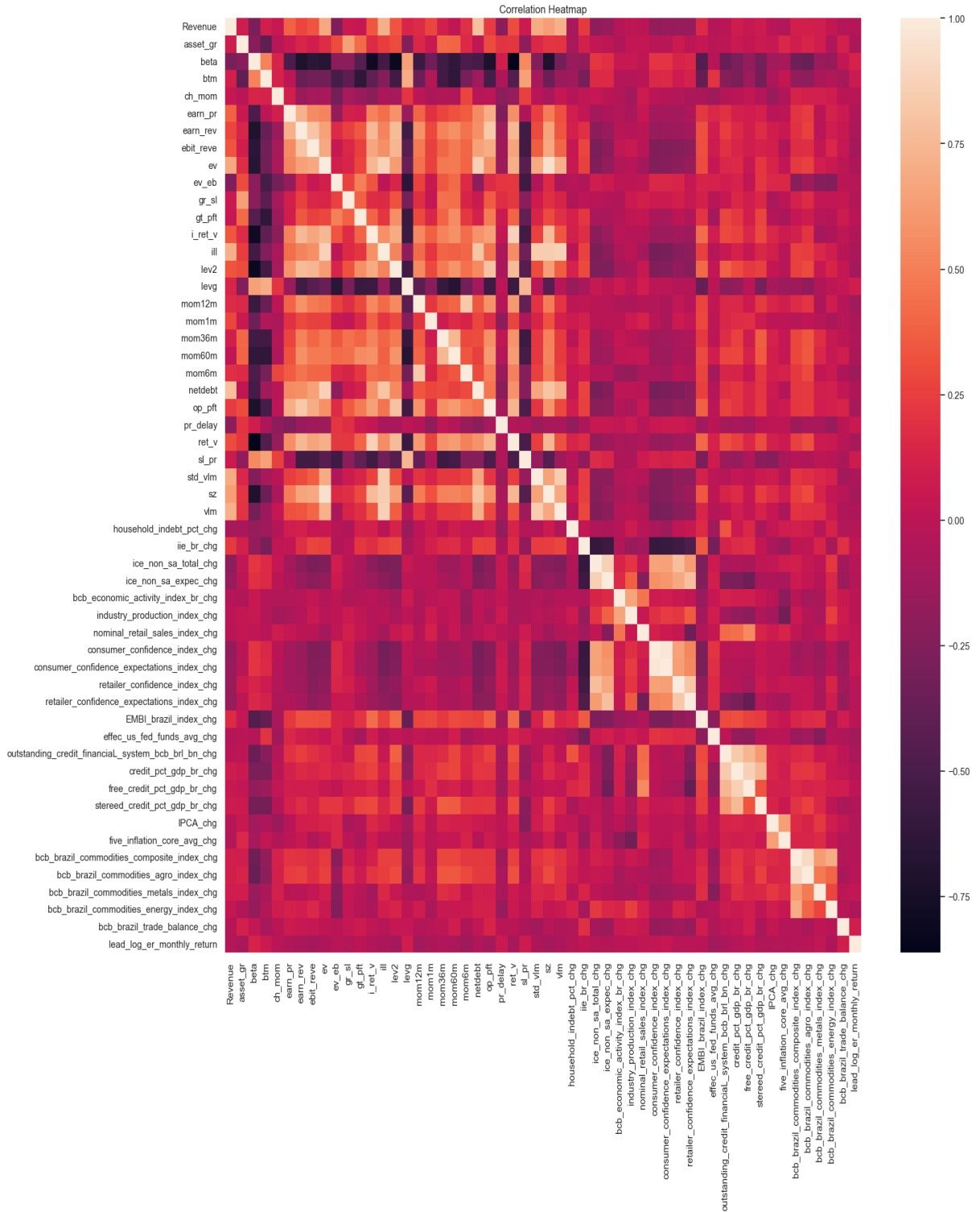


Figure 7 – Factors correlation matrix

3.1.4 Hyperparameter tuning

We closely follow [Gu, Kelly e Xiu \(2020\)](#) on defining the range for hyperparameter tuning used in our models. We slightly differ in GBRT's number of trees and maximum depth for Random forest due to computational limitations.

Model	OLS-3	OLS	PLS	PCR	ENet	GBRT	RF	NN1 - NN5
Huber loss ξ	✓	✓				✓		
Others			max components = 20	max components = 10	$\rho = 0.5$	Depth=1 ~ 4 #Trees=10	Depth=1 ~ 4 #Features in each split $\in \{1, 2, 3, 4, 5, 10, 20, 30, 40 \dots\}$ #Trees=100	L1 penalty Batch size = 1000
Learning Rate					$\lambda \in (10^{-5}, 10^{-3})$	LR $\in \{0.01, 0.1\}$		LR $\in \{0.01, 0.1\}$
Epochs								100
Adam Para.								Default

Table 1 – Hyperparameters tuning

3.1.5 Data Sampling

We used a rolling sampling scheme, similarly to [Gu, Kelly e Xiu \(2020\)](#). In this method, the training and validation sets shift forward in time with each iteration, ensuring that the total number of time periods in each set remains constant. For instance, we used 60 months of data for training, 24 for valuation and 12 for testing. For each rolling window, the model is re-fitted using the current training and validation sets, and its performance is evaluated on the test data not yet included in the rolling windows. This process generates a sequence of performance metrics for each rolling window. The key advantage of this approach is that it incorporates more recent data for predictions, making it more adaptive compared to the fixed split scheme, in which the model is estimated only once from the training and validation samples, attempting to fit all points in the testing sample.

