UCLouvain

NOVA SCHOOL OF BUSINESS & ECONOMICS

LOUVAIN School of Management

**Louvain School of Management**
**and Nova School of Business & Economics**

# Using Machine Learning tool in Fund Selection : Review and Empirical test

Project/Research Master's Thesis submitted by
Anouk Reinhardt

With a view of getting the degrees:
Master's in international finance, Financial Market Expertise
Master's in business engineering, Professional Focus, Major in Financial Engineering

Supervisor :
Anh Nguyen

Academic Year [*2020-2021*]

## Abstract

Since the Finance Industry is, through the years, growing tremendously, the willingness to understand, predict and trade on financial markets has led to an urgent demand for developing innovative models. Hence, Machine Learning Methods have gained in popularity. Yet, considered underperforming compared to traditional statistical approaches, the real-life application of machine learning on financial decision-making is currently lagging behind its promise. Indeed, poor performance was demonstrated when dealing with a large amount of data with intrinsic complexity and high dimensionality. This thesis aims to scrutinise opinions on machine learning, especially on its potential of being a tool for selecting, predicting and optimising portfolios. The study will evaluate the machine learning method by selecting and forecasting United States open-ended mutual funds from June 1990 to January 2021. This evaluation will be made on the seven following Machine Learning Methods: OLS regression, Lasso, Ridge, Elastic Net, XG Boost, Deep Neural Network and Random Forest. To select the open-ended mutual funds, the characteristics used as predictors are the following variables: monthly realised alpha, alpha t-stat, total net asset, flows, value-added, R2, and the t-stats of the market, profitability, investment, size, value, and momentum betas.

The results obtained were regarding cumulative returns that Random Forest gains the first position with 37.21%, followed by XG Boost with 30.76%, and closely by Deep Neural Network with 29.43%. Hence, the results demonstrated the efficiency and robustness of machine learning methods in the selection strategy of mutual funds. Thus, the findings support that those machine learning methods will find their way into finance due to their reliability and good performance. Yet, the results should also be nuanced as the high turnover and implying high transactions cost could adversely affect the investment strategy's performance.

# Table of Contents

LIST OF FIGURES ..... V

LIST OF TABLES ..... VI

1.INTRODUCTION ..... 1

1.1 CONTEXT AND MOTIVATION ..... 1
1.2 STRUCTURE AND SCOPE ..... 3

2. PREVIEW OF EMPIRICAL FINDINGS AND CONTRIBUTIONS ..... 4

2.1 PREDICTION OF FUTURE MUTUAL FUNDS RETURNS WITH PAST PERFORMANCE, HISTORICAL INFORMATION AND ALPHA ..... 4
2.2 RELATIONSHIP OF MUTUAL FUND CHARACTERISTICS AND PERFORMANCE ..... 6
*2.2.1 Effect of Mutual fund size on performance ..... 6*
*2.2.2 Effect of Mutual Fund Age on performance ..... 9*
*2.2.3 Effect of Mutual Funds Flows on performance ..... 10*
*2.2.4 Effect of the Portfolio turnover on performance ..... 12*
*2.2.5 Effect of the Expense ratio on performance ..... 13*
2.3 ESTIMATION OF FUND PERFORMANCE WITH MACROECONOMIC DERIVATIVE-FACTORS ..... 16

3. METHODOLOGY ..... 18

3.1 EMPIRICAL RESEARCH ON THE APPLICATION OF MACHINE LEARNING TECHNIQUES ..... 19
3.2 MACHINE LEARNING METHODS ..... 21
*3.2.1 Ordinary Least Square Regression (OLS) ..... 21*
*3.2.2 Shrinkage-Based Methods ..... 22*
3.2.2.1 Lasso (Least Absolute Shrinkage and Selection Operator) ..... 23
3.2.2.2 Ridge Regression ..... 24
3.2.2.3 Elastic Net ..... 25
*3.2.3 Tree-Based Method ..... 25*
3.2.3.1 Random Forest ..... 26
3.2.3.2 XG Boost – Extreme Gradient Boosting ..... 27
*3.2.4 Deep Neural Network ..... 28*

4. DATA AND PREDICTIVE FEATURES ..... 32

4.1 DATA DESCRIPTION AND PREPARATION ..... 32
4.2 MODEL BUILDING AND MUTUAL FUNDS SELECTION STRATEGY ..... 36

5. EMPIRICAL RESULTS ..... 43

5.1 ANALYSE OF MACHINE LEARNING METHODS ON MUTUAL FUND RETURNS ..... 43
5.2 ANALYSE OF MACHINE LEARNING METHODS ON MUTUAL FUND SELECTION STRATEGY ..... 49
5.3 ANALYSE OF MACHINE LEARNING METHODS ON THE FIVE-FACTOR MODEL OF FAMA AND FRENCH ..... 55
5.4 ANALYSE OF MACHINE LEARNING METHODS ON ACCURACY ..... 58

6. DISCUSSION ..... 61

7. CONCLUSION ..... 66

8. BIBLIOGRAPHY ..... 68

9. APPENDIX ..... 79

## List of Figures

*Figure 1: Machine learning problems* ____ *18*
*Figure 2: Components of a decision tree* ____ *26*
*Figure 3: Architecture of a single neuron in a neural network* ____ *29*
*Figure 4: Neural network architecture* ____ *30*
*Figure 5: Illustration of Deep Neural Network architecture with multiple layers* ____ *31*
*Figure 6: 5-fold Cross-Validation* ____ *38*
*Figure 7: Evolution of the XG Boost and actual mutual funds drawdown period.* ____ *46*
*Figure 8: Cumulative return of machine learning methods, S&P500 and actual mutual funds returns.* ____ *48*
*Figure 9: Comparison of selected mutual funds (Group 8) of XG boost with actual returns on the whole period (July 2018 to June 2021).* ____ *52*
*Figure 10: Comparison of selected mutual funds (Group 8) of Deep Neural Network with actual returns on the whole period (July 2018 to June 2021).* ____ *53*
*Figure 11: Cumulative returns of the best mutual funds (group 8) depending on the different machine learning methods from July 2018 to June 2021.* ____ *54*
*Figure 12: Summary of Boruta analysis on the importance of variables* ____ *62*
*Figure 13: Evolution of the actual monthly returns of the equal-weighted portfolio composed of 288 mutual funds with the S&P 500 index from July 2018 to June 2021.* ____ *64*
*Figure 14: Comparison of Shrinkage-based methods, OLS and Random Forest machine learning methods* ____ *65*

## List of Tables

*Table 1: Principal investment strategy of mutual funds from our sample* 33
*Table 3: Comparison of monthly returns predicted by Machine Learning methods with the actual returns of the mutual funds and the benchmark S&P 500 from July 2018 to June 2019* 43
*Table 4: Summary of the performance of the statistics of the machine learning methods, S&P500 and mutual funds actual returns* 44
*Table 5: Summary of the analysis of drawdown period from July 2018 to June 2021* 45
*Table 6: Summary of statistics performance of machine learning methods from July 2019 to June 2020.* 47
*Table 7: Summary of statistics performance of machine learning methods from July 2020 to June 2021.* 47
*Table 8: Summary of average annual return of the machine learning methods performance from July 2018 to June 2019* 50
*Table 9: Summary of results of Standard Deviation, Sharpe Ratio and positive months of the machine learning methods from July 2018 to June 2019* 50
*Table 10: Summary of average annual return of the machine learning methods performance from July 2019 to June 2020* 51
*Table 11: Summary of results of Standard Deviation, Sharpe Ratio and positive months of the machine learning methods from July 2019 to June 2020* 51
*Table 12: Summary of average annual return of the machine learning methods performance from July 2020 to June 2021* 51
*Table 13: Summary of results of Standard Deviation, Sharpe Ratio and positive months of the machine learning methods from July 2020 to June 2021* 51
*Table 14: Summary of raw alpha (intercept coefficient from the 36-month rolling window regression) of the best selected mutual funds (group 8) depending on the predicted year and machine learning technique* 56
*Table 15: Summary of raw betas (beta coefficient from the 36-month rolling window regression) of the best selected mutual funds (group 8) depending on the predicted year and machine learning techniques* 57
*Table 16: Summary of statistics error measures of the mutual fund FAGOX US EQUITY with the highest return on the whole period (July 2018 to June 2021)* 59
*Table 17: Summary of statistics error measures of the mutual fund FSESX US Equity with the lowest return on the whole period (July 2018 to June 2021)* 60
*Table 18: Summary of correlation between features* 63

# 1.Introduction

## 1.1 Context and motivation

In the last years, the Finance Industry has experienced fundamental changes. With the Big Data Era breakout and technology progress, settlement transactions can be done in real-time and borderless via T2S (Trans- European Automated Real-time Gross Settlement Express Transfer). T2S, a pan-European settlement platform implemented by the International Eurosystem Central Securities depository, can process on a daily basis 40,000 instructions in less than 20 seconds of receipt on average. Data and analytics, considered a significant driver of performance, are at the centre of attention, especially in the asset management industry. With the explosion in data availability, Big data represents a positive trend for active managers who embrace the opportunities presented by this development. Indeed, data and analytics is a tool that can sharpen their decision making in the selection of the portfolio.

Based on two popular approaches, the traditional and the quantitative, portfolio selection aims to optimise wealth allocation across a set of assets. Both methods share common characteristics, such as analysing historical data to estimate critical drivers or market trends, adopting deterministic criteria in the decision of asset selection, and evaluating performance over time. The traditional approach relies heavily on the judgement depth analysis, key characteristics and qualitative factors. Markowitz provides the foundation for portfolio selection with the Modern Portfolio theory or so-called mean-variance analysis. In contrast, the quantitative approach is based on quantitative investment analysis relying on mathematical models and concepts. Selecting a portfolio over a set of assets with specific targets is challenging, such as predicting the market's movement or a particular asset. Although forecasting has been considered one of the most challenging tasks in the financial market, it plays a vital role in selecting assets. As a result, developing an accurate forecasting model is challenging due to its complexity. Thus, a fundamental research problem in computational finance and financial engineering is the selection of the portfolio.

Interest in machine learning methods for finance has grown tremendously in academia and industry. Studies have been carried out to enhance prediction accuracy with statistical and machine learning algorithms, recognised as tools that provide full access and analysis of a large amount of

unstructured data. Although machine learning has been extensively studied, its application in the era of investments remains scarce. It is mainly for its controversial topic in quantitative finance, notably whether machine learning can be a practical investment tool.

This thesis aims to scrutinise opinions on machine learning, especially on its potential of being a tool for selecting, predicting and optimising the portfolio. This thesis will focus on United States open-end mutual funds with the asset-class focus of equity to achieve this goal.

Over the past decade, mutual funds have increased drastically in popularity. Mutual funds are defined as various collective investment vehicles, managed professionally, that pool money from individuals or professional investors to purchase securities such as bonds, stocks, and derivatives. Hence, worldwide open-end mutual funds regulated assets sharply rise from $32.25 trillion in 2012 to $54.9 trillion in 2019 (ICI, 2020). Indeed, mutual funds offer numerous advantages as opposed to direct investment in individual securities. The fund's nature provides a high level of diversification, ease of comparison to investors, and the accessibility of participation in investments unachievable for small investors. Other advantages are associated with mutual funds, such as higher liquidity, stability in price, economies of scale, know-how, and skilled fund managers' experiences.

There are two types of mutual funds, open-end mutual funds and closed-end funds. An open-end fund can issue an unlimited volume of shares and redeem existing shares on demand allowing the investor to buy shares at the chosen quantity and at any time. One hallmark of this mutual fund type is that the shares trade at their current net asset value (NAV). The NAV is calculated by dividing the fund's assets' total market value, minus its liabilities, by the number of mutual fund shares outstanding at the closing market. In contrast, closed-end funds, also known as publicly-traded funds or closed-end investment trusts, issue a fixed number of shares that trade intraday on a stock exchange and does not issue redeemable securities. As a result, the share price of a closed-end fund is determined by the market. Another distension between these two types is the management style. Open-end managers must comply with multiple objectives, including maintaining cash on hand and a minimum degree of liquidity of underlying assets to accommodate redemptions.

## 1.2 Structure and scope

The goal of this master thesis is to describe and evaluate machine learning methods on the capacity of selecting the best mutual fund. By predicting future mutual funds returns, we will compare the performance of each Machine Learning algorithms on their accuracy and efficiency. Indeed, the forecast will be compared to the original values to outline the performance or underperformance of the machine learning tools. The model to achieve this goal will be explained in chapter 4, and machine learning methods algorithms used will be presented in chapter 3.

Machine learning is defined as an umbrella term encompassing methods, models, and algorithms to uncover patterns and trends with the vast array of data processing tasks without explicit programming instructions. Thus, this thesis will focus only on the following machine learning techniques: OLS regression, Least Absolute Shrinkage and Selection Operator called Lasso, Elastic, Random Forest, XG Boost and Neural Network. The reason for these particular chosen methods will be explained in chapter 3. The selection and forecast model will be fed with inputs variables called target variables that could carry high-quality predictive information. In such ways, the model can select and forecast future returns of mutual funds.

This master thesis, structured with a combination of theoretical and empirical aspects, initiates by presenting the context and background of the goal's formulation. In the second place, a summary of the finding of existing and recent literature on features for selecting and forecasting mutual funds is presented. In Chapter 3, the methodology of machine learning techniques is explained. Findings on the application of machine learning methods in the finance sector will also be discussed. The description of data, the input features, and the model description are given in Chapter 4. In Chapter 5, the empirical study on a sample of open-ended mutual funds during the time frame 1990-2021 will be analysed, and the results will be presented. In Chapter 6, the results of the empirical study will be further discussed and reflected. The thesis concludes with Chapter 7 and suggests new paths to be undertaken in this research area.

## 2. Preview of empirical findings and contributions

Extensive literature has explored the identification of high-performing mutual funds using a wide array of factors. Hence, the possibility of using fund characteristics to pre-determined mutual funds' performance such as fund size (Berk and Green, 2004; Chen, Hong, Huang, and Kubik, 2004), fund age (Otten and Bams, 2002; Yap and Pierce, 2008), fund flows (Zheng, 1999; Sirri and Tufano, 1998), expense ratio (Walker & Droms, 2001; Bers, Madura, 2000) has been widely studied. In addition, numerous fund performance studies provided evidence on predicting future mutual funds returns by using past performance (Grinblatt and Titman, 1992; Ferreira, Keswani, Miguel, and Ramos, 2012) as past alphas, R-squared (for example, Hendricks, Patel, and Zeckhauser, 1993; Carhart, 1997; Mamaysky, Spiegel, and Zhang, 2006), t-statistic of alpha measuring the significance of fund performance (Elton, Gruber, and Blake, 2011). Other papers exploit the underlying stock's momentum (Grinblatt, Titman, and Wermers, 1995). In this section, in a more detailed view, we will see empirical studies of each factor on identifying performance or future returns of mutual funds.

### 2.1 Prediction of future mutual funds returns with past performance, historical information and alpha

In the choice of mutual fund investment decision, the most obvious and most straightforward analysis for investors is to use past performance and historical information in their decision-making process (Zheng, 1999). Researchers believe that the study of historical returns contains information on the future and could help to predict future returns.

This effect of past performance on the current performance of the mutual fund has been termed as "persistence in performance" (Kaur, 2018). Bollen and Busse (2005) commented that mutual fund managers possess a significant persistent ability to select a mutual fund, allowing an astute investor to predict future performance based on past results. Gruber (1996) stated that if management ability exists and is not included in the price of open-end funds, the performance of mutual funds should be predictable. For him, future performance is predictable from past performance because the price of buy and sell of mutual funds is equal to the net asset value, which does not change to reflect superior management.

Grinblatt and Titman (1992) study the relation of past performance on mutual fund performance. They find evidence of long-run persistence in performance. Indeed, the differences in performance between funds persist over time, and this long-run persistence is consistent with the ability of fund managers to earn abnormal returns (Grinblatt and Titman, 1992). As a result, for every 1% abnormal return achieved on a rollback of five years, mutual funds are expected to realize 0.28% greater abnormal returns in the second five-year period. In addition, Hendricks, Patel, and Zeckhauser (1993) analyze the extent to which the relative performance of mutual funds can be predicted by studying shorter evaluation horizons. Although on a short-run phenomenon, peaking each four-quarter approximatively, they also found significance on relatively superior fund performance.

Moreover, Elton, Gruber, and Blake (1995) examine predictability for stock mutual funds using risk-adjusted returns. They exhibit that past performance is predictive of future risk-adjusted-performance. For them, employing modern portfolio theory to construct a portfolio based on past returns improves the selection of a portfolio of mutual funds. Elton, Gruber, and Blake (1995) concluded that the past carries information about the future. The performance of a fund contains information on its future performance. Thus, a fund overperforming in the past tends to overperform in the future and on a risk-adjusted basis.

Yet, Ferreira, Keswani, Miguel, and Ramos (2012) illustrate short-term persistence in the performance of mutual funds through the analysis of mutual fund flows across 28 countries. They found that only in the case of the USA, mutual fund flows depend on past performance. However, not in the other countries. This result was also found previously by Bollen and Busse (2004). No persistence in performance other than the USA was found. Besides, there has been no evidence for persistence in performance measured with returns contrary to risk-adjusted return. Otten & Bams (2002) investigate 506 European mutual funds from 1991 through 1998. This paper shows an overview of the European mutual fund industry by exploring mutual fund persistence performance in France, Germany, Italy, the Netherlands and the U.K. They applied the Carhart (1997) 4-factor asset-pricing model. The United Kingdom funds was the only European country demonstrating statistically significant mutual funds persistence performance effects. From the results of studies on the impact of fund past performance on future performance, most authors seem to agree on the

existence of short-term performance persistence in the U.S. However, outside of the United States, the results are mitigated.

Although past performance can also be measured on historical information and returns, alpha can also be a measure for quantifying past performance. In portfolio performance assessment, the two popular performance indices are Jensen's alpha and the Sharpe index.

In the context of a single index model, Jensen's single criterion was introduced as a regression of the capital market excess returns on the fund excess return. Defined as the "average incremental rate of return on the portfolio per unit time" (Jensen, 1968), it measures abnormal performance. If the CAPM is the correct model of equilibrium returns, then the portfolio should lie on the Security Market Line, and the value of alpha should be zero (Cuthbertson, Nitzsche & O'Sullivan, 2010). It is hypothesized that if the alpha value is positive and statistically significant, it indicated superior risk-adjusted performance or a good manager skill to select funds.

According to Cuthbertson, Nitzsche & O'Sullivan (2010), a positive alpha demonstrates a portfolio performing better than a random selection buy-and-hold strategy. Berkowitz, Schorno & Shapiro (2017) findings reveal that both positive (hot-hand funds) persistence tend to have higher alpha. On the contrary, negative (icy-hand funds) persistence is associated with a higher expense ratio, lower alpha, and lower capital inflow. This single factor modelling can be extended to a multi-factor framework to improve the portion of variance explained by the regression. The four-factor model is primarily based on the empirical findings of Fama and French (1992 and 1993) and Carhart (1997). This model will be explained in further parts.

## 2.2 Relationship of Mutual fund characteristics and performance

### *2.2.1 Effect of Mutual fund size on performance*

In recent research, the relationship between a fund's size and the performance of funds gained notoriety. Yet, arguments pointed out by literature are contradicting each other. For instance, Tang et al. (2012) discovered an inverted U-shaped relationship while Ding, Zheng, Zhu (2015) found a U-shaped relationship. In opposition, a negative relationship was established by Chen et al. (2004) Otten and Bams (2002).

Using a sample of Brazilian funds markets from January 2009 to March 2014, Malaquias, Duarte de Sarvas, Oliveira (2016) concluded the size as an essential variable to address the performance of investments funds in Brazil. This conclusion result from the empirical evidence of a U-shaped relationship. The "U-shaped" size-performance represents the outperformance of small funds compared to large counterparts when the size of the fund is relatively small or the outperformance of large funds when the size of funds is relatively large. For instance, in terms of the size of a threshold. The "inverted U-shaped" size-performance is the opposite: when the size of the fund is relatively small or when the size of funds is relatively large, the fund's performance is underperforming. In contrast to Malaquias, Duarte de Sarvas, Oliveira (2016), Tang, Wang, Xu (2012) found evidence of an inverted U-shaped relationship between fund size and performance on the Chinese capital market.

A negative relationship between size and performance is argued as the flexibility of trading small funds in the financial market. As Grinblatt & Titman stated, small mutual funds "have an advantage over large funds in that they can more easily purchase and sell securities without altering securities price" (Grinblatt & Titman, 1989, p. 407). According to Ding, Zheng, Zhu (2015), funds with small sizes are more efficient in exchanging ideas among different layers of management, which contributes positively to their overall performance. In contrast, the positive relationship is explained as their large fund having significant market power that drives the stock price to increase or decrease depending on their trading activities. Thus, small funds having lower market power to influence price are subject to the price impact caused by the trading activities of large funds (Ding, Zheng, Zhu, 2015). Indeed, in the study of Grinblatt & Titman, the portfolio of mutual funds with the smallest net asset value realized the most extensive performance with an average performance of 2.5% per year. However, this high performance is minimized due to the high transaction costs. They will display the highest transaction costs and the most prominent survivorship bias along with the other portfolios. Survivorship bias is described by Prather, Bertin, Henker (2004) as the overstated average performance measure due to the records of poorly performing funds lost through liquidation or merger. For instance, Morningstar Principia Pros (MS)[1] annual data files include previous years' returns for only existing funds. Also, large counterparts have the possibility to

[1] Tools for in-depth investment research and portfolio analysis. It provides equity mutual fund data, including performance information, reporting and fund characteristics.

diversify their portfolios and decrease risk to a further degree than small funds. In a period of market downturn such as financial crisis, this is an essential element to consider. Large funds with comparatively large cash and liquid assets holdings are more competitive in seizing new investments opportunities.

According to numerous investigations, the potential effect of size on fund performance implies economies of scale and diseconomies of scale. According to the comprehension of Busse, Chordia, Jiang, Tang (2014), diseconomies of scale is explained as the following: a fund faces significant percentage transaction costs for a given set of stock holdings due to the increase of total net asset of funds. As its size increase, the fund will potentially increase the number of its holdings. To alter a particular fraction of the portfolio, a higher requirement of transactions of larger dollar amounts will be needed. As a result, it will increase transaction costs and thereby reduce subsequent performance. Keim & Madhavan (1997) express similar reasoning. Large funds tend to be involved in large transactions that are difficult to execute because of their impact on price. This difficulty of execution leads to high transaction costs that undermine performance. But higher transaction costs are not the only reason for diseconomies of scale in the mutual fund industry. Busse, Chordia, Jiang, Tang (2014) found that characteristics such as holding smaller market capitalization stocks, stocks with greater book-to-market ratios and higher momentum account for diseconomies of scale in the mutual fund industry. Grinblatt and Titman (1989), Yap and Pierce (2008), Ferreira, Keswani, Migues and Ramos (2012) demonstrate that the size of the fund to be inversely related to the performance of the fund, implying diseconomies of scale. It suggests reflecting the ability of managers to select funds as due to their increasing size, large funds face challenges to locate meaningful and sensible investments.

Many other researchers have empirically found out the effect of economies of scale on fund performance. They argue that large funds can improve their performance by reducing their average fixed cost. Stoll and Whaley (1983) show that, based upon portfolios formed on firm size, the average price per share is monotonically increasing in size. Besides, both the mean portfolio return, and beta are monotonically decreasing in size. Chen, Hong, Huang, and Kubik (2004) study the effect of scale on performance in the fund size on the US equity funds. They discover that funds size erode performance as the net fund returns declined with lagged fund size. Chen, Hong, Huang, and Kubik (2004) explain this association due to liquidity. Tang et al. (2012) concluded that the

inverted U-shaped between size and performance resulted in the effect of the economy of scale and liquidity. In its research, it is observed that in Chinese mutual funds, for small funds, economies of scale play an essential role; “however, for large funds, the role of liquidity is substantial” (Tang et al., 2012, p.246). Grinblatt & Titman suggested that small funds “may experience higher transaction costs than larger funds because they cannot take advantage of certain economies of scale” (Grinblatt & Titman, 1989, p. 407). This suggestion is similar to the founding of Busse, Chordia, Jiang, Tang (2014). They observed that large funds experience lower percentage transaction costs than smaller funds, with the top quintile having an annual performance drag of 0.19 percent and the bottom quintile 0.30 percent.

Finally, the finding of Fama and French (1992) reveals a strong negative relationship between stock returns and size. For him, smaller firms tend to have higher average returns with a spread of 0.74% per month on average based on their size rankings. Indeed, the size factor SMB of Fama and French )1992) (“small minus big”) measure the difference between the returns on small versus big stocks.

### *2.2.2 Effect of Mutual Fund Age on performance*

The fund age is defined as the duration for which the fund has been in existence and measures if the fund can survive an extremely competitive environment (Babbar, Sehgal, 2018). Recent literature has studied the association between the age of the funds with mutual funds performance. This association can be distinguished by two opposing views: a positive association and a negative one. Otten and Bams (2002) found a positive relationship between fund age and performance in five European countries. Babalos et al. (2007) showed a positive effect of age of fund on performance in Greece. As Kaur (2018) suggest, this significant correlation can be explained as older funds may perform better than newer funds due to the “learning by doing” effect. Ferreira et a;.(2012) comments that this negative relationship can result since more recent funds have to face higher start-up costs and challenges. As a result, new funds could be put at a disadvantage compared to older funds. As older funds have more knowledge, they can bear the costs or plan unexpected ones.

In the contrary, Otten and Bams (2002) and Yap and Pierce (2008) find a negative relationship. For instance, Peterson, Petranico, Riepe, and Xu (2001) and Prather, Bertin, and Henker (2004)

showed that younger funds perform better than older funds or that fund age and performance bear little relation. Ferreira et al. (2012) suggested that since young funds are new to the industry, they will tend to work harder and be more dedicated to achieving recognition and maintaining their survival in the industry.

Nevertheless, Białkowski and Otten (2011) and Murcia (2011) find an insignificant relationship between age and performance. This was also the case for Ferreira et al.(2012), which didn't find any relation between age and performance by analyzing the US and non-US funds. Both categories haven't any significant correlation.

### *2.2.3 Effect of Mutual Funds Flows on performance*

The mutual funds' flow-performance relationship has constituted the subject of numerous articles. The so-called flow of mutual funds is defined as the net growth in fund assets beyond reinvested dividends (Sirri and Tufano,1998). The flow, formally calculated in Chapter 4, reflects the percentage growth of a fund in excess of the growth that would have occurred if all dividends have been reinvested and no new funds had flowed in (Sirri and Tufano,1998).

Gruber (1996) termed the "Smart money as the net money flows in the mutual funds earning better returns. He studied this aspect by ranking actively managed funds from 1985 to 1994. This study put in evidence the ability of investors to select funds by analyzing risk-adjusted returns of new cash flows invested in these actively managed funds. Indeed, the return on the newly invested money is higher than the average return for all investors in these funds. Zheng finds the same result by employing a test of performance introduced by Grinblatt & Titman (1993). In aggregate, investors are able with the use of cash flow information to make mutual funds selection decisions based on the good assessment of short-term future performance (Zheng, 1999). Gruber's so-called "smart money" effect is, for Zheng, an effect related to fund-specific information used by investors in their mutual fund investment decisions (Zheng,1999). This effect is also linked to the size effect such that superior returns are principally earned by new money flows into and out of small funds than large funds. Hence, the information of aggregate new money flows, determined by Zheng as the information effect, resulting in abnormal returns, can be used to outperform the market, especially for small funds.

Siri and Tufano (1998) remark the similar reasoning: consumers base their fund purchase decisions on prior performance information. However, he commented that consumers were invested asymmetrically, such their investment is disproportionately more in equity funds that outperformed in previous periods. While Gruber (1996) findings have shown that the aggregate pattern of consumer investing behavior is rational such that investors invest in funds receiving inflows and disinvest from those experiencing outflows (Gruber, 1996). Hence, they would earn a risk-adjusted return that outperforms passive index funds even after fees. This result was also consistent with Zheng's finding on small funds.

Sirri and Tufano (1998) provide evidence on the fact that mutual fund flows are affected by factors related to the search and information costs bear by consumers. Sirri and Tufano (1998) demonstrate that high-fee funds beat their rivals on the performance-flow relationship. As flows are fee-sensitive, they assumed that high-fee funds spend more on marketing than their rivals as their percentage fees decline as the fund grows. Thus, consumers are highly informed about their current and potential investments. With such a marketing spotlight, consumers can visualize their performance. Since investors tend to redeem more of their shares of a fund that has shown a good performance, Sirri and Tufano (1998) concluded that performing funds were attracting inflows. As a result, a convex relationship between flows and performances has been proved. Throughout their study, Goldstein et al. (2017) also confirm this convex shape for equity funds' flow-to-performance.

Other research focuses on the mutual funds flow-performance correlation in the case of a negative relationship. For instance, Berk & Green (2004) believe that there is a correlation between performance and mutual fund flows. Chou & Hardin (2014) examine the relationship of fund size, and fund flows with the performance of real estate mutual funds (REMFs). Their expectations of findings were to meet the same result as Berk and Green (2004). For them, real estate mutual funds return exceed relevant benchmarks on a before expenses basis and match benchmark returns after expense. According to Chou & Hardin (2014), the negative correlation between the increase of fund flows and size with returns is explained by the rise of the expense ratio, which reduces the flow of capital. The flow of capital was measured by the difference of increase between the current and previous month in assets under management with the fund's monthly return divided by assets under management of the previous month.

In addition, Goldstein et al. (2017) demonstrate a concave flow-to-performance relationship on corporate bond funds as the sensitivity of outflows on bad performance is higher than inflows on a good performance. Besides, if a corporate bond fund possesses more illiquid assets and an overall high illiquid market, it will significantly impact the sensibility of outflows on bad performance. Goldstein et al. (2017) show in an illiquid corporate bond market, as one standard deviation increase in outflows, a decline in corporate bond returns by 31 to 36 basis points result. In contrast, for funds with liquid asset holdings and when the corporate bond market is liquid, as one standard deviation increase in outflows, only a decline in fund returns by only 5 to 7 basis points results. This Sharpe negative flow-performance relationship is due that investors remaining in the fund will bear higher costs such that these remaining investors will be encouraged to redeem before the others (Goldstein et al., 2017)

### *2.2.4 Effect of the Portfolio turnover on performance*

The turnover ratio is another important characteristic of the fund in analysing the performance and selection of mutual funds. According to Prather et al. (2004), turnover measures the trading activity or the propensity of a manager to trade. Hence, a high turnover ratio should reflect an active market timing and stock selection strategy compared to a low turnover ratio indicating a buy-and-hold strategy. Therefore, the portfolio turnover ratio reflects the active trading strategy effect on performance. The turnover ratio is formally calculated on the first step on individual funds as the lesser of monthly purchases or sales of securities divided by average monthly assets (Prather et al., 2004). The second step is divided by the mean turnover ratio for all funds with the same investment objective (Prather et al., 2004).

Carhart (1997) mentioned that mutual fund managers claim that turnover does not reduce performance. Since investors are paying for the quality of the manager's information, mutual fund managers argue that they only trade to increase expected returns net of transaction costs.

The perception of mutual fund managers is similar to Prather et al. (2004). Prather et al. (2004) consider that high turnover significates high transaction costs like brokerage fees and bid-ask spreads. The following suggest that turnover rate has a negative association with fund performance. However, Prather et al. (2004) nuanced it by suggesting it could be a signal that a manager is

gathering and trading on the information. Grinblatt and Titman (1994) proved a positive relationship between performance and turnover. They suggested that a high turnover ratio would significantly increase funds in an active trading way and represent underpriced securities. Thus, turnover may have a positive impact on performance even if it occurs high transaction costs.

Bers & Madura (2000) poses the hypothesis that the turnover ratio had a positive influence on net asset value persistence. Their reasoning was explained that as managers have superior knowledge and are constantly buying and selling funds, high turnover funds should reflect higher closed-end net asset value fund returns. Indeed, investors could believe that the signal given by a high turnover the manager has superior knowledge compared to rivals and will thus outperform regarding future net asset value return. Therefore, it could help increase investors' confidence, reduce overreaction to negative news and be associated with superior market price return persistence. Malkiel (1995) demonstrated that there is a strong correlation between turnover and unrealized appreciation. He explains that fund turning rapidly their portfolio are unlikely to a numerous amount of unrealized appreciation. Thus, high turnover funds can be expected to sell at smaller discounts, reducing overreaction and generating market price persistence.

However, Bers & Madura (2000) also suggested that higher transaction costs caused by higher turnover will offset more than the gains from selling and buying new stocks. As a result, higher turnover would persistently provoke lower net asset value performance. Bers & Madura (2000) concluded that in either case, the turnover ratio had an impact on net asset value return performance persistent as long as turnover also stays persistent.

Carhart (1997), instead, found a negative influence of portfolio turnover on performance. On average, he found that mutual funds do not bear their investment costs through higher returns. For every 100 basis points increase in turnover, performance drops by about 95 basis points for every buy-and-sell transaction.

### *2.2.5 Effect of the Expense ratio on performance*

According to the literature, high expenses should be justified by superior investment services, called higher risk-adjusted returns. Expenses paid for obtaining portfolio management services

provided by mutual funds are reported as the expense ratio. Hence, the authors believe that mutual funds can be predictable with the factor expense ratio. Carhart, for instance, explains that the predictability of mutual funds is only related to two elements: common factors in stock returns and persistent differences in mutual fund expenses and transaction costs (Carhart, 1997). Thus, the study of this factor on the potential of predicting future mutual fund returns flourished.