For robustness checks, we also created two additional data sets containing large and small stocks, using a top/bottom 30% market capitalization rule.

3.2 Results

3.2.1 The cross-section of individual stocks

We compare twelve models in total, namely OLS-3 with Huber Loss that contains size, book-to-market, and momentum as the only predictors, OLS with Huber containing all factors, Partial Least Squares (PLS), Principal Component Analysis (PCR), Elastic net (ENet), Gradient Boosted Regression Trees (GBRT), Random Forest (RF), and Feed-forward Neural Networks from one to five hidden layers.

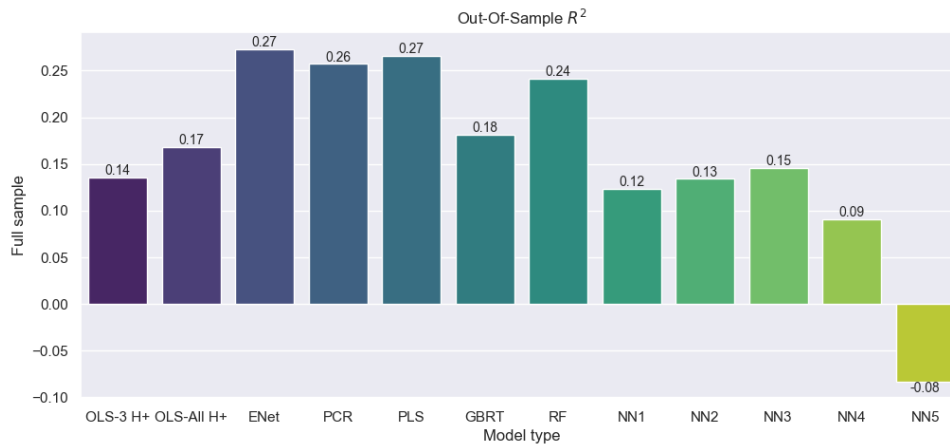
Figures [8a](#), [8b](#) and [8c](#) and [Table 2](#) we show out-of-sample R^2 for both the full sample of stocks and for the subset of large and small ones, as defined previously. It is notable the predictive power gains obtained from most of machine learning models (+0.24% in R^2_{OS} on average for ENet, PCR, PLS, GBRT and RF vs +0.17% for traditional OLS

with all factors Huber Loss, OLS-All H+, while feed-forward Neural Networks performed worse), both in the full sample and in the large and small stocks sub-samples, added as robustness checks.

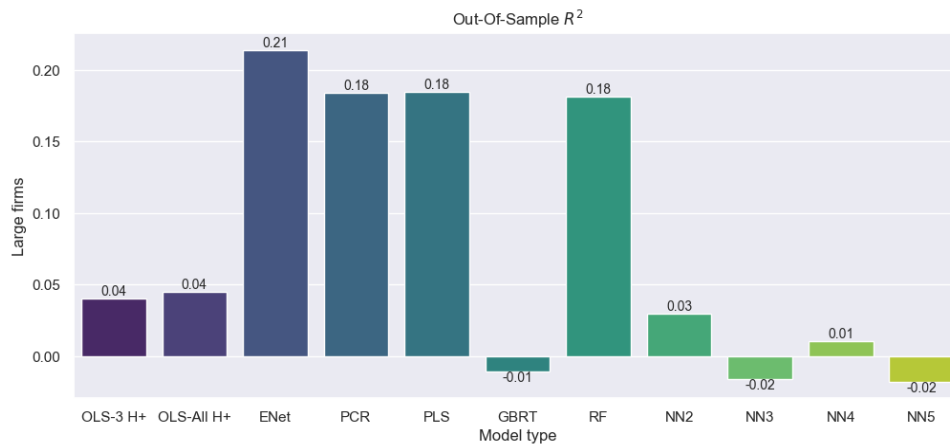
ENet and PLS performed best across all sub-samples with R_{OOS}^2 of 0.27% each, but PCR and Random Forest performed similarly with R_{OOS}^2 of 0.26% and 0.24%, respectively. We can see that the usage of parameter shrinkage and variable selection $l1_{ratio} = 0.5$ to limit the regression's degrees of freedom, greatly improved OLS-H+ results. Nonetheless, differently from Gu, Kelly e Xiu (2020), feed-forward neural network models did not show gains over OLS H+, but we reach similar findings on the higher predictive power of more shallow neural networks compared to deeper ones (+0.15% in the 3-hidden layer version being the best performing one compared to 0.09% and -0.08% for the four-and-five hidden layer versions). The results seem robust as within large and small stocks, as obtained R_{OOS}^2 also performed better with ML models, in the exception of Gradient Boosted Regression Trees. Finally, we found that R_{OOS}^2 were higher for small stocks compared both the full sample and large stocks, despite their theoretically lower signal-to-noise ratio.