Grinblatt & Titman (1994) study the performance of 5 characteristics reported in the Wiesenberger Investment Companies Service via regressions such as the following: net asset value, load, expense ratio, turnover and management fee on a reliable and unreliable benchmark. The first goal of this research is to find how the performances of mutual funds are related to these 5 characteristics. The second goal was to discover whether these mutual funds performances are due to differences in transaction costs. The regressions result illustrates that there is no statistically reliable relation between the performance and expense ratios or net asset values. However, they found a significant association of combined management fees and load with mutual funds performances (Grinblatt & Titman, 1994). For every 10 bp reduction in management fee, the performance increases by 50 bp per year, indicating that low fee managers are superior performers. This result was also proved by Porter & Trifts (1998). Porter & Trifts (1998) examines the performance of 93 funds managed by experienced managers over 10 years from 1986 to 1995. They didn't find a causality link between the persistent performance of the fund and managers managing the same fund for a long period. Indeed, managers with 10-year track records of the same fund didn't outperform managers with shorter track records. Similar to Grinblatt & Titman (1994), try to determine in a second step to what extend expense ratios are associated with the performance rankings of the experienced managers. However, contrary to Grinblatt & Titman (1994), they discover that funds with low management fees have shown persistent positive performance while those with high fees has shown persistent negative performance. For instance, over the 4 first years of the study period, superior funds had significant lower expense ratios across funds with on average 0.47% less than the expense ratios of inferior funds (Porter & Trifts, 1998).

Ranked actively managed funds into deciles based on monthly four index alpha and monthly returns from 1985 to 1994. Gruber (1996) demonstrates that the best performing funds have an expense ratio similar to the expense ratio of average funds and lower than bottom deciles funds. In his reasoning, he indicated that investors are better off buying low expense ratio funds and in

parallel that good managers have lower expense ratios than bad managers. It presumes that there is a negative correlation between fund performance and expense ratios.

The study of Droms & Walker (2001) tests the persistence of mutual fund returns characteristics such as turnover rates and expense ratios over the 20-year period from 1971 to 1990. Using Jensen's equation (1968) as a risk-adjusted total returns measure, they find that the expense ratio has no significant effect on the performance for any of their multivariate models and that expense ratios do not persist across mutual funds between the period analyzed. Nevertheless, Indro et al. (1999) found also that mutual funds returns are negatively correlated with expense ratios. This significant negative association, according to them, means that on average mutual funds are overinvesting in information and, more particularly of acquiring and trading information. As a result, Indro et al. (1999) think that prior studies on mutual fund returns related to the cost of acquiring and trading on information failed to recognize the dependency of net assets under management of the funds for the efficiency of an active investment management strategy. In order to achieve sufficient returns to cover the expenses such as costs for acquiring and trading on the information, Indro et al. (1999) suggest that actively managed mutual funds have to attain a minimum fund size.

The analysis of the fund size in terms of net asset value (NAV) on the mutual fund's returns has been done by Bers & Madura (2000). This study examines if the persistence of both net asset value (NAV) and market price returns of U.S registered closed-end funds is related to funds characteristics (Bers & Madura (2000). Bers & Madura (2000) showed that the significant persistence of strong net asset value (NAV) and market price-performance was stronger on funds with lower expense ratios and funds traded on the NYSE. After a deeper analysis, they even discovered that with the management of closed-end funds combined with lower expense ratios, they could achieve more persistent positive returns, particularly for net asset value performance persistence. In addition, closed-end funds traded on the NYSE, managed by more experienced managers, are more likely to perform persistently well than other closed-end funds. This result was seen at least for a three-month holding period. Moreover, the variation in performance persistence levels among winner closed-end funds instead of loser closed-end funds was more effectively explained by fund-specific characteristics. This last result can be interpreted as the ability of investors to distinguish winner funds that will experience persistent performance by simply looking at fund specific features.

In more practical and recent research, Kaur (2018) investigates Indian equity mutual funds' performance during 2004–2013. She assumed that expense ratios have a negative effect on the different fund performance measures used during that period. However, the results show a significant positive impact of expense ratio on gross returns and Sharpe ratio (at 5 percent level of significance) Yet, the result also leads to an insignificant negative effect on conditional Carhart alpha. Kaur (2018) nuanced the results regarding the effect of the fund's expense ratio on fund performance. For her, the effect of fund characteristics depends upon the performance measure chosen. For instance, gross returns (unadjusted returns) can be subject to manipulations. In comparison, alpha provides risk-adjusted returns relative to the market.

In another country such as Pakistan, Ahmad, Sun & Khan (2018) obtained evidence of the significant positive effect of management fees on equity mutual funds performances. In opposition, the expense ratio had a negative correlation with the performance of the equity mutual funds. However, these observations should be carefully analysed. Indeed, the magnitude of both effects is different. The management fee effect positively affected the mutual fund performance with a minimum of 0.3 basis points and a maximum of 0.5 basis points. Contrary to the expense ratio decreasing from 15 basis points until 26 basis points depending on the performance measure used (CAPM, FF 3-factors and Carhart 4-factors) (Ahmad et al., 2018). Thus, it indicates an expense ratio having a stronger effect than the weakened positive management fee effect.

## 2.3 Estimation of fund performance with macroeconomic derivative-factors

Volatility is another factor that can be used for estimating fund performance. Indeed, Livingston et al. (2018) made hypotheses such as that high expense funds have higher volatility in performance than low expense funds. They argue that trading volume is positively related to asset price and return volatility as it is established by Kandel & Pearson (1995) or Karpoff (1987). This association comes from two theories of differences of opinion and overconfidence. The differences-of-opinion hypothesis states that investors substantially interpret public information differently, leading to a higher trading volume and larger price or return volatility (Harris & Raviv, 1993; Kandel & Pearson, 1995). Harris & Raviv (1993) find absolute price change and volume are positively correlated. The overconfidence theory posits those large differences of opinion result from the overconfidence of investors and traders in the accuracy and precision of their information (Odean,

1998). Odean (1998) believes that overconfidence increases the expectation of trading volume, increases market depth, decreases overconfident traders' expected utility, lower risk-adjusted returns, and higher return volatility. The overconfident person has high importance on the effect on volatility and price-quality (Odean, 1998): Overconfident traders could cause the market to underreact to the information of rational traders. Markets underreact to abstract, statistical, and highly relevant information, and they overreact to salient, anecdotal, and less relevant information.

Livingston et al. (2019) demonstrated that high turnover funds have higher volatility in performance than low turnover funds. This can be explained as actively managed funds are likely to take more idiosyncratic risk as they have the tendency to pursue investment and trading strategies that deviate from benchmarks. Berkowitz, Schorno & Shapiro (2017) focus on characteristics of mutual funds with extreme performance periods, considered as the period with high volatility. Funds with either positive or negative persistence tend to have portfolio similarities consistent with riskier positions: holding few stocks, investing more in top ten holdings, and having a higher portfolio beta. Beta shows the amount of systematic risk managers hold (Berkowitz, Schorno & Shapiro, 2017). For Livingston et al. (2018), double adverse effects are described: high expense ratios and turnover ratios reduce fund performance and increase volatility. Thus, Livingston et al. (2018) concluded that actively managed funds have higher volatility in performance as the superior average performance of actively managed funds is accompanied by higher volatility in performance.

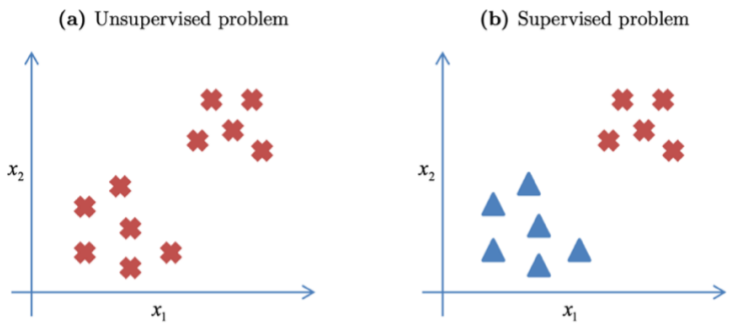
## 3. Methodology

In this section, machine learning will be described by starting with the definition of machine learning. Then, we will enunciate the empirical research found on machine learning techniques. In the last part, we will explain the seven different machine learning models we will use throughout this thesis.

Machine learning is defined as an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment (El Naqa I., Murphy M.J., 2015). Machine learning methods and algorithms allow to exercise and learn in an automated way without explicit programming instructions. It is commonly known that the more the size of the data is significant, the more efficient and successful machine learning can be.

Machine learning is composed of two main types of tasks: supervised and unsupervised. Unsupervised learnings designate a set of algorithms employed when the target variable is undefined and unknown. This learning algorithm is meant to discover hidden structures or patterns in a dataset without labels. This type of machine learning is shown in Figure 1.a, where the input features x1,x2 are similar or with no labels. Machine learning algorithms will then try to find patterns, relations or clusters to separate the data set without knowing the correct answer beforehand.

*Figure 1: Machine learning problems*




Source: Jakob Biehl Kristoffersen (2016) Applying artificial intelligence in equity selection. Facing the challenge to beat the S&P 500 index by applying machine learning on equity fundamental data Master thesis. (pp10). Accessed as : https://research-api.cbs.dk/ws/portalfiles/portal/58423508/jakob_biehl_kristoffersen.pdf

While, in supervised learning, the target variable is known such that the correct answer is initially given. In Figure 1.b, the supervised problem is demonstrated. For a given set of data, all the samples have been labelled with a symbol according to the class it belongs to. In a supervised learning problem, according to the target's nature, the prediction tasks differ. This distinction in output type has led to a naming convention for the prediction tasks: regression when the output is a quantitative measurement, classification when it is a qualitative measurement. The classification model is used when the goal is to set the input as the target resulting in a qualitative output. A classification model can have categorical, real-valued or discrete input variables.

In contrast, the model of regression is achieved for predicting targets with continuous values. An example of such a case would be predicting future stock prices based on historical firm characteristics such as revenues or size. Algorithms like neural networks can solve both models, while linear regression can only solve regression problems with labels corresponding to discrete categories.

## 3.1 Empirical Research on the Application of Machine Learning Techniques

Research on the application of machine learning techniques in finance is relatively active. Dhagash, Desai, Pradeep (2020) used a Random Forest model to categorise mutual funds. The authors used S&P Dow Jones Benchmark and FTSE/Russell Benchmark with 89 unique features for S&P Dow Jones Benchmark and 42 unique values for the other. For them, the wide variety of mutual funds from various investment management firms and diversification strategies have become available due to the surge of popularity of mutual funds as diversified financial investment. They defined accuracy as the fraction of correct predictions predicted by the model. The accuracy is assigned 1.0 if the samples match the valid label, else zero. The accuracy achieved by the methods is the following: 90.20% for the Random Forest, 87. Wolff, Echterling (2020) 70% for the decision tree and 90.70% for the deep neural network method. implement machine learning algorithms for stock selection with data of the S&P500 over the period from January 1999 to March 2021. Using techniques like Deep Neural Networks, Long Short-term Neural Networks, Random Forest, Gradient boosting and regularized logistic regression, they found a significant outperformance of machine learning-based stock selection compared to an equally weighted benchmark. The neural networks and tree-based models outperform logistic regression methods.

In addition, Moritz and Zimmerman (2016) also establish the effectiveness of tree-based models to predict portfolio returns. They ascertain that strategy returns and information ratios based on this model were more effective in terms of accuracy of the expected return predictions than alternative models. Moreover, Fu et al. (2018) construct technical and fundamental features to characterize each stock and label stocks in order to distinguish the "good" stocks from "bad" ones. They apply machine learning algorithms such as logistic regression, Random Forest, Deep Neural Networks. They demonstrated significant effectiveness in stock selection within the Chinese stock market. Besides forecasting returns, Alberg and Lipton (2017) use the deep neural networks technique to forecast future fundamentals based on a trailing 5-years window. They assume that financial data such as revenue, operating income, and debt provide insights into the company's financial health. Thus, they proposed forecasting company fundamentals as they believe that the long-term success of an investment depends on the current stock price with respect to its future fundamentals. As a result, the improvement in mean square error compared to a naïve strategy[2] was found significant.

Lastly, Gu, Kelly, and Xiu (2019) discover the efficacy of machine learning methods in the context of asset pricing. For them, the use of machine learning by investors for forecasting represents enormous economic gains. For instance, they found that an investor which competes with the S&P 500 by investing in bottom-up neural network forecasts has a 26-percentage point increase in annualized out-of-sample Sharpe ratio (0.77) relative to a buy-and-hold investor (0.51). They analyse the following machine learning techniques: Ridge Regression, Lasso, Elastic Net, Random Forest, Gradient Boosting, (Deep) Neural Networks. They identify that the best performing methods are regression trees and neural networks.

Further, they reveal evidence of the reason of the aids of the accuracy of machine learning in return predictions compared with ordinary least-squares (OLS) regressions. Indeed, they attributed the improved accuracy of machine learning models to uncovering non-linear patterns and their robustness to multicollinear predictors. In particular, for the "elastic net" technique, they found that as this technique uses shrinkage parameters and variable selection, it helps to limit the regression's

[2] Naive strategy is simplistic strategy where all the stocks have been bought equally.

degrees of freedom, which solves the OLS inefficiency problem. Besides, they confirm that neural network performance peaks at three hidden layers. Afterwards, more layers are added, less performant is the neural network. Likewise, for the algorithms as a Boosted Tree and Random Forest, increasing the leaves of the selected trees reduces the accuracy of the prediction. Finally, they concluded that the most successful informative stock-level predictors among all the methods studied are price trends, liquidity and volatility variables. Price trend variables concern stock momentum, industry momentum and short-term reversal. For liquidity, variables are such as market value, dollar volume, and bid-ask spread. Lastly, volatility variables include return volatility, idiosyncratic volatility, market beta, and beta squared.

## 3.2 Machine Learning Methods

Machine learning was invented 64 years ago by Frank Rosenblatt with the "Perceptron". This invention is a single-layer neural network, an algorithm that would be further detailed (3.3.3). For selecting and forecasting the designated open-ended mutual funds, the following machine learning methods were used: OLS regression, Lasso (Least Absolute Shrinkage and Selection Operator), Ridge Regression, Elastic Net, Random Forest, XG Boost and Neural Network. This section will introduce these seven methods.

### *3.2.1 Ordinary Least Square Regression (OLS)*

The linear model, often called linear regression, remains one of the most commonly supervised learning algorithms from the field of statistics. Linear models are the simplest parametric methods and deserve the proper attention for their simplicity, flexibility and unbiasedness.

A linear model is based on the assumption that it can approximate the output values through a regression process. In general, regression models predict an output Y from a series of N inputs X = $(X_1, X_2, \dots, X_n)$ as follow:

$$Y = \beta_0 + \sum_{i=1}^{N} \beta_i X_i + \varepsilon \qquad (1)$$

There is a relationship between two variables such that when one variable increases, it will impact another variable to increase or decrease proportionately. The term $\beta_0$ is the intercept, also know as bias in machine learning. The independent variable must be linearly related to dependent

variables. There should not be any outliers in values among independent variables. If there are any outliers, they should either be capped, or a new variable needs to be created flagging the outliers' data points. Linear methods have different types which differ from each other in the way they estimate the coefficients.

The ordinary least-squares model (OLS) technique is the most commonly used method for estimating linear regression's unknown parameters. It can be used as a simple linear regression model assessing the relationship between independent (input) and dependent variables (output). The relationship between input and output variables can be calculated by using the following formula:

$$Y = \alpha + \beta x_i + \varepsilon_i \qquad (2)$$

This technique can also be seen as an unconstrained quadratic programming problem. It estimates β that minimised the residual sum of squares (RSS), also known as training mean squared error, as follow :

$$\hat{\beta}_{OLS} = arg\min_{\beta} RSS(\beta) = \sum_{t=1}^{T}(\mathcal{Y}_i - \beta_0 - \sum_{i=1}^{N}\beta_i \mathcal{X}_{ij})^2 \quad (3)$$

The OLS estimator is famous due to its features. It is unbiased when the linear model is correct, and due to the implication of the Gauss- Markov theorem, it has a minor variance out of all unbiased estimators. The error values are distributed on both sides of the fitted line. Due to homoscedasticity, errors cannot get larger as the value of an independent variable increase. Errors should be normally distributed.

However, the standard OLS estimation of the linear regression coefficients may not be appropriate for forecasting because of its significant variance and its underperformance to outliers.

### *3.2.2 Shrinkage-Based Methods*

Shrinkage-based methods aim to improve prediction accuracy by shrinking the estimated parameters or setting them to 0. As a result, it decreases the variance and avoiding overfitting. There are three methods: Ridge, Lasso and Net Elastic.

These methods are regularised subject to three norms l0, l1 and l2. These three norms are defined by their canonical form as follow :

$$\| \beta \|_0 = \sum_{j=1}^{p} I(\beta_j \neq 0)$$
$$\| \beta \|_1 = \sum_{j=1}^{p} |\beta_j|$$
$$\| \beta \|_2^2 = \sum_{j=1}^{p} \beta_j^2 \qquad (4)$$

The l1 regularization adds an l1 penalty equal to the absolute value of the magnitude of coefficient or simply restricts coefficients' size. The l2 regularization adds an l2 penalty which is equal to the square of the magnitude of coefficients.

*3.2.2.1 Lasso (Least Absolute Shrinkage and Selection Operator)*

Introduced by Robert Tibshirani in 1996, the Lasso method, or so-called Least Absolute Shrinkage and Selection Operator, is a penalised least squares technique based, as the name suggests, a shrinkage model.

This method puts the l1- norm penalty on the sum of the model parameters' absolute values by shrinking the coefficient of the regression variable to 0. Thus, with an upper bound on the sum of the model parameters' absolute values, the sum of squared errors is minimised. The Lasso estimator $\hat{\beta}$ for the linear regression model (5) is given as follow :

$$\hat{\beta}_{Lasso} = arg\min_{\beta} RSS(\beta) + \lambda \sum_{i=1}^{N} |\beta_i| \quad (5)$$

The parameter alpha controls the type of shrinkage, which has important consequences for the estimation method's properties. The regularisation parameter λ controls the extent of this penalty. If λ is sufficiently large, then coefficients are set to be precisely equal to zero due to the l1- norm penalty geometry. Thus, Lasso has an l1 penalty on the parameters and performs both parameter shrinking and variable selection. In such a way, the curse of dimensionality is reduced.

Increasing inputs that are truly associated with the outputs should lead to reduced set error. However, if these inputs add only noisy inputs, then the set error would only increase. Dimensionality refers to the fact that additional inputs, even relevant, engender variance. Indeed, the variance incurred in fitting coefficients may outweigh the reduction in bias that they would add.

The Lasso method provides accurate prediction due to the shrinkage process. However, the more significant the parameter λ, the more the coefficients are shrunk to zero. In this way, the prediction error is minimised. But on the other hand, if the regularisation parameter λ = 0, an exceptional case result. Indeed, it will be no more than just a simple OLS regression that doesn't bring value to the forecast. The key motivation for this method is the bias-variance tradeoff in mean squared error. By shrinking the estimator in Lasso, it trades off an increase in bias with decreased variance.

*3.2.2.2 Ridge Regression*

Ridge regression is another method based on shrinkage. Ridge regression is a popular parameter estimation method used to address frequently the collinearity problem in multiple linear regression. This method adds a constraint on the l2 norm of the coefficients as follow (6) :

$$\hat{\beta}_{Ridge} = arg \min_{\beta} RSS(\beta) + \lambda \sum_{i=1}^{N} \beta_i^2 \quad (6)$$

This method shrinks the coefficients of regression variables on their size by penalising large coefficients. Given the l2 norm on the parameters, the alpha parameter is set to 0. Higher the λ is, more the coefficients $\beta_i{}'s$ are shrunk toward 0 without shrinking $\beta_0$. Simplistic, it is due to the mean value of the response variable. In turn, it makes the coefficients more robust to collinearity. This condition permits the differentiation with the Lasso method as the coefficients are not exactly 0. It means that Ridge regression will not enforce the irrelevant variable coefficients to become 0. Instead, it will reduce the impact of these variables on the model. Ridge regression and Lasso regression also differ in how they handle correlated variables. While Ridge regression shrinks correlated variables toward each other, Lasso typically selects one. Therefore, Ridge regression tends to perform better than Lasso when the predictors are highly correlated. However, both

methods need their inputs to be standardised so that the coefficients' size is comparable and penalised fairly on the same scale.

#### *3.2.2.3 Elastic Net*

To resume, the Lasso performs variable selection, but it tends to select only one input from a group of highly correlated variables ignoring all the other inputs. The method of Ridge alleviates the significant coefficient that can arise for highly correlated inputs, but it does not perform variable selection. Thus, the last method based on the shrinkage method is proposed by Zou and Hastie in 2005 called Elastic Net. This method is the combination of both the Lasso method and the Ridge Method.

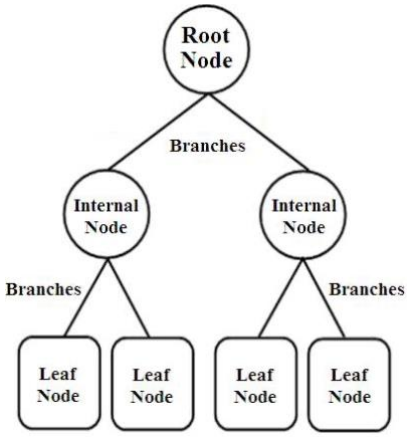
The elastic net involves two non-negative hyperparameters, λ and ρ, and includes two well-known regularizers as special cases (l1-norm and l2-norm). Hence, the Elastic net overcomes these limitations by limiting the number of selected variables and stabilising the selection from grouped variables. This is done by putting penalties on both l1 and l2- norm. Thus, it accepts biased coefficient estimates in return for lower variance. When the alpha parameter is set to 0 (ρ = 0 case), the penalty function reduces the l1 term, corresponding to the lasso regression. According to DeMiguel (2021), the fortunate geometry of the Lasso sets coefficients on a subset of covariates to precisely zero. In this sense, the Lasso imposes sparsity on the specification. Therefore, the l1-norm term can control the degree of sparsity and be thought of as a variable selection method. While, by setting the parameter to 1 (ρ = 1 case), the penalty function reduced the l2 term, which corresponds to the ridge method. The ridge regression, which uses an $l_2$ parameter penalization, draws all coefficient estimates closer to zero but does not impose exact zeros anywhere. In this sense, Ridge is a shrinkage method that helps prevent coefficients from becoming unduly large in magnitude and can thus increase stability. Therefore, to optimise the Elastic Net, the alpha value needs to be between 0 and 1 as it will encourage shrinkage and selection. As a result, some coefficients will be shrunk, and others will be exactly 0.

### *3.2.3 Tree-Based Method*

Tree-Based methods are unique and valuable for interpretation. These types of approaches are known as decision-tree methods as the set of rules used to segment the predictor space can be

summarized in a visual representation of a tree. Decision Tree is one of the most straightforward yet powerful machine learning algorithms for classification and regression tasks. Figure 2 describes the components of a decision tree. A decision tree is composed of two main elements: Nodes and Branches. Nodes represent the possible attributes associated with an event. Nodes can be differentiated into three different types: root nodes, internal nodes and leaf nodes. Root nodes represent the decision to be made. It represents the attribute with the largest information gain. Internal nodes represent one of the possible choices available given the tree structure. Leaf nodes represent the outcome of the combination of decisions or events. Finally, branches represent the outcomes emanating from the root and internal nodes.

*Figure 2: Components of a decision tree*




Source : José Alberto Silva de Sá (2016) Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy. (pp2). doi : 10.21528/CBIC2011-27.1

### *3.2.3.1 Random Forest*

Random Forest is an ensemble learning method again based on Decision Trees. Random Forest consists of multiple decision trees such that each tree depends on the values of a random vector. This random vector is independent of the past random vectors and with identical distribution for all trees in the forest (Breiman, 2001)

Concretely, each tree in a Random Forest is first trained simultaneously on the sample data using dawns replacements, known as bootstrapping. The term bootstrap refers to selecting a few rows randomly. Thus, each decision tree will be trained on a slightly different set of observations. It is

such that every single tree will grow differently according to the diverse bootstrapped sample and that the final predictions will be made by averaging each decision tree. By averaging across the prediction of the numerous decision trees, the prediction variance is reduced, and thus the predictive performance of trees are stabilized. Therefore, by increasing the number of trees in the Random Forest, out-of-sample performance is often improved. However, as previously written, according to Gu, Kelly, and Xiu (2018), the marginal improvement is decreasing with the number of trees growing as well as the accuracy of the prediction with the increase of the leaves. Moreover, a larger Random Forest requires more computational resources.

#### *3.2.3.2 XG Boost – Extreme Gradient Boosting*

XG Boost stands for Extreme Gradient Boosting. It is a specific implementation of the Gradient Boosting method. The Gradient Boosting algorithm is a model used for regression and classification problems. Gradient boosting uses ensembles of decision trees. Gradient refers to the error or residual obtained after building the model. Boosting refers to improving. The technique is known as a way to improve and reduce error gradually.

Contrary to aggregating independent decision trees and taking an average like Random Forests, Gradient boosting aggregates decision trees sequentially to connect week decision trees ("week learner") to form single strong trees ("strong learner"). Thus, the GBM algorithm iteratively adds a new decision tree at each step that best reduces the loss function. This latest decision tree gives more weight to poorly observations predicted by previous trees that failed to achieve a good performance. The logic behind the scenes is that subsequent learners learn from their predecessors' mistakes and correct these errors by being trained on the residual of the prediction of the previous trees. The algorithm continues to iterate until the prespecified number of trees by the user are reached.

In this fashion, the Gradient Boosting Method improved predictions each time by fitting on the residuals of decision trees. According to Schapire and Freud (2012), boosting improves predictions by reducing the prediction variance and prediction bias. To avoid overfitting, a regularization technique requiring additional hyperparameter can be employed. For instance, a key hyperparameter in Gradient boosting could be the learning rate. The learning rate determines how

fast the method learns from its most recent tree, meaning how much weight it gives to the most recent decision tree. Or it could also impose constraints on the number of decision trees aggregated, the depth and number of nodes of each tree, and the minimum number of observations in a leaf node.

Based on the original framework of the Gradient Boosting Method, according to Meng, Yang, Qian and Zhang (2020), it uses K additive trees to approximate the output. Contrary to Gradient Boosting Method, XG boost is a technique aiming to minimise the regularised objective function, divided into two components. The first component is the training loss, which measures the difference between the predicted and actual values. (Meng, Yang, Qian and Zhang, 2020, p.475). The second component is the regularisation term which permits to measures the complexity of the model. The regularisation term can help to smooth the final learnt weight to avoid overfitting. (Meng, Yang, Qian and Zhang, 2020, p.475).

The two other features of XG Boost are the Shrinkage and the Column Subsampling. (Meng, Yang, Qian and Zhang, 2020, p.475). At each step of boosting, the shrinkage technique scales the newly added weights. (Meng, Yang, Qian and Zhang, 2020, p.475). Therefore, the influence of each tree is declined and thus permits to avoid overfitting. (Meng, Yang, Qian and Zhang, 2020, p.475). Further, the method of Column Subsampling for building a given tree will select only a random subset of input features for reducing the training process time (Meng, Yang, Qian and Zhang, 2020, p.475).

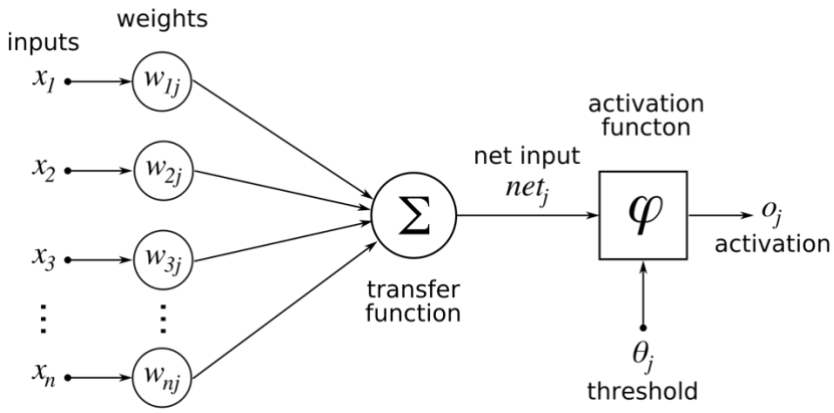
### *3.2.4 Deep Neural Network*

The final non-linear method we will analyze in this thesis is the neural network or artificial neural network. Commonly known as the most powerful modelling device in machine learning, they are used for complex natural language processing or computer vision problems. The neural network is a system of artificial neurons connected and consisted of multiple layers. Between neurons in consecutive layers, there are connections that disappear between neurons in the same layer.

From a mathematical point of view, a neuron is essentially a non-linear transformation of its input. Inside a neuron, there are two stages of operations which are illustrated in Figure 3. These stages

have the goal to weigh, aggregate and transform the inputs. The first step is a linear transformation of the inputs, while the second step is a non-linear activation function that is applied to the linearly transformed inputs.

*Figure 3: Architecture of a single neuron in a neural network*

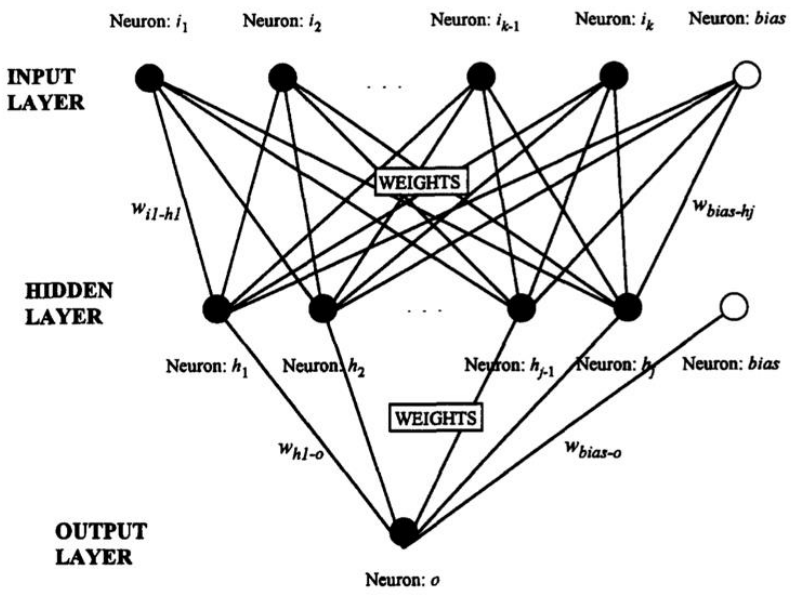



Source : Chrislb (2005) Modelling Systemic Risk using Neural Network Quantile Regression (pp4). Master Thesis. doi: 10.18452/19318

In the beginning, each neuron makes a weighted sum of its inputs. Then, it applies a non-linear activation function to its aggregated signal before sending its output to the next layer. Thus, these activation functions control the output as a result depend on them. Also, they provide non-linearities effects for the network. Finally, the results from each neuron are linearly aggregated into an ultimate output forecast.

The most straightforward neural network consists of three layers: one input layer, one output layer, and one hidden layer in the middle, which is demonstrated in Figure 4.

*Figure 4: Neural network architecture*

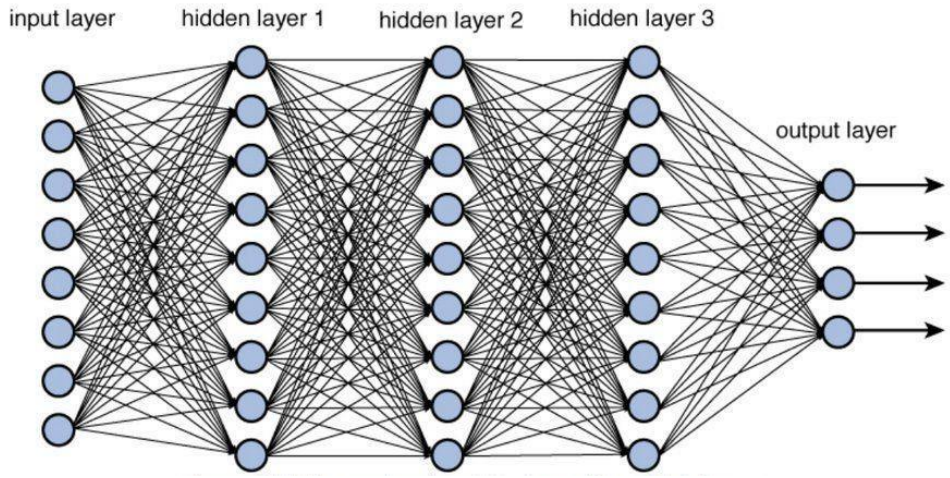



Source: Chiang, W. C., Urban, T., & Baldridge, G. (1996). A neural network approach to mutual fund net asset value forecasting. Omega, 24(2), 205–215. (pp207).. https://doi.org/10.1016/0305-0483(95)00059-3

The input layer collects the input patterns and values from the features of the data. Then, it transmits directly to the hidden layer. The number of neurons in the input layer is the same as the number of features by construction. The hidden layer gets its values from the input layer. After nonlinearly transforming the inputs, it passes those values to the output layer. The hidden layer is used to optimize the input variable's weight and bias until the network lowers its loss error. Lastly, the output layer is the layer that aggregates hidden layers into an ultimate outcome prediction.

A deep neural network is a neural network with many hidden layers that requires a massive amount of data to train. As the example of a deep neural network shown in Figure 5, the number of parameters grows extremely fast with number of hidden layers and number of neurons in each layer. Thus, while it increases tremendously, the complexity of the model engenders an increase of difficulty to illustrate it in a visual representation.

*Figure 5: Illustration of Deep Neural Network architecture with multiple layers*




Source : Ravindra Parmar (2018). Training Deep Neural Networks. Deep Learning Accessories. Website as: https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964, consulted on 13 August 2021

# 4. Data and predictive features

In this section, we will explain the data set that we used in this thesis and the model for selecting our mutual funds.