	Full sample	Large firms	Small firms
OLS-3 H+	0.1353	0.0405	0.2302
OLS-All H+	0.1678	0.0449	0.2274
ENet	0.2732	0.2140	0.3403
PCR	0.2572	0.1839	0.3054
PLS	0.2656	0.1848	0.3202
GBRT	0.1810	-0.0103	0.1502
RF	0.2418	0.1814	0.2854
NN1	0.1230	-6.8824	-1.7040
NN2	0.1339	0.0298	-0.1058
NN3	0.1457	-0.0156	0.0400
NN4	0.0908	0.0108	0.0458
NN5	-0.0834	-0.0177	0.0494

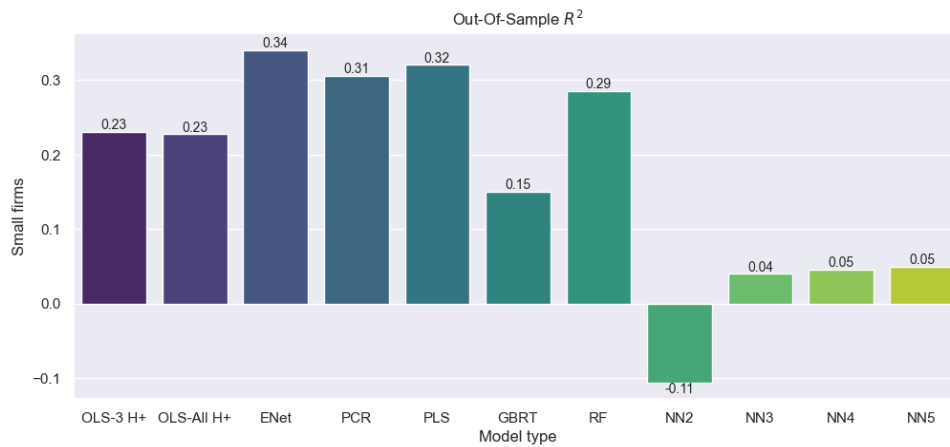
Table 2 – R_{OOS}^2 performance metrics for different models.



(a) Out-of-Sample R^2 for full sample of stocks.

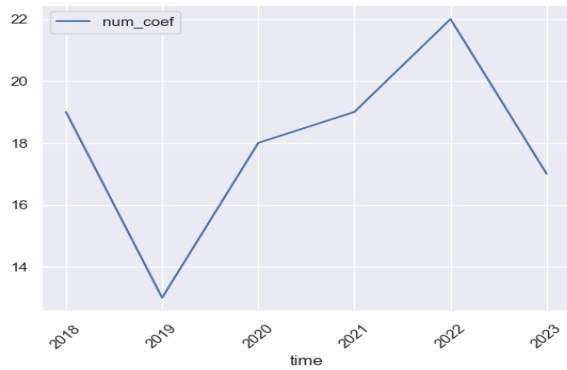


(b) Out-of-Sample R^2 for large stocks.

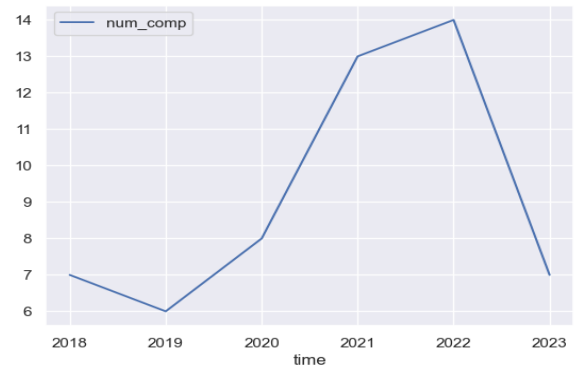


(c) Out-of-Sample R^2 for small stocks.

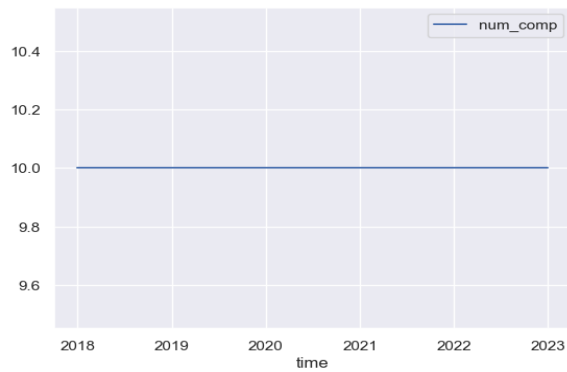
Figure 8 – Out-of-Sample R^2 for different subsets of stocks.



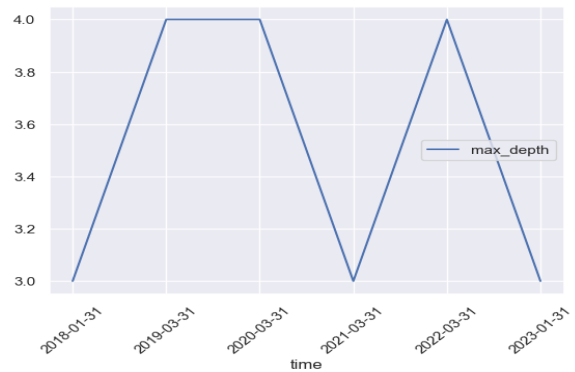
(a) ENet's time-varying model complexity.