## 4.1 Data description and preparation

We obtain monthly U.S equity mutual funds data as well as any other data from Bloomberg. Kaminsky, Lyons, & Schmukler (2001) study the performance of funds and discover that equities have outperformed bonds as the largest share of mutual fund portfolio beginning in the 1990s. With this discovery, we supposed that we would obtain the best results by starting in 1990. Thus, our sample starts in January 1990 and ends in June 2021, totalling 31 years and six months. We have specified that the country of domicile is the United States and the fund type as the open-end funds to extract these data. For the fund asset class focus, we precise it to be equity and that this equity is the fund primary share class. Thus, from a match of 39518 U.S mutual funds, we end up after all specifications written above to a result of 3408 funds. After cleaning and removing all the data who didn't possess the last price of the mutual fund, the final sample has a number of 289 mutual funds. Table 1 summarizes all types of principal investment strategies that our mutual funds are composed of according to the classification system of Bloomberg of our sample

According to Bloomberg fund classifications[3], the asset class focus of equity is divided into two categories: strategy and market capitalization. Strategy is composed of the following strategies: value, growth, blend, bear market, long/short, geographical focus, industry sector focus and market neutral. There are small, mid, large and broad (Multi) capitalizations in the market capitalization categories. In the appendix 1, the definition of each investment strategy is explained.

[3] Source: Bloomberg (2013). Bloomberg Fund Classifications: Guide to our new funds Classification System. PDF Document available on the website : https://cpb-us-e1.wpmucdn.com/wp.txstate.edu/dist/2/916/files/2020/01/tmp14.pdf. Consulted on 13 August 2021

*Table 1: Principal investment strategy of mutual funds from our sample*

| Objective mutual funds | Number of mutual funds |
|---|---|
| Asian Pacific Region | 2 |
| Basic Materials Sector | 2 |
| Blend Broad Market | 6 |
| Blend Large Cap | 27 |
| Blend Mid Cap | 4 |
| Blend Small Cap | 9 |
| Communications Sector | 3 |
| Consumer Discretionary | 3 |
| Consumer Staples | 1 |
| Emerging Market Stock | 1 |
| Energy Sector | 6 |
| European Reg. ex UK | 1 |
| European Region | 1 |
| Financial Sector | 4 |
| Foreign Blend | 14 |
| Foreign Growth | 9 |
| Foreign Value | 9 |
| Growth Broad Market | 19 |
| Growth Large Cap | 30 |
| Growth Mid Cap | 15 |
| Growth Small Cap | 12 |
| Health Care Sector | 8 |
| Industrial Sector | 2 |
| Japan | 1 |
| Natural Resources Sector | 3 |
| North American Region | 1 |
| Precious Metal Sector | 8 |
| Single Country | 2 |
| Technology Sector | 5 |
| Thematic Sector | 9 |
| Utilities Sector | 3 |
| Value Broad Market | 19 |
| Value Large Cap | 33 |
| Value Mid Cap | 6 |
| Value Small Cap | 10 |

To calculate the monthly return, we compute the continuously compounded return or so-called logarithmic return. It is calculated by taking the logarithm of the mutual funds returns at time T divided by the mutual funds return at time T-1.

To predict the future data of the mutual funds, we choose multiple features such as benchmarks of stock markets from various countries. We took benchmarks such as SPX Index, NKY Index, DAX Index, FDTR Index. We believe that the benchmarks could impact the selection of United States funds.

The SPX Index is the American stock market index S&P 500 ( Standard & Poor's 500) ticker. Widely regarded as the best gauge of large-cap U.S equities, there is over 11.2 trillion indexed or

benchmarked to the index[4]. The index, including 500 leading companies, covers approximately 80% of available market capitalization. Therefore, the S&P 500 is recognized by investors as the best representation of the U.S stock market and a bellwether for the U.S. economy. On the contrary, the Nikkei Stock Average (NKY Index) is considered the benchmark and indicator of the Japanese Stock Market[5] movement. The index includes 225 stocks listed on the Tokyo Stock Exchange First Section. While the DAX Index is the German stock market index. The DAX[6] shows the performance of the 30 largest and most liquid companies called blue chips listed on the Frankfurt Stock Exchange. Like the S&P 500, the DAX components also represent around 80% of the total market capitalization of the listed companies in Germany.

FDTR Index was also taken as we believe that it has an impact on the U.S mutual funds. FDTR is the upper bound index of the federal funds' target range. Depository institutions lend reserve balances to other depository institutions overnight on an uncollateralised basis at the federal funds rate. Thus, the higher this interest rate is, the more attractive are investments in the United States as credit cards, mortgages, and bank loans are affected by this rate. According to Guo (2004), during 1974–1979 and in the late 1980s, they found that stock prices reacted significantly to unanticipated changes in the federal funds rate target. From this discovery, we hypothesised that the federal funds rate target could potentially influence mutual funds with an asset-class focus of equity. The federal funds rate also affects the currency, which a higher rate increases the value of the U.S dollar. Therefore, it could potentially affect mutual funds denominated in the U.S dollar and thus on our sample set of U.S mutual funds.

We also took the volatility, such as the VIX Index, as a feature for our model. The VIX measure the market's expectation of future volatility. Based on options of the S&P 500, it is known as an indicator of the U.S stock market and, more particularly, the U.S equity market volatility.

[4] S&P Dow Jones Indices. S&P500 (2021). Available on the website : https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview. Consulted on 13 August 2021.
[5] Nikkei Indexes. Nikkei Stock Average (Nikkei 225) (2021). Available on the website: https://indexes.nikkei.co.jp/en/nkave/index/profile?idx=nk225. Consulted on 13 August 2021.
[6] Deutsche Börse Group. Dax indices. The leading German equity index and more (2021). Available on the website: https://www.deutsche-boerse.com/dbg-en/products-services/ps-indices-analytics/ps-dax-indices. Consulted on 13 August 2021

In addition, we compute the monthly flows, calculated as below. The flow reflects the percentage growth of a fund in excess of the growth that would have occurred if all dividends have been reinvested and new funds haven't flowed in (Sirri and Tufano,1998) :

$$FLOW_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1} * (1 + R_{i,t})}{TNA_{i,t-1}} \quad (7)$$

where $TNA_{i,t}$ is fund *i's* total net assets or the dollar value of all shares outstanding at time *t*, and $R_{i,t}$ is the fund's return over the prior year.

Thus, our features to predict and select are composed of SPX Index, NKY Index, DAX Index, FDTR Index, Flow and the median of all mutual funds returns.

Moreover, we downloaded Fama French data to obtain several characteristics associated with the time-series regression of mutual funds returns on the five Fama and French (2015) and momentum factors. In particular, for each mutual fund selected in our sample, we run a "rolling- window" regression of the returns on the Fama French 5 factors and momentum factor for the previous 36 months. Further, we compute alpha (the intercept of the equation) as well as all betas' coefficients. Then, we also compute the $R^2$ from the FF5+MOM rolling- window regression as a characteristic to evaluate machine learning methods, as proposed by Amihud and Goyenko (2012). Lastly, we compute the monthly realized alpha as suggested by DeMiguel (2021) :

$$\alpha_{i,m} = r_{i,m} - \hat{\beta}_{MKT_{i,m}} MKT_m - \hat{\beta}_{SMB_{i,m}} SMB_m - \hat{\beta}_{HML_{i,m}} HML_m - \hat{\beta}_{RMW_{i,m}} RMW_m - \hat{\beta}_{CMW_{i,m}} CMW_m - \hat{\beta}_{MOM_{i,m}} MOM_m \quad (8)$$

where $MKT_m$, $SMB_m$, $HML_m$, $RMW_m$, $CMW_m$, and $MOM_m$ are the returns in month m of the five Fama-French and momentum factors, and $\hat{\beta}_{MKT_{i,m}}$, $\hat{\beta}_{SMB_{i,m}}$, $\hat{\beta}_{HML_{i,m}}$, $\hat{\beta}_{RMW_{i,m}}$, $\hat{\beta}_{CMW_{i,m}}$, $\hat{\beta}_{MOM_{i,m}}$ are the factor loadings of the mutual funds' excess return with respect to the five Fama-French and momentum factors estimated using the 36-month estimation window.

Finally, we use the monthly realized alpha defined in Equation (7) to compute the value-added for each mutual fund and month. Berk and Van Binsbergen (2015) define value-added as the below mathematical formula :

$$value - added = \left(\alpha_{i,m} + ER_{i,\frac{m}{12}}\right) \times TNA_{i,m-1} \quad (9)$$

However, due to a lack of resources, we could only get Bloomberg's last annual expense ratio. This expense ratio is the yearly average of the total expense. Thus, we calculate the value-added as below :

$$value - added = (\alpha_{i,m} + ER) \times TNA_{i,m-1} \quad (10)$$

Thus, the characteristics used to study the performance of machine learning methods in Chapter 5 are the following variables: monthly realized alpha, alpha, flows, value-added, $R^2$, and the coefficient betas of the market, profitability, investment, size,value, and momentum.

### 4.2 Model building and mutual funds selection strategy

Our model was built with the goal of explaining if machine learning methods were a good investment tool to select mutual funds for investors. Our model selects mutual funds depending on the predicted future returns from the machine learning methods. We believe that the selection should be made according to the future returns, predicted with the characteristics and the different machine learning methods explained above. We wanted to see if the selected mutual funds were the best choice and the best investment that could make the investor at a given time. We decided to be as realistic as possible. For instance, we choose only characteristics that could be available for investors. Hence, we can put in the shoes of a random investor who isn't a professional but willing to invest in mutual funds. Finally, if it wasn't a success, we wanted to know what could improve or change. For example, if it was due to wrong features.

As previously written, we implement our approach using monthly data US equity mutual funds from January 1990 until June 2021. We use the first 28- and 6-month years of data to train the machine-learning methods to predict three-year ahead returns.After importing our features, the

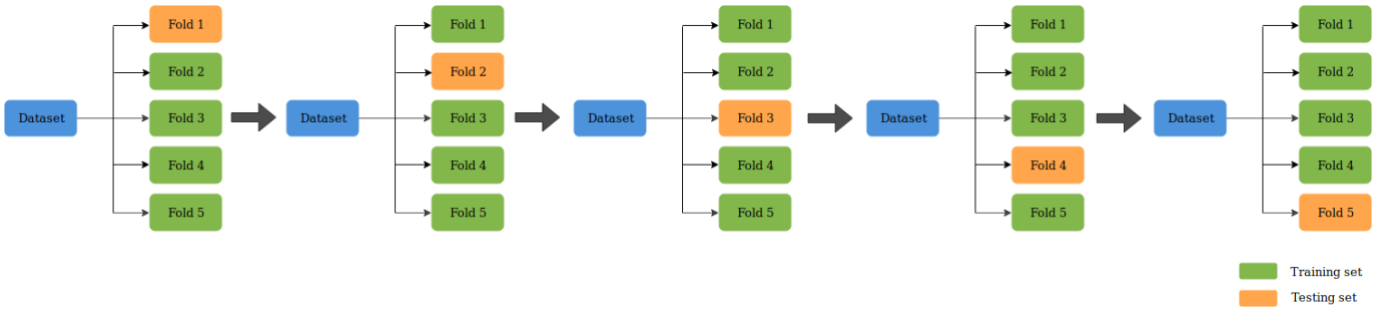
model scales the data according to our data set's mean and standard deviation. Then, it will ask which mutual fund will be chosen as the target variable to run the seven machine learning methods. We will run this code for all mutual funds.

Separating the dataset is an essential step in building a machine learning model. The objective is to divide the database into two parts: the first part where the classification algorithm will train and discover links between our explanatory variables and our explained variable, and the second part where the algorithm will attempt to predict the dependent variable, namely the target variable, on these new data. The more data the algorithm has to practice on, the more accurate it will offer. Thus, the dataset will be split with a ratio of 80 % for the training set and 20% for the testing set.

We use well-known R packages to implement the different machine-learning methods such as caret, xgboost, prophet. All the machine methods explained above will be run, such as OLS, Random Forest, Lasso, Ridge, Elastic Net, XG Boost and Deep Neural Network. The hyperparameters of the OLS, Ridge, Lasso, Elastic net, Random Forest are tuned using five-fold repeated cross-validation.

The K-fold cross-validation estimates the mean squared error of the test by simulating prediction on the training set. The first step will randomly shuffle the training set into a copy of the training set and divide the training set into k blocks or folds of approximatively equal size where each is used as a testing set. For each iteration, it will, for each block j, remove the jth fold and use the remaining folds to obtain the predictions corresponding to the different values of the hyperparameters. For instance, in the first iteration, the first block is used to test the model. The other blocks serve as the training set. This process is repeated until each fold of the k folds is used as the testing set. The below Figure 6 summarizes the k- fold cross- validation that we used.

*Figure 6: 5-fold Cross-Validation*




We then evaluate the prediction error of the prediction associated with each value of the hyperparameters on the jth fold. After completing this process for each of the k folds, we select the value of the hyperparameters that minimizes the average cross-validation error. To go deeper in the approach, we use the repeated k-fold cross-validation process. Similar to k-fold cross-validation, the only difference, it will repeat the process multiple times and report the mean performance across all folds. We decide to use repeated k-fold cross-validation as it allows the selection of models working well out of sample with low prediction error. Further, the model calculates the Root Mean Square Error, the Mean Square Error, the Mean Absolute Error, and the R-squared of each machine learning method for each mutual fund.

The Root Mean Square Error (RMSE), applied to a model, allows us to define the level of error that the system will emit in its predictions. Larger errors have greater weight in the forecasts in the model. This measurement is expressed in the following mathematical form:

$$\text{RMSE} = \sqrt{\frac{1}{\text{m}}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2} \quad (11)$$

where m is the number of variables in the database, $\hat{y}_i$ is the predicted value of the dependent variable for variable i and $y_i$ is the observed, expected value for item i. A value of 0 would indicate that the regression model fits the data perfectly. Therefore, a small value is preferable.

Mean squared error (MSE) evaluates the squared units of predicted and observed values. According to Sammut C., Webb G.I. (2011), the mean squared error of a model with respect to a test set is calculated with the mean of the squared prediction errors over all instances in the testing set.

Thus, it will tell how far the predicted value is from the true value, which is represented as the prediction error. Mean squared error is defined as follows:

$$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (12)$$

Where $\hat{y}_i$ is a predicted value, and $y_i$ is the observed value.

Mean Absolute Error is the mean of the absolute values of the individual prediction errors. Each prediction error is the difference between prediction and actual observation, where all individual differences have equal weight.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i| \quad (13)$$

Where $\hat{y}_i$ is a predicted value, and $y_i$ is the observed value.

A final tool that is often used is the R-squared. According to Miles (2014), R squared permits explaining the predictor variables in the sample as it represents the proportion of the variance. The independent variables permit the prediction of the portion of the variance in the dependent variable. The closer the value is to 1, the better the mode will be at predicting out dependent variable. However, in some specific cases, a low r-squared value doesn't illustrate a bad prediction. Suppose the independent variables are statistically significant such as high variance and highly dispersed residuals, while the r-squared has a low value. It is still possible to make the conclusion that the relationships between the variables and the regression model are the best fit for a given dataset. For instance, studies on human behavior that possess unexplained variations. R-squared is expressed mathematically according to the following formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i^m(\hat{y}_i - y_i)^2}{\sum_i^m(y_i - \bar{y})^2} \quad (14)$$

where $\hat{y}$ is the mean of the values of the dependent variable.

This formula makes the comparison between the fitted regression line and a baseline model. The baseline model represents the "worst" model and is a flat line assuming that all predicted values of $\mathcal{Y}$ will be the mean value of $\mathcal{Y}$. Thus, the R-squared will be able to see if the fitted regression line is better at predicting y than the mean.

After running all machine learning methods, we assemble the result in a table. The machine learning method with the lowest root mean squared error will be chosen to predict performance in the following year. However, since we wanted to see the difference in prediction return of all different techniques, we will predict mutual funds returns for every machine learning method. After prediction, we rank the prediction set according to returns. We divided the funds into 8 groups according to this ranking. Thus, each group will have 36 mutual funds categorized by their order. Group 1 represents the funds in the bottom decile and Group 8 with the top decile of the predicted performance distribution. We then form an equally weighted portfolio of the funds according to their groups.

We expand the training sample forward one year with the new predicted data, train the algorithms again and make new predictions for the following year. Then again, we construct groups according to the latest ranking of the new predicted set and track its return for the next 12 months. This procedure is done for every remaining year until June 2021. Thereby, we have a forecast of three years. Thus, we can track mutual funds and see if their ranking has evolved through time. We can also see which machine learning methods have the most accurate prediction by comparing them with the actual returns in the same period. Further, we are able to analyze the performance of the selection of the mutual funds based on regression of the returns on the Fama French 5 factors and momentum factor.

To analyze which feature is essential in our model to predict future mutual funds, we created a separate model that tells us the variable had highly predictive information. Different techniques will indicate which variable is the best one to predict future mutual funds returns. For this programming code, we use the median of all mutual funds returns as the target variable. The reason was to see which feature variable will indicate information to get as closer as possible to the actual returns of the past 28 and 6 months years.

As previously written, the separate model is based on different techniques, which will be explained further: Stepwise Regression, Multivariate Adaptive Regression Splines and Boruta.

Stepwise regression is a technique to select features in linear regression models. There are three strategies of stepwise regression such as forward selection, backward selection or so-called backward elimination and lastly, stepwise selection. In our model, we will use the forward selection and the combination of both, forward selection with backward elimination. The forward stepwise regression approach uses a sequence of steps permitting features to enter or leave the regression model one at a time. The entry and exit criteria are commonly based on a p-value threshold. Usually, a typical entry criterion will be a p-value less than 0.15 for a feature entering the model and greater than 0.15 for a feature leaving the model. The process ends when there are no features with a p-value inferior to 0.15. However, if a feature has a p-value inferior to 0.15, the feature with the lowest value will be retained. The process begins by creating a number N of linear regression models. Each linear regression model uses precisely one feature. The features will be ranked according to their ability to explain variation in the outcome. After the first cycle, the process will start again to build linear regression models with a number of N-1 of linear regression models. The process will continue to run these steps until the scope is meet.

Contrary to the forward stepwise regression, the backward selection will start with all model variables and remove features with the largest p-value that are the least statically significant. However, the process will then do, as the forward stepwise regression, create a number of N linear regression models. After the cycle, it will start again to create but with a number of N-1 linear regression models. Similar approach as before, the process will continue until the goal is reached. The criteria needed to perform backward selection is to have n greater than p, meaning more observations than variables. As the combination of both methods is more flexible than the backward selection, we decided to select this strategy.

Multivariate adaptive regression splines, simplified with the acronym MARS, is a flexible method for modelling high dimensional data. It will randomly separate 80% of the data as training samples and 20% of the data as testing samples. Simplistic, MARS can be imagined as an ensemble of linear functions joined together by one or more hinge functions. From the point of view of tensor

product splines methods[7], it looks similar to a stepwise backward linear regression or a decision tree CART. The model takes the form of an expansion in product spline basis functions. As Friedman (1991) explained, this technique aims to combine recursive partitioning (e.g. decision tree CART) and spline fitting to will optimize the positive aspects of both. With the help of data, the number of basic functions and the parameters associated with each one (product degree and knot locations) is determined. Thus, MARS is capable of tracking the dependencies of parameters and thus capturing high order interactions. Also, the model can be represented in a form that individually identifies the additive contributions and those associated with different multivariable interactions.

Boruta algorithm is an extension of the random forest method, presented by Kursa, Jankowski and Rudnicki (2010). It permits the identification of the true essential variables in an information system. The idea behind this algorithm is to copy the system in a randomized way and then assess the significance of the variable in the iterative fashion by comparing it with the randomized variables. Variables with higher importance than randomized variables will be considered important. Thus, in the first step, the system is extended, permitting variables to be replicated. The values of replicated variables are randomly permuted with objects such that the correlation of the replicated variables and the decision attribute is also random. In a second step, Random Forests are run with the specification of randomly choosing replicated variables before each run. Then, the importance of all attributes is computed. If the attribute for a single rum has higher importance than the whole maximum importance value of all randomized attributes, then this attribute is deemed important. Or the variable is considered important by counting the importance of the attributes being higher than MIRA (maximal importance of the random attributes) is significantly higher than the expected value. The ones that are considered unimportant are removed from the system. This procedure will end when the predefined number of iterations is reached, or all attributes are conclusively regarded as important or rejected.

[7] Tensor product splines provide a one-dimensional spline function.

# 5. Empirical Results

This section examines the performance of the seven-machine learning (methods explained in Section 3) for forecasting and selecting mutual funds as the model described in Section 4.

## 5.1 Analyse of Machine Learning Methods on mutual fund returns

Firstly, we examine the performance of our mutual fund portfolios between the different Machine Learning methods from July 2018 until June 2019. We assume that we invested in an equally weighted portfolio composed of the 288 U.S equity mutual funds with a weight of 0.35%.

As we can see in table 3, in the first year, all the 7 Machine Learning Methods have underperformed compared to the real monthly returns of the mutual funds. Apart from XG Boost, machine learning methods didn't achieve to forecast the up and downs of the actual returns. However, XG Boost manages to forecast this trend slightly. For instance, in December 2018, the equally weighted portfolio of all mutual funds had a drastic decrease of 17.07%, which XG Boost forecasted with a reduction of 2.90%. Yet, in January 2019, when the mutual funds achieve a performance of 8.67%, XG Boost forecasted a negative one equivalent to -1.42%. In appendix 2, the monthly returns for the second and third years are shown.

*Table 2: Comparison of monthly returns predicted by Machine Learning methods with the actual returns of the mutual funds and the benchmark S&P 500 from July 2018 to June 2019*

| Year 1: 07/2018 - 06/2019 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Date | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic | Actual | S&P 500 |
| 29/07/2018 | 0,33% | 0,33% | 0,58% | 0,99% | 0,80% | 0,33% | 0,33% | 2,31% | -3,54% |
| 29/08/2018 | 0,29% | 0,28% | 0,53% | 0,97% | -0,17% | 0,26% | 0,28% | 1,85% | -2,98% |
| 29/09/2018 | 0,24% | 0,24% | 0,48% | 0,82% | -1,08% | 0,23% | 0,24% | -0,27% | -0,43% |
| 29/10/2018 | 0,52% | 0,50% | 0,72% | 0,81% | 4,58% | 0,52% | 0,52% | -8,58% | 7,19% |
| 29/11/2018 | 0,32% | 0,32% | 0,57% | 0,69% | 0,47% | 0,33% | 0,33% | 0,73% | -1,77% |
| 29/12/2018 | 0,18% | 0,22% | 0,38% | 0,72% | -2,90% | 0,24% | 0,23% | -17,07% | 9,63% |
| 29/01/2019 | 0,22% | 0,22% | 0,48% | 0,88% | -1,42% | 0,21% | 0,22% | 8,67% | -7,57% |
| 01/03/2019 | 0,32% | 0,30% | 0,62% | 1,17% | 0,09% | 0,30% | 0,31% | 3,45% | -2,93% |
| 29/03/2019 | 0,41% | 0,39% | 0,66% | 0,79% | 2,54% | 0,40% | 0,40% | 0,59% | -1,78% |
| 29/04/2019 | 0,34% | 0,34% | 0,60% | 0,76% | 1,11% | 0,35% | 0,34% | 3,11% | -3,86% |
| 29/05/2019 | 0,42% | 0,41% | 0,67% | 0,85% | 2,80% | 0,43% | 0,42% | -6,39% | 6,80% |
| 29/06/2019 | 0,20% | 0,21% | 0,42% | 0,79% | -2,12% | 0,20% | 0,21% | 6,66% | -6,67% |

Table 3 compares the monthly returns predicted by Machine Learning methods with the actual returns of the mutual funds and the benchmark S&P 500 from July 2018 to June 2019. Table 3 reports the following measures for the equal-weighted portfolio of all mutual funds: Annual average return, Standard deviation of returns, Sharpe ratio (average return divided by the standard deviation), Positive Months, Value-at-risk (VaR) with 95% confidence and Conditional value-at-risk, known as the expected shortfall (CVaR). The conditional value at risk is more a conservative approach in terms of risk exposure. It will take the weighted average of the extreme losses in the tail of the normal distribution of possible returns. By comparing the average annual return and standard deviation of the same period shown in Table 4 (July 2018 to June 2019), machine learning methods were not convincing. Even XG Boost, with a standard deviation of 7.51% didn't properly show the reality as the exact standard deviation of this year is equivalent to 24.60%. However, the machine learning method which achieved to predict the positive month closely is the XG Boost method. The Sharpe ratio of the XG Boost is closely similar to the actual but don't have proper signification since it predicted a positive average annual return when actually it was negative. From the Value at Risk measure (VaR) at a significance level of 95%, all machine learning methods predicted a negative VaR as the reality. However, they minimized since the real VaR is at -40.46%. This same trend is also done for parametric CVaR at a significance level of 95%.

*Table 3: Summary of the performance of the statistics of the machine learning methods, S&P500 and mutual funds actual returns*

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic | Actual | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|
| Average Annual Return | 3,78% | 3,78% | 6,71% | 10,24% | 4,70% | 3,79% | 3,82% | -4,94% | -7,90% |
| Standard Deviation | 0,34% | 0,30% | 0,36% | 0,47% | 7,51% | 0,34% | 0,32% | 24,60% | 19,22% |
| Sharpe Ratio | 11,21 | 12,46 | 18,80 | 21,91 | 0,63 | 11,16 | 11,85 | -0,20 | -0,41 |
| Positive Months | 100,00% | 100,00% | 100,00% | 100,00% | 58,33% | 100,00% | 100,00% | 66,67% | 25,00% |
| VaR (95%) | -0,56% | -0,50% | -0,59% | -0,77% | -12,35% | -0,56% | -0,53% | -40,46% | -31,62% |
| CVaR (95%) (parametric) | -0,70% | -0,63% | -0,74% | -0,96% | -15,48% | -0,70% | -0,66% | -50,74% | -39,65% |
| CVaR (95%) (historical) | 0,32% | 0,31% | 0,56% | 0,85% | 0,39% | 0,32% | 0,32% | -0,41% | -0,66% |

As mentioned previously and demonstrated in Table 4, the XG boost method performed better than the other machine learning methods on forecasting the positive months for the year 1 from July 2018 to June 2019. Thus, we further study this characteristic by analyzing the drawdown period from the high watermark for the three years predicted (from July 2018 to June 2021).). These results are illustrated in Table 5.

*Table 4: Summary of the analysis of drawdown period from July 2018 to June 2021*

| POSITIVE PERIOD COUNT | | |
|---|---|---|
| **Statistics** | **XG Boost** | **Actual returns** |
| Positive Months | 20 | 24 |
| in % | 55,56% | 66,67% |
| Worst DD (%) | -4,73% | -33,03% |
| Maximum Days in DD (%) | 3,33% | 5,06% |
| Average Drawdown (%) | -1,03% | -7,25% |

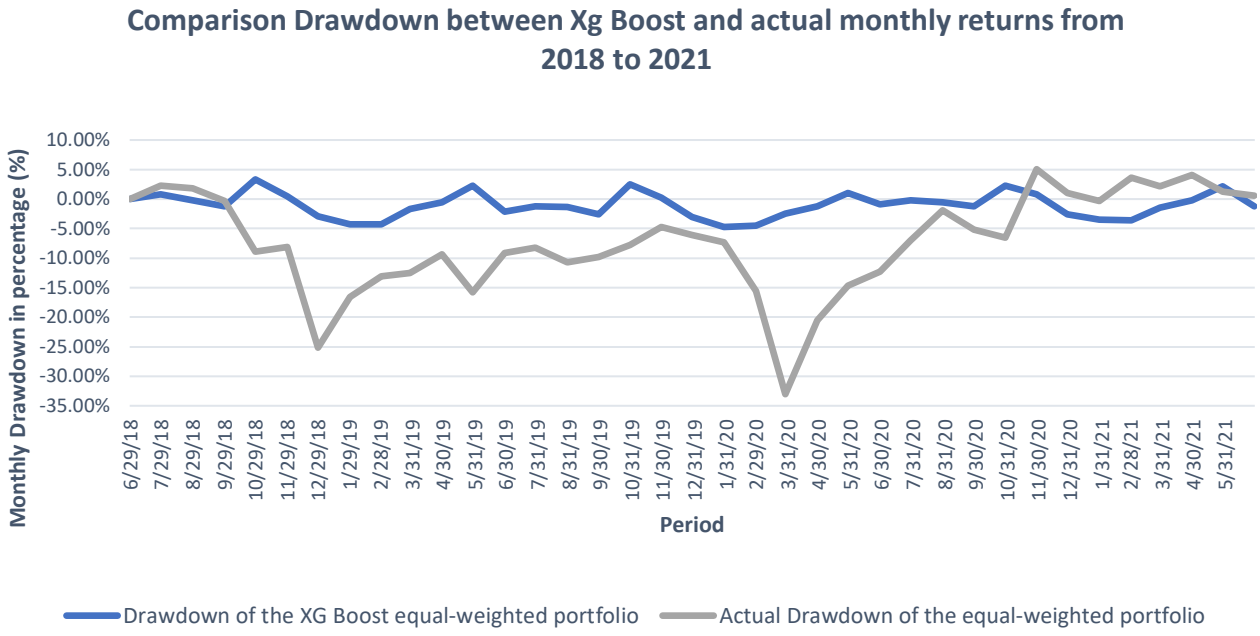
Table 5 shows that the XG boost predicted the actual positive months on the whole period (55.56 % compared to 66.67% in reality). 20 months out of 36 predicted months are positively months predicted by XG Boost, when in reality mutual funds beneficiated 24 positive months. In terms of the worst drawdown month, XG Boost minimizes it. The worst month was actually -33.03% which XG Boost predicted -4.73%. This difference is due to the Coronavirus Crisis's unexpected event that happened in March 2020, which XG Boost was unable to oversee. However, it did better on the maximum days in the drawdown period (5.06% in reality compared to 3.33% with XG Boost). As shown in Figure 7, XG Boost is trying to follow the same trend as the real mutual fund's drawdown period. However, in October 2018, May 2019, March 2020 and November 2020 until January 2021, XG Boost prediction was going in the opposite direction as in reality. Especially in March 2020, XG boost didn't handle to overview the drastic drop.

While, in December 2018, although slightly, it has recognized the decreasing tendency of the mutual funds. This is a good indicator of the performance of XG Boost since 2018 was a year fraught with volatility, characterized by record highs and sharp reversals such that it becomes the worst year since the Great Recession for investors. In December 2018, the S&P 500 dropped 20 percent from its record high on an intraday basis. Mutual funds of all Bloomberg classification types[8] (explained previously in chapter 4) declined. For example, Vanguard's Total Stock Market Index fund had its first loss since the financial crisis in 2008. In our set, all Vanguards open-ended mutual funds dropped in December 2018. For instance, Vanguard Explorer Fund investment had a decrease of 22% in December 2018. Also, in our set, in December 2018, the worst performance

[8] Source: Bloomberg (2013). Bloomberg Fund Classifications: Guide to our new funds Classification System. PDF Document available on the website : https://cpb-us-e1.wpmucdn.com/wp.txstate.edu/dist/2/916/files/2020/01/tmp14.pdf. Consulted on 13 August 2021

was Harbor International with a return of -53,98% and a fund objective of Foreign Value. As described in Appendix 1, the aim of Foreign value means that the fund primarily holds stocks deemed to be undervalued in price with foreign currencies and is likely to pay dividends. Thus, we can suppose that indexes of other countries than the United States, such as the DAX Index or NKY Index, were able to get predictive information to predict mutual funds of this type of classification.

*Figure 7: Evolution of the XG Boost and actual mutual funds drawdown period.*




Thus, we can conclude that in our set, XG Boost can better oversee mutual funds returns compared to other machine learning methods. Yet, it fails to foresee when unexpected events occurred like Coronavirus effects in March 2020.

By assuming again that we invested in an equally weighted portfolio of the 288 U.S equity mutual funds, we analysed the performance of machine learning in years 2 and 3 (from July 2019 until June 2021). During the second and third predicted years, like the first year, the results are not convincing in the direct comparison with actual results. As seen in Table 6, in the second year, all seven machine learning methods foreseen a positive trend contrary to the actual trend. In the third year, shown in Table 7, machine learning methods minimised the actual monthly mutual funds

returns. From July 2020 until June 2021, mutual funds had a drastic increase equivalent to 30.03%. The machine learning technique that was the closest was the deep neural network technique with 10.45%. Table 6 showed that XG Boost forecasted the real positive months but didn't handle the annual average return well. While in the third year (Table 7), XG Boost was 25% under the real positive months. Similar to the first year, the VaR, historical and parametric CVaR was minimised for all machine learning methods.

*Table 5: Summary of statistics performance of machine learning methods from July 2019 to June 2020.*

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic | Actual | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|
| Average Annual Return | 3,79% | 3,79% | 6,70% | 10,28% | 5,08% | 3,79% | 3,83% | -3,16% | -5,25% |
| Standard Deviation | 0,34% | 0,30% | 0,34% | 0,44% | 7,42% | 0,26% | 0,32% | 25,42% | 22,16% |
| Sharpe Ratio | 11,24 | 12,53 | 19,98 | 23,30 | 0,68 | 14,35 | 11,94 | -0,12 | -0,24 |
| Positive Months | 100,00% | 100,00% | 100,00% | 100,00% | 58,33% | 100,00% | 100,00% | 58,33% | 33,33% |
| VaR (95%) | -0,56% | -0,50% | -0,55% | -0,73% | -12,21% | -0,43% | -0,53% | -41,82% | -36,46% |
| CVaR (95%) (parametric) | -0,70% | -0,62% | -0,69% | -0,91% | -15,31% | -0,54% | -0,66% | -52,44% | -45,72% |
| CVaR (95%) (historical) | 0,32% | 0,32% | 0,56% | 0,86% | 0,42% | 0,32% | 0,32% | -0,26% | -0,44% |

*Table 6: Summary of statistics performance of machine learning methods from July 2020 to June 2021.*

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic | Actual | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|
| Average Annual Return | 3,78% | 3,78% | 6,73% | 10,45% | 4,80% | 3,79% | 3,82% | 30,03% | -32,65% |
| Standard Deviation | 0,26% | 0,24% | 0,28% | 0,56% | 5,97% | 0,26% | 0,25% | 13,39% | 14,16% |
| Sharpe Ratio | 14,37 | 16,02 | 23,85 | 18,57 | 0,80 | 14,35 | 15,28 | 2,24 | -2,31 |
| Positive Months | 100,00% | 100,00% | 100,00% | 100,00% | 50,00% | 100,00% | 100,00% | 75,00% | 25,00% |
| VaR (95%) | -0,43% | -0,39% | -0,46% | -0,93% | -9,82% | -0,43% | -0,41% | -22,02% | -23,28% |
| CVaR (95%) (parametric) | -0,54% | -0,49% | -0,58% | -1,16% | -12,31% | -0,54% | -0,52% | -27,62% | -29,20% |
| CVaR (95%) (historical) | 0,32% | 0,31% | 0,56% | 0,87% | 0,40% | 0,32% | 0,32% | 2,50% | -2,72% |

Secondly, we analyse the performance of the cumulative returns of the machine learning methods of the equal-weighted portfolio from July 2018 to June 2021, illustrated in Figure 8. Figure 8 represents only the predicted results of the machine learning methods. It is intentionally compared with the actual mutual funds' returns and the benchmark S&P 500 since we want to reflect the accuracy of the machine learning methods.