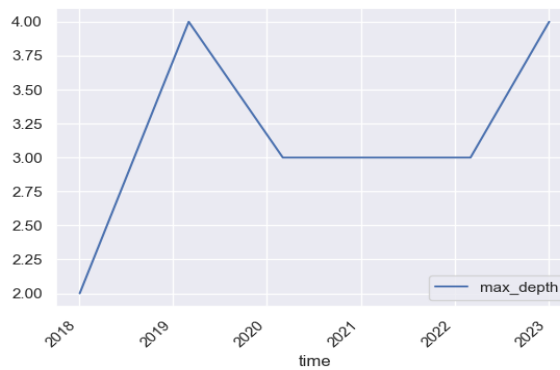
(b) PLS's time-varying model complexity.



(c) PCR's time-varying model complexity.



(d) Gradient Boosted Regression Trees' time-varying model complexity.



(e) Random Forest's time-varying model complexity.

Figure 9 – Time-varying model complexity for various methods. For ENet we report the number of features selected to have nonzero coefficients; for PCR and PLS, we report the number of selected components; for RF, we report the average tree depth; and, for GBRT, we report the number of distinct characteristics entering into the trees. ENet, PLS and PCR used substantially more non-zero components (i.e. minimum 6 and maximum of 22) than did GBRT and RF, which topped at four.

3.2.2 The relevance of covariates

We investigate the relative importance of individual covariates in contributing to the performance of each model, applying the importance measures outlined in the data treatment section. Following Gu, Kelly e Xiu (2020) we calculate the reduction in R_{OOS}^2 results for each ML method from setting all values of a specific predictor to zero within each training sample.

These reductions are then averaged to obtain a single importance measure for each predictor. Figure 10-16 display the resulting importance scores for all factors utilized in our models. To ensure comparability, the variable importance scores within each model are normalized to sum to one, enabling the assessment of relative importance within each model.

We found macroeconomics factors overweight firm-related ones, with a predominance of country risk (EMBI Brazil Index), in ENet and PLS models, followed by the expectations of macroeconomic conditions in PCR model, Brazil's commodities composite index in Random Forest model, and credit-to-GDP ratio in neural network . Among Brazil equity risk factors, beta and price trends ranked the most relevant for the best performing methods. Our findings suggest a high relevance of macroeconomic factors when predicting monthly stock returns for Brazilian stocks.