*Figure 8: Cumulative return of machine learning methods, S&P500 and actual mutual funds returns.*

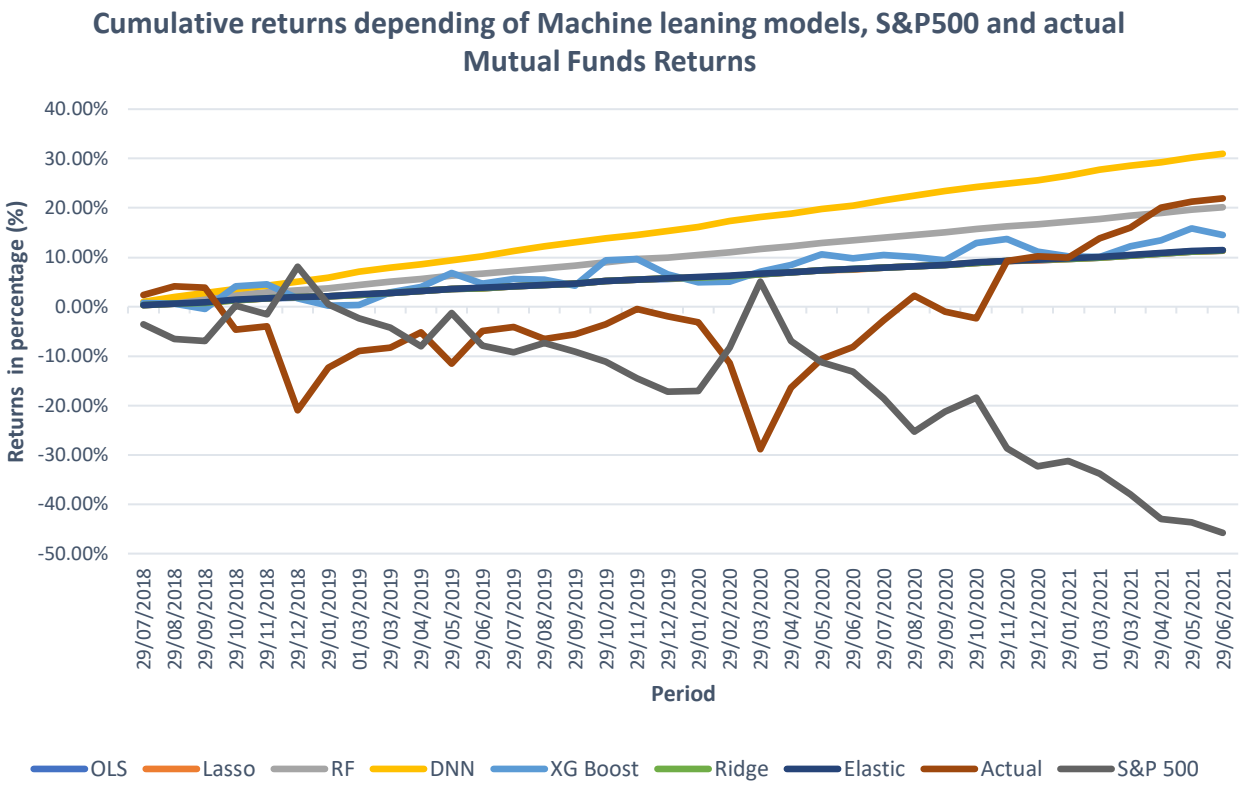



Figure 8 reveal that the machine learning method Deep Neural Network is the one that has the highest cumulative returns on the set period with 30.97%. OLS, Lasso, Ridge and Elastic have almost the exact results with a difference of approximatively 0.1%. For this reason, in Figure 8, the curves cannot be differentiated. Compared to the actual results, we can see that the machine learning methods are not accurate enough. Indeed, they are forecasting linearly, contrary to the reality that they have monthly returns with ups and downs. As shown in Figure 8, the XG Boost is the only machine learning method that didn't follow this linear and straight line. Besides, Random Forest is the strategy that was the closest to the actual results (21.94% compared to 20.14% in reality). Finally, we can also realize that by investing in these 288 mutual funds with a weight of 0.35%, we outperform the S&P 500 at an investment of 100% on the index between July 2018 and June 2021. The S&P 500 reached -45.80%, contrary to 21.94%. Yet, this performance is not attributed to machine learning methods as it was the pre-selection of mutual funds.

To conclude, we have analyzed the monthly returns of an equal-weighted portfolio composed of 288 mutual funds. Deep Neural Network predicted the highest cumulative return with 30.97%. In truth, 21.94% was achieved, which Random Forest predicted the closest with 20.14%. XG Boost was the machine learning method that outstands due to its ability to adapt more realistically compared to the other six machine learning methods. Indeed, the study of positive months and drawdown, we have seen that XG Boost follow slightly the same path as the actual results. Yet, XG Boost and the other methods struggled to select the appropriate mutual funds during the bear market and hard times.

## 5.2 Analyse of machine learning methods on mutual fund selection strategy

As explained in Chapter 4, each machine learning method predicts July 2018 to June 2021. At the end of each year, the predictions of each machine learning technique will be ranked. The mutual funds with the highest rank will be formed as a portfolio (called the group 8 in the algorithm) and be selected for investment. Thus, we evaluate the performance of machine learning methods by studying the portfolio chosen "Group 8". Group 8 represents the portfolio composed of 36 mutual funds with the highest ranking according to the machine learning methods selected for investment.

Table 8 and Table 9 represents the first predicted year (from July 2018 to June 2019). The details of the monthly returns of the Group 8 can be found in Appendix 3. Although the seven machine learning methods forecasted positive returns, only the Random Forest technique actually achieved positive average annual returns. Random Forest predicted an average annual return of 18.57%, which, in reality, by investing in the same mutual funds, would have achieved an average yearly return of 15.85%. The other machine learning methods overestimated the results. For instance, the XG Boost stated a 14.01% average annual return from July 2018 to June 2019, which actually represents a decline of average yearly return equivalent to -2.78%. In Table 9, we can see that by investing in Group 8 of the mutual funds chosen by Random Forest, it would have reached a Sharpe ratio of 5.02, which is very high. The portfolio would have achieved positive returns at 83.33% on the whole year.

On the contrary, the shrinkage-based methods and XG boost had positive returns only at 66.67%. The Deep Neural Network has the worst positive months out of all machine learning methods with

58.33%. It also had the worst results in terms of mutual funds selection. Hence, the Deep Neural Network stated a 49.57% average annual return on the choice of mutual funds, which actually represents -8.51% average yearly return on the whole first year.

*Table 7: Summary of average annual return of the machine learning methods performance from July 2018 to June 2019*

| Average Annual Return | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OLS** | **Actual** | **Lasso** | **Actual** | **RF** | **Actual** | **DNN** | **Actual** | **XG Boost** | **Actual** | **Ridge** | **Actual** | **Elastic** | **Actual** |
| 9,15% | -4,31% | 7,62% | -4,31% | 18,57% | **15,85%** | 49,57% | -8,51% | 14,01% | -2,78% | 9,13% | -4,14% | 9,18% | -4,08% |

*Table 8: Summary of results of Standard Deviation, Sharpe Ratio and positive months of the machine learning methods from July 2018 to June 2019*

| | **OLS** | **Actual OLS** | **Lasso** | **Actual Lasso** | **RF** | **Actual RF** | **DNN** | **Actual DNN** | **XG Boost** | **Actual XG Boost** | **Ridge** | **Actual Ridge** | **Elastic** | **Actual Elastic** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard Deviation | 0,38% | 26,45% | 1,87% | 26,45% | 0,60% | 3,16% | 1,98% | 25,49% | 8,97% | 27,19% | 0,36% | 26,37% | 0,35% | 26,43% |
| Sharpe Ratio | 24,28 | -0,16 | 4,08 | -0,16 | 30,98 | 5,02 | 25,02 | -0,33 | 1,56 | -0,10 | 25,25 | -0,16 | 26,44 | -0,15 |
| Positive Months | 100,00% | 66,67% | 83,33% | 66,67% | 100,00% | 83,33% | 100,00% | 58,33% | 66,67% | 66,67% | 100,00% | 66,67% | 100,00% | 66,67% |

Table 10 and Table 11 represents the second predicted year (from July 2019 to June 2020). The details of the monthly returns of the Group 8 can be found in Appendix 2. For this second year, shown in Table 10, Random Forest didn't achieve a similar performance as the previous year. Indeed, it had selected mutual funds with an annual average of -7.25% from July 2019 to June 2020. Deep Neural Network, which performs poorly previously, for this second year, predicted 8.26% and reached 7.21% average annual returns.

XG Boost reaches the same result as it has chosen the same 36 mutual funds as Deep neural Network. However, these performances should be nuanced as the volatility of these mutual funds being very high, demonstrated in Table 11. Deep Neural Network and XG Boost have a standard deviation of 24.82%, meaning these investments are risky. Thus, the Sharpe ratio is at 0.29, significate a low risk-adjusted-performance. The Shrinkage-based methods and OLS performed positive returns but disappointing as the average annual return of the portfolio is around 0.20%. These methods selected almost similar mutual funds (approximatively 2 to 5 different mutual funds) and thus have identical Sharpe ratios (0.01 for Elastic, OLS, Lasso and 0.02 for Ridge). In contrast to the first year, the number of months having positive returns is higher than in the second year. Although the selecting mutual funds had lower positive months, they still achieved positive average annual returns. This points out that the positive months had high returns to compensate for

the negative returns in negative months. As shown in Figure 7, the drawdown of the mutual funds was the deepest in March 2020, representing the Coronavirus crisis. Thus, even with the low returns of the whole set of mutual funds, machine learning techniques still choose suitable funds to accomplish positive returns.

*Table 9: Summary of average annual return of the machine learning methods performance from July 2019 to June 2020*

| Average annual return | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OLS** | **Actual** | **Lasso** | **Actual** | **RF** | **Actual** | **DNN** | **Actual** | **XG Boost** | **Actual** | **Ridge** | **Actual** | **Elastic** | **Actual** |
| 9,15% | 0,31% | 9,14% | **0,31%** | 18,42% | -7,25% | 8,26% | **7,21%** | 14,74% | **7,21%** | 9,14% | **0,64%** | 9,19% | **0,16%** |

*Table 10: Summary of results of Standard Deviation, Sharpe Ratio and positive months of the machine learning methods from July 2019 to June 2020*

| | **OLS** | **Actual OLS** | **Lasso** | **Actual Lasso** | **RF** | **Actual RF** | **DNN** | **Actual DNN** | **XG Boost** | **Actual XG Boost** | **Ridge** | **Actual Ridge** | **Elastic** | **Actual Elastic** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard Deviation | 0,37% | 25,46% | 0,32% | 25,46% | 0,57% | 29,35% | 0,69% | 24,82% | 8,69% | 24,82% | 0,36% | 25,53% | 0,35% | 25,68% |
| Sharpe Ratio | 24,42 | 0,01 | 28,39 | 0,01 | 32,27 | -0,25 | 11,97 | 0,29 | 1,70 | 0,29 | 25,45 | 0,02 | 26,60 | 0,01 |
| Positive Months | 100,00% | 50,00% | 100,00% | 50,00% | 100,00% | 66,67% | 100,00% | 58,33% | 66,67% | 58,33% | 100,00% | 50,00% | 100,00% | 50,00% |

Table 12 and Table 13 represents the final, and third predicted year (from July 2020 to June 2021). The details of the monthly returns of the Group 8 can be found in Appendix 3. All the machine learning methods outperformed for this last year. Indeed, the actual average annual returns are around 31%. OLS and shrinkage-based methods realized approximatively the same performance as they choose similar mutual funds. Apart from Deep Neural Network, all the six machine learning techniques underestimated the actual monthly returns. By comparing the Sharpe ratio, all machine learning methods realized around 2.30.

*Table 11: Summary of average annual return of the machine learning methods performance from July 2020 to June 2021*

| Average annual return | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OLS** | **Actual** | **Lasso** | **Actual** | **RF** | **Actual** | **DNN** | **Actual** | **XG Boost** | **Actual** | **Ridge** | **Actual** | **Elastic** | **Actual** |
| 9,14% | 31,95% | 9,13% | 31,95% | 18,88% | 28,61% | 50,28% | 30,74% | 14,77% | 26,33% | 9,13% | 31,90% | 9,18% | 31,96% |

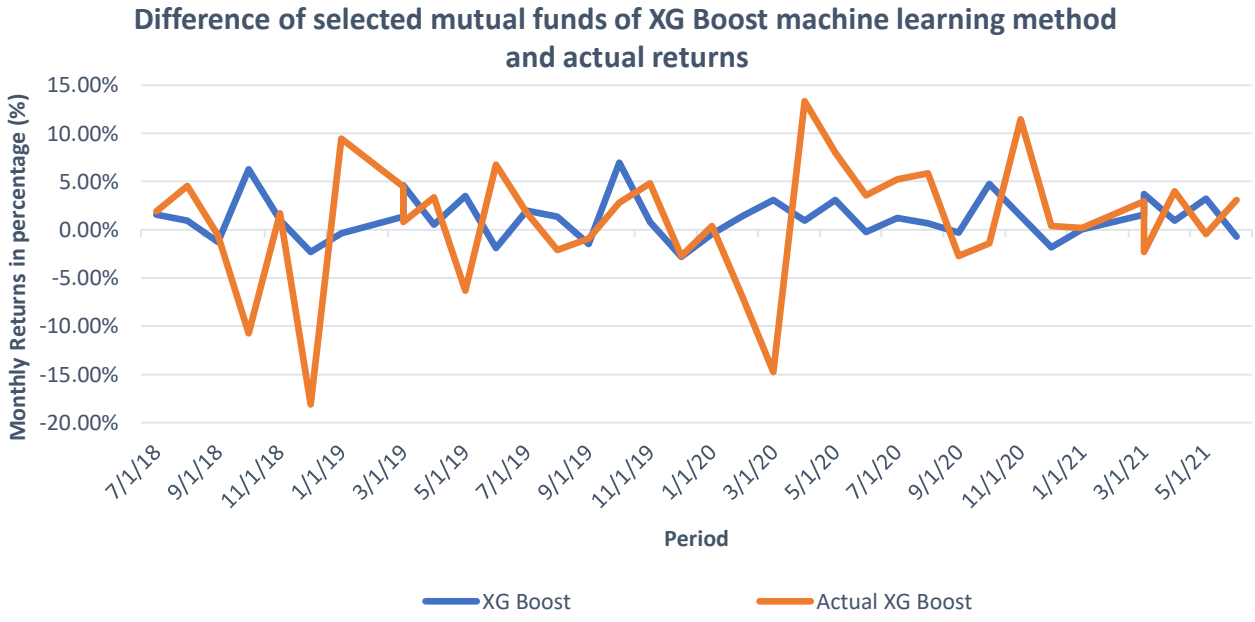
*Table 12: Summary of results of Standard Deviation, Sharpe Ratio and positive months of the machine learning methods from July 2020 to June 2021*

| | **OLS** | **Actual OLS** | **Lasso** | **Actual Lasso** | **RF** | **Actual RF** | **DNN** | **Actual DNN** | **XG Boost** | **Actual XG Boost** | **Ridge** | **Actual Ridge** | **Elastic** | **Actual Elastic** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard Deviation | 0,29% | 13,83% | 0,25% | 13,83% | 0,55% | 14,70% | 1,61% | 14,32% | 6,61% | 14,17% | 0,28% | 13,87% | 0,27% | 13,82% |
| Sharpe Ratio | 31,68 | 2,31 | 36,90 | 2,31 | 34,25 | 1,95 | 31,18 | 2,15 | 2,24 | 1,86 | 32,45 | 2,30 | 34,29 | 2,31 |
| Positive Months | 100,00% | 83,33% | 100,00% | 83,33% | 100,00% | 66,67% | 100,00% | 66,67% | 75,00% | 66,67% | 100,00% | 83,33% | 100,00% | 83,33% |

To sum up, in the first year, only Random Forest selected the suitable mutual funds to achieve superior returns in terms of average annual returns. While, in the second year, it is the turn of Deep Neural Network and XG boost to realize high average annual returns. In the third year, all machine learning techniques outperformed the selection strategy. However, as each year the selected mutual funds changed, the good results should be nuanced. Deep Neural Network varies 32 mutual funds from year 1 to year 2. If transaction costs were to be accounted for and high, the strategy would not have been profitable. Thus, the benefit of the machine learning algorithms could be adversely affected by the high turnover and transactions costs.