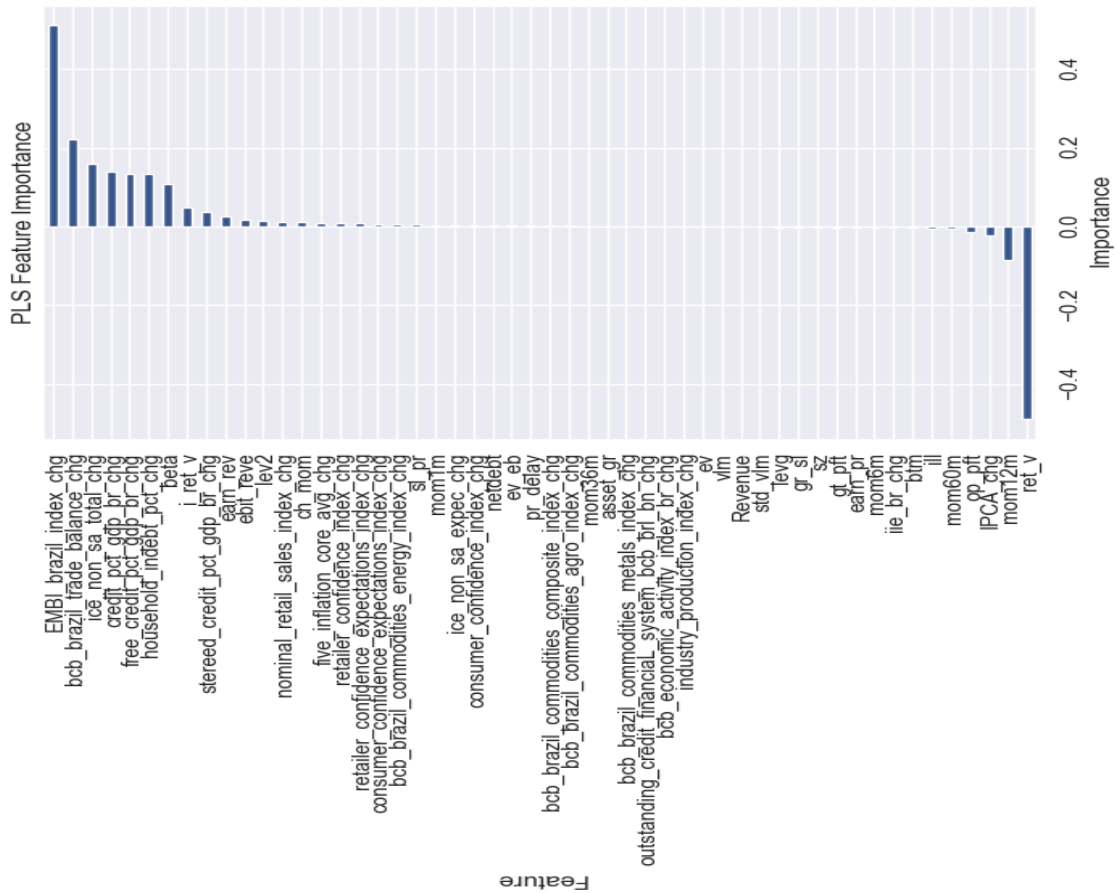


Figure 11 – PLS OOS Variable Importance

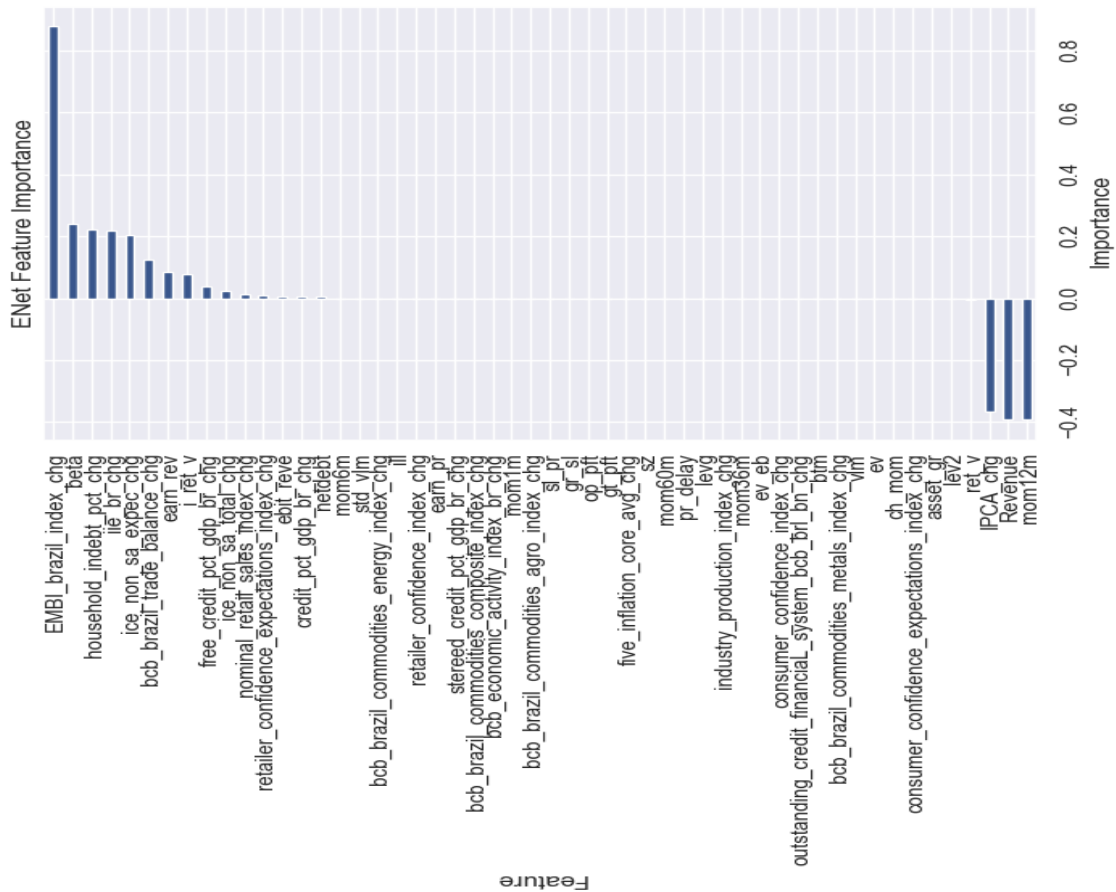


Figure 10 – ElasticNet OOS Variable Importance

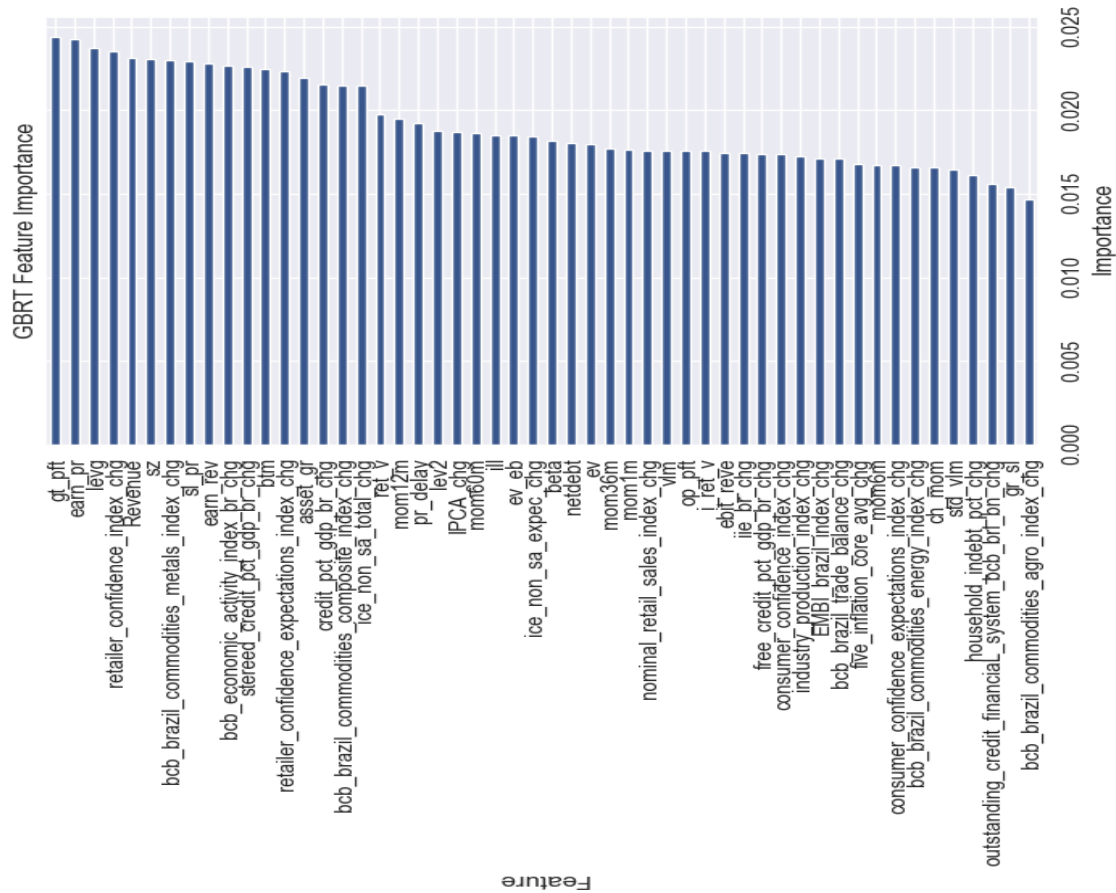


Figure 13 – Gradient Boosted Regression Tree OOS Variable Importance

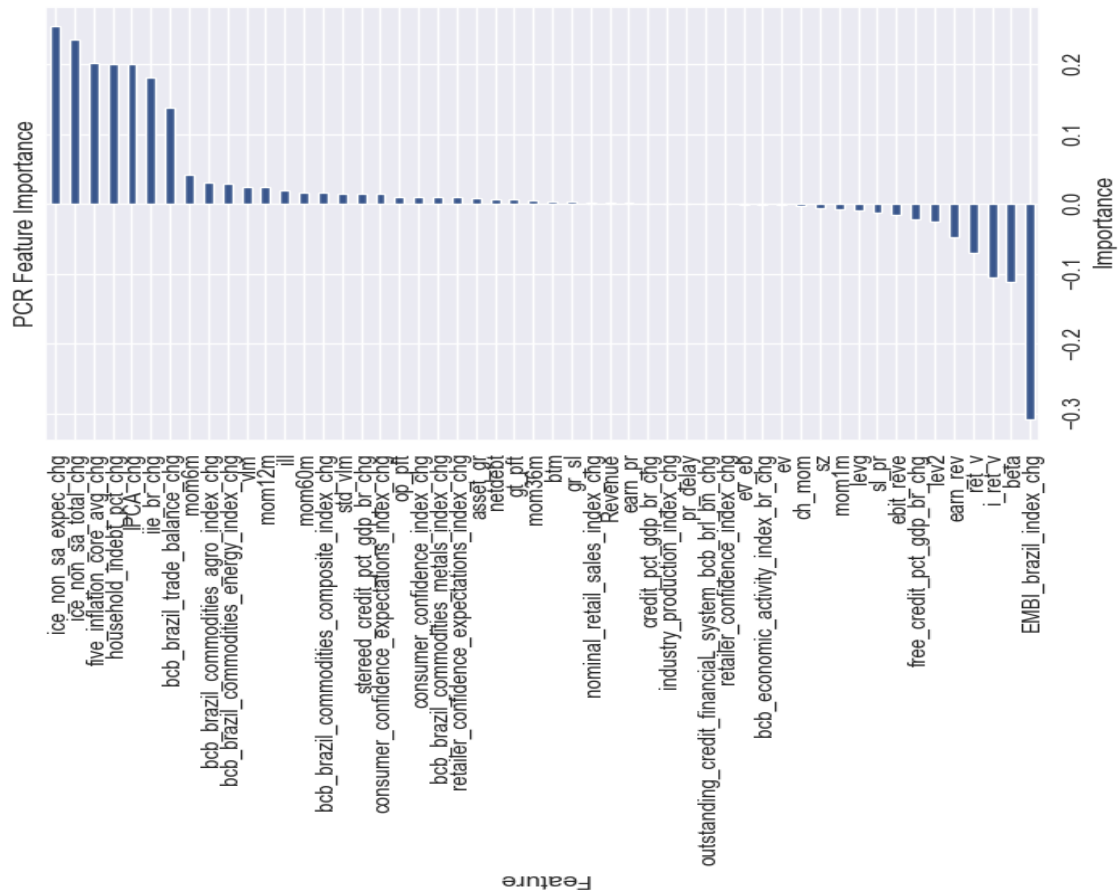


Figure 12 – PCR OOS Variable Importance

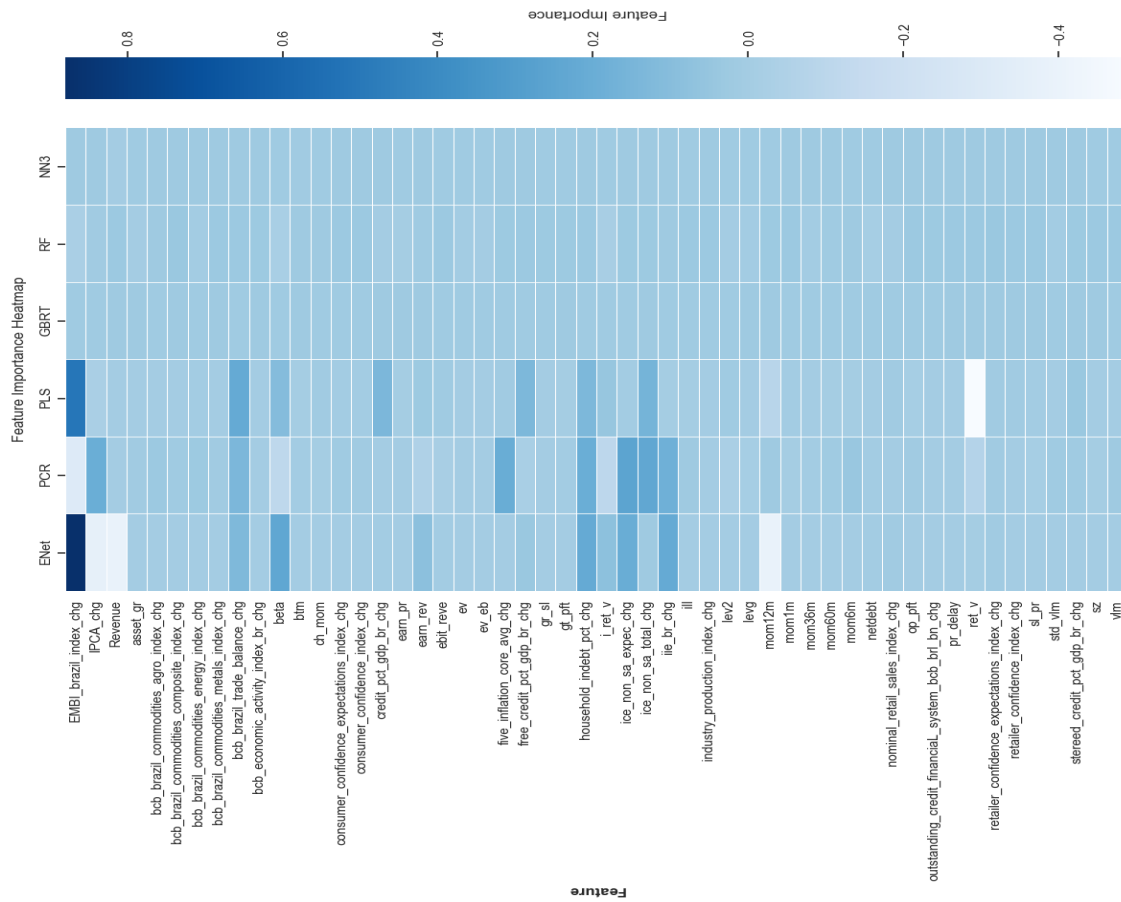


Figure 16 – All Models OOS Variable Importance

4 Conclusion

Backed by empirical asset pricing approach on cross-section stock return prediction, we have conducted a comparative analysis of methods in Machine Learning for Brazilian stocks. We've run twelve models, including Brazil equity market risk factors and innovating through the addition of a vast set of Brazilian macroeconomics factors. We confirm literature results of gains in predictive power over traditional OLS models as measure by Out-of-Sample R^2 for the vast majority of analyzed models, markedly variable selection and shrinkage methods, such as Elastic Net and Partial Least Squares, which while Feed-Forward Neural Networks underperformed, although we could see shallow learning outperformed deeper learning as also seen in the literature. We have also found that the employed methods delivered higher performance for small stocks, despite the theoretically higher signal-to-noise ratio. Finally, running an out-of-sample variable importance analysis, we also found macroeconomic factors overweight firm-related ones, with a slight predominance of country risk (EMBI Brazil Index), followed by the expectations of economics conditions, Brazil's commodities composite index, and credit-to-GDP ratio. Among Brazil equity risk factors, beta and price trends ranked the most relevant for the best performing methods. Our findings suggest a high relevance of macroeconomic factors when predicting monthly stock returns for Brazilian stocks.

References

- ANG, A.; HODRICK, R. J.; XING, Y.; ZHANG, X. The Cross-Section of Volatility and Expected Returns. *Journal of Finance*, v. 61, n. 1, p. 259–299, 2006.
- ARISMENDI-ZAMBRANO, J.; GENARO, A.; ALEXANDRE, H. L. Intraday returns forecasting using machine learning: Evidence from the brazilian stock market. *SSRN*, 2023.
- ASNESS, C. S.; MOSKOWITZ, T. J.; PEDERSEN, L. H. . Value and momentum everywhere? *Journal of Finance*, v. 638, n. 3, p. 929–985, 2013.
- BACK, K. E. *Asset Pricing and Portfolio Choice Theory*. US: Oxford University Press, 2017.
- BAGNARA, M. Asset pricing and machine learning: A critical review. *Journal of Economic Surveys*, v. 38, n. 1, p. 27–56, 2024.
- BLUME, M. E. On the assessment of risk. *Journal of Finance*, v. 26, n. 1, p. 1–10, 1971.
- BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.
- CAROSIA, A. E. de O.; SILVA, A. E. da; COELHO, G. P. Predicting the brazilian stock market with sentiment analysis, technical indicators and stock prices: A deep learning approach. *Comput Econ*, 2024.
- CEFIM. *Brazil Stock Market Risk Factors*. 2024. <https://cefim-insper.github.io/projects/GHZ_output.html>.
- CHEN; ROLL, N.-F.; R.; ROSS, S. A. . Economic forces and the stock market. *Journal of Business*, v. 59, n. 3, p. 383–403, 1986.
- COCHRANE, J. H. *Asset Pricing. Revised*. NJ, US: Princeton University Press, 2005.
- E.SCHAPIRE, R. The strength of weak learnability. *Machine Learning*, v. 5, n. 2, p. 197–227, 1990.
- FAMA, E. F.; FRENCH, K. R. The cross-section of expected stock returns. *Journal of Finance*, v. 47, n. 2, p. 427–465, 1992.
- FAMA, E. F.; FRENCH, K. R. Five-factor asset pricing model. *Journal of Financial Economics*, v. 116, n. 1, p. 1–22, 2015.
- FERREIRA, F. G. D. C.; GANDOMI, A. H.; CARDOSO, R. T. N. Financial time-series analysis of brazilian stock market using machine learning. p. 2853–2860, 2020.
- FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001.
- GU, S.; KELLY, B.; XIU, D. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, v. 33, n. 5, p. 2223–2273, 2020.

- HARVEY, C. R.; LIU, Y.; ZHU, H. Editor's Choice ... and the Cross-Section of Expected Returns. *The Review of Financial Studies*, v. 29, n. 1, p. 5–68, 2016.
- HASTIE ROBERT TIBSHIRANI, J. F. T. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. [S.l.]: Springer Science & Business Media, 2013.
- HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 12, n. 1, p. 55–67, 1970.
- HONG, H.; LIM, T.; STEIN, J. C. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of Finance*, v. 55, n. 1, p. 265–295, 2000.
- HOU, K.; XUE, C.; ZHANG, L. Editor's Choice Digesting Anomalies: An Investment Approach. *The Review of Financial Studies*, v. 28, n. 3, p. 650–705, 2015.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; TAYLOR, J. *An Introduction to Statistical Learning with Applications in Python*. 1. ed. [S.l.]: Springer Cham, 2023.
- KELLY, B.; PRUITT, S. Market expectations in the cross-section of present values. *Journal of Finance*, v. 68, p. 1721–1756, 2013.
- KELLY, B.; PRUITT, S. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, v. 186, p. 294–316, 2015.
- KELLY, B. T.; XIU, D. *Financial Machine Learning*. [S.l.], 2023.
- LINTNER, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, v. 19, n. 3, p. 425–442, 1965.
- MARKOWITZ, H. Portfolio selection. *Journal of Finance*, v. 7, n. 1, p. 77–91, 1952.
- MEHRA, R.; PRESCOTT, E. C. The equity premium: A puzzle. *Journal of Monetary Economics*, v. 15, n. 2, p. 145–161, 1985.
- MOSSIN, J. Equilibrium in a capital asset market. *Econometrica*, v. 34, n. 4, p. 768–783, 1966.
- N., J.; TITMAN, S. . Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, v. 48, n. 1, p. 65–91, 2015.
- NOVY-MARX, R. The other side of value: The gross profitability premium. *Journal of financial economics?* *The Accounting Review*, v. 108, n. 1, p. 1–28, 2013.
- RIBEIRO, R.; COSTA, J.; KAEBI, M. M.; MARTINS, I.; NOBREGA, T. Firm characteristics and stock returns in brazil. *SSRN*, 2024.
- ROBITAILLE, P. T.; ROUSH, J. E. *How Do FOMC Actions and U.S. Macroeconomic Data Announcements Move Brazilian Sovereign Yield Spreads and Stock Prices?* [S.l.], 2006. Available at SSRN: <<https://ssrn.com/abstract=946776>>. Disponível em: <<https://ssrn.com/abstract=946776>>.
- ROSS, S. A. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, v. 13, n. 3, p. 341–360, 1976.

- SHANKEN, J. The arbitrage pricing theory: Is it testable? *Journal of Finance*, v. 37, n. 5, p. 1129–1140, 1982.
- SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, v. 19, n. 3, p. 425–442, 1964.
- SILVA, J. E.; SARTIRO, R. M. The prediction ability of machine learning boosting models in the brazilian stock market. *Brazilian Journal of Quantitative Methods Applied to Accounting*, 2024.
- SLOAN, R. G. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, v. 71, n. 1, p. 289–315, 1996.
- TABAK, B. M.; LAIZ, M. C.; CAJUEIRO, D. O. The impact of the global financial crisis on the brazilian stock market. In: *Emerging Market Economies and Financial Globalization*. Springer, 2017. cap. 11. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-64885-9_11>.
- TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996.
- WONGSWAN, J. The reaction of international stock markets to federal reserve policy. *Springer Journal of Financial Markets and Policies*, 2009. Disponível em: <<https://link.springer.com/article/10.1007/s11408-012-0204-3>>.
- ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 67, n. 2, p. 301–320, 2015.