As seen in Figure 9, we can see that the XG boost try to follow the same trend as the actual returns. We reckon that XG Boost manages to forecast the movement one month ahead. For instance, in October 2019, XG Boost predicted 6.97%, while 4.84% were realized in November 2019. The same occurs in November 2020. In October 2020, XG Boost forecasted 4.74%, while 11.45% was achieved in November 2020 (Appendix 4).

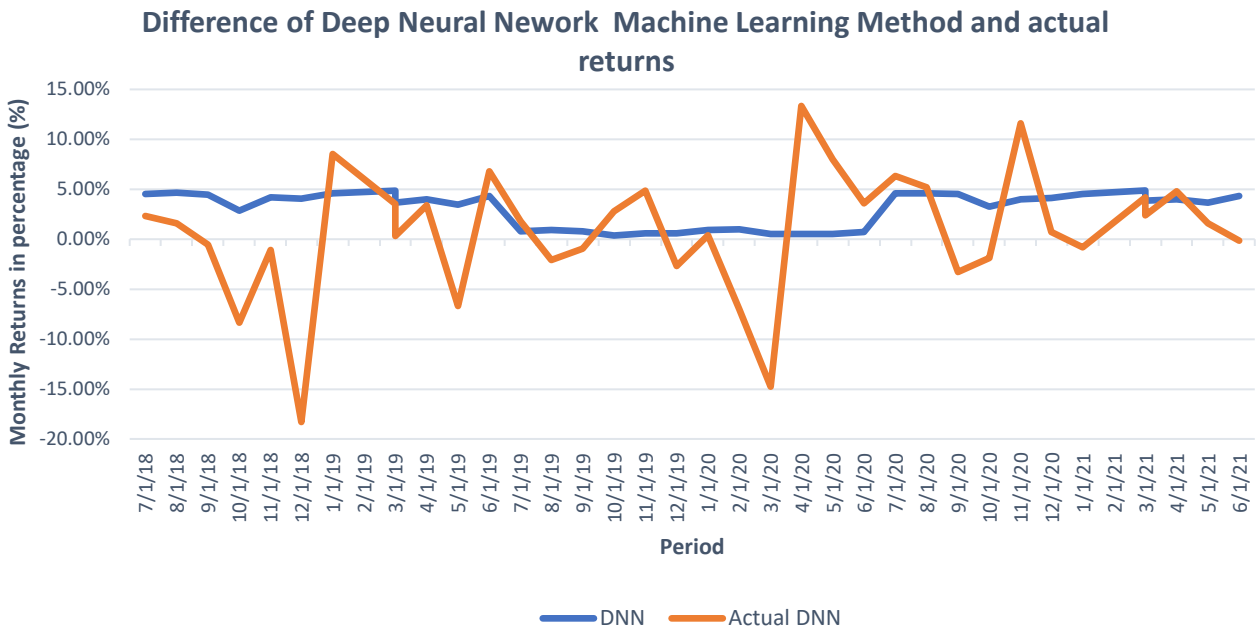
*Figure 9: Comparison of selected mutual funds (Group 8) of XG boost with actual returns on the whole period (July 2018 to June 2021).*




Conversely, in the Figure 10, Deep Neural Network didn't achieve the same performance as XG Boost. In Appendix 5, the plot of the comparison between OLS, Shrinkage-base methods and

Random Forest with actual returns is illustrated. These methods didn't behave similarly to the actual returns and thus failed to be accurate and a realistic method.

*Figure 10: Comparison of selected mutual funds (Group 8) of Deep Neural Network with actual returns on the whole period (July 2018 to June 2021).*




Finally, to complete our analysis on the performance of selecting the mutual funds, we computed the cumulative returns of each machine learning method. In appendix 6, the details of the cumulative returns are shown.

*Figure 11: Cumulative returns of the best mutual funds (group 8) depending on the different machine learning methods from July 2018 to June 2021.*

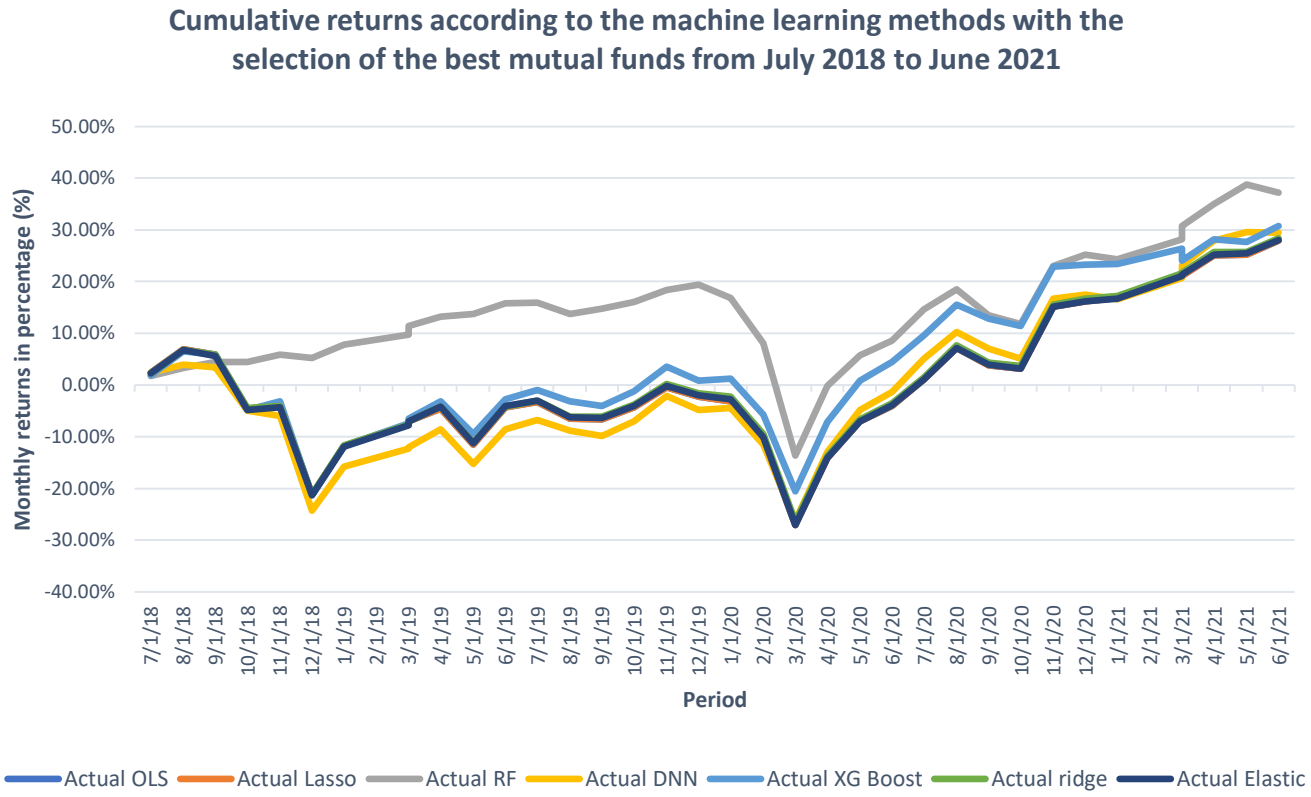



Figure 11 shows the evolution of the seven machine learning methods for selecting the best mutual funds based on predictions. We can conclude that Random Forest is the strategy that achieves the highest cumulative return, followed by Ridge and closely Deep Neural Network. The OLS, Shrinkage-based methods are in the last position. However, as mentioned before, XG boost is the method that is more realistic as it predicts the actual results accurately compared to the other methods. As illustrated in Table 11, the Random Forest has more volatility than XG Boost ( 29.35% in year 2 for Random Forest while 8.69% for XG Boost). The same occurs for Deep Neural Network (24.82% in year 2 for Deep Neural Network while 8.69% for XG Boost). Besides, contrary to Deep Neural Network, which keeps only four mutual funds of the 36 selected from year 1 to year 2, 13 mutual funds have remained with XG Boost. Thus, the high turnover suggesting high transaction costs could offset more than the gains from selling and buying new mutual funds. Thereby, it could reduce subsequent the performance of these machine learning methods. Yet, we

can mention that Random Forest achieved high returns in 2018 when the mutual had an average annual return equivalent to -4,94%. All other machine learning methods had on this period negative monthly returns, especially in December 2018. In March 2020, Random Forest had also the less profound drop contrary to the other machine learning methods. This could be related to the high turnover, which permits to switch the mutual funds that didn't perform well with others.

Random Forest was the machine learning method in terms of cumulative returns with the highest position and thus the best at selecting mutual funds. XG Boost reached the second position with 30.76%, followed closely by Deep Neural Network with 29.43%. In terms of average annual returns, Random Forest was the greatest method to achieve superior performance by picking mutual funds in the first year. In the second year, Deep Neural Network and XG Boost were the best ones. In the third year, all machine learning techniques had outstanding performance. However, Elastic was the best technique with a realization of 31.96%.

## 5.3 Analyse of Machine Learning methods on the five-factor model of Fama and French

The regression equation in Table 14 shows the results of regression testing of the Five-Factor Model as mutual funds excess return is the dependent variable, and independent variables are the Fama and French (2015) 5-factor (FF5) model augmented with momentum (MKT-RF, SMB, HML, CMA, RMW + MOM). According to Kenneth R.French Website[9], the Fama French 5 factors are the market, size (SMB – Small Minus Big), value (HML – High Minus Low), profitability (RMW – Robust Minimum Weak), investment (CMA – Conservative Minus Aggressive ) factors.

[9] Kenneth R. French (1995).Description of Fama/French 5 Factors (2x3. Fama French. Available on the website: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_5_factors_2x3.html. Consulted on the 13 August 2021.

*Table 13: Summary of raw alpha (intercept coefficient from the 36-month rolling window regression) of the best selected mutual funds (group 8) depending on the predicted year and machine learning technique*

| Alpha | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **OLS** | **Lasso** | **RF** | **DNN** | **XG Boost** | **Ridge** | **Elastic** |
| Year 1 | 0,19% | 0,21% | -0,21% | -0,23% | 0,31% | 0,19% | 0,19% |
| Year 2 | -0,14% | -0,13% | -0,67% | 0,11% | 0,11% | -0,14% | -0,14% |
| Year 3 | -0,65% | -0,65% | -1,18% | -1,12% | -0,45% | -0,63% | -0,63% |

According to Table 14, XG Boost has the highest alpha in year one and year 3 with 0.31% and -0.45%, respectively. In year 2, Deep Neural Network and XG boost have the highest alpha with 0.11%. For year 1, XG boost has an equation with a constant value of 0.31%, which significates that if all independent variables are considered constant then mutual funds excess return will increase by 0.31%.

A first glance at the factor loadings demonstrated in Table 15 reveals significant negative Market-rf loadings for all mutual funds techniques and predicting years, indicating the returns of funds being negatively driven by the market excess return. The factor loading of SMB indicate positive SMB factor loadings for all mutual funds techniques and predicting years, indicating the returns of funds being positively driven by the smaller cap mutual funds. The coefficient of SMB for Deep Neural Network in the third predicted year has a value of 1.02%. This shows that this variable has increased or decreased by 1 point. Thereby, the variable SMB will increase or reduce by 1.02 % points with the assumption that other variables remain constant. The fourth RMW factor also indicates positive RMW factor loadings for all mutual funds techniques and predicted years, indicating the funds' returns being positively associated with robust stocks. In the last predicted year, the third factor, HML, shows up significantly in all cases, while the sign of the coefficients is negative, indicating preferences for a low book to market mutual funds. Thus, the machine learning techniques will choose preferably low- book to market and small-cap mutual funds that are negatively associated with the value-weighted excess return on the market portfolio

*Table 14: Summary of raw betas (beta coefficient from the 36-month rolling window regression) of the best selected mutual funds (group 8) depending on the predicted year and machine learning techniques*

Beta's coefficient Market-rf

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic |
|---|---|---|---|---|---|---|---|
| Year 1 | -0,26% | -0,26% | -0,35% | -0,29% | -0,25% | -0,26% | -0,26% |
| Year 2 | -0,22% | -0,22% | -0,25% | -0,23% | -0,23% | -0,22% | -0,22% |
| Year 3 | -0,06% | -0,06% | -0,04% | -0,06% | -0,08% | -0,06% | -0,06% |

Beta's coefficient SMB

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic |
|---|---|---|---|---|---|---|---|
| Year 1 | 0,09% | 0,10% | 0,33% | 0,23% | 0,08% | 0,09% | 0,09% |
| Year 2 | 0,43% | 0,43% | 0,56% | 0,38% | 0,38% | 0,43% | 0,43% |
| Year 3 | 1,01% | 1,01% | 1,02% | 1,09% | 0,94% | 1,01% | 1,01% |

Beta's coefficient HML

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic |
|---|---|---|---|---|---|---|---|
| Year 1 | 0,04% | 0,05% | 0,09% | -0,07% | 0,05% | 0,04% | 0,04% |
| Year 2 | -0,40% | -0,40% | -0,45% | -0,34% | -0,34% | -0,40% | -0,40% |
| Year 3 | -0,87% | -0,87% | -0,93% | -0,91% | -0,85% | -0,86% | -0,86% |

Beta's coefficient RMW

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic |
|---|---|---|---|---|---|---|---|
| Year 1 | 0,31% | 0,31% | 0,61% | 0,39% | 0,23% | 0,30% | 0,30% |
| Year 2 | 0,46% | 0,44% | 0,68% | 0,33% | 0,33% | 0,45% | 0,45% |
| Year 3 | 1,23% | 1,23% | 1,16% | 1,44% | 1,28% | 1,23% | 1,23% |

Beta's coefficient CMA

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic |
|---|---|---|---|---|---|---|---|
| Year 1 | -0,23% | -0,23% | -0,42% | -0,18% | -0,20% | -0,23% | -0,23% |
| Year 2 | 0,65% | 0,65% | 0,59% | 0,58% | 0,58% | 0,65% | 0,65% |
| Year 3 | 1,95% | 1,95% | 2,05% | 2,01% | 1,87% | 1,94% | 1,94% |

Beta's coefficient MOM

| | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic |
|---|---|---|---|---|---|---|---|
| Year 1 | -0,16% | -0,16% | -0,13% | -0,16% | -0,14% | -0,16% | -0,16% |
| Year 2 | 0,07% | 0,07% | 0,13% | 0,04% | 0,04% | 0,07% | 0,07% |
| Year 3 | 0,20% | 0,20% | 0,21% | 0,24% | 0,15% | 0,20% | 0,20% |

For the conservative minus aggressive coefficient, the results show that during the first predicted year, all machine learning methods preferred the selection of mutual funds with an aggressive strategy, while in the second and third years, the conservative approach is preferred. Meaning, during a bad period and bearish market development, an aggressive approach will be chosen while a conservative one in a bullish market, characterized by an upward trend with high investors' confidence. This interpretation is also seen in the coefficient of momentum. According to the

Website of Fama French[10], the Momentum factor is the average return on the two high prior return portfolios minus the average return on the two low previous return portfolios. Thus, it reflects in the first year the negative rate of change in the price movement of mutual funds, which exhibit a bearish momentum. Conversely, in the second and third years, the coefficient Momentum (MOM) reveals a positive velocity of price changes in mutual funds representing a bullish momentum. This has been perceived in the analysis of the machine learning method for the selection of mutual funds. Indeed, in year 2, six machine learning methods achieve positive returns, and in year 3, all machine learning methods achieved positive returns. In opposite, the only Random Forest had managed to achieve positive returns. This reflects the potential of machine learning in selecting mutual funds at the appropriate time.

To summarize, in our data set, machine learning techniques will choose low-book preferably to market and small-cap mutual funds that are negatively associated with the value-weighted excess return on the market portfolio. We reckon the efficiency of changing strategy according to the mutual funds' returns. It will prefer a conservative strategy with an overall upward trend in broad market indices and an aggressive strategy during a downturn. Lastly, by analysing the Momentum factor, we can thus demonstrate during bad periods the bearish momentum and the bullish momentum during good periods. This could be permitted to investors the ability to securitize the movement and trend of the mutual funds.

## 5.4 Analyse of Machine Learning methods on accuracy

Table 16 reports the following measures: the Mean Square Error, the Root Mean Square Error, the R-Squared and the Mean Absolute Error. This was done on the mutual fund that had the highest returns from July 2018 to June 2021. Fidelity Advanced Growth Open-ended, a mutual fund with the abbreviation FAGOX US EQUITY, has the objective Growth Broad Market. According to Bloomberg classification[11], the portfolio is investing in companies that are reinvesting into expansion, acquisitions due to their above-average growth in earnings.

[10] Kenneth R. French (1995).Description of Fama/French 5 Factors (2x3. Fama French. Available on the website: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_5_factors_2x3.html. Consulted on the 13 August 2021.

[11] Source: Bloomberg (2013). Bloomberg Fund Classifications: Guide to our new funds Classification System. PDF Document available on the website : https://cpb-us-e1.wpmucdn.com/wp.txstate.edu/dist/2/916/files/2020/01/tmp14.pdf. Consulted on 13 August 2021

*Table 15: Summary of statistics error measures of the mutual fund FAGOX US EQUITY with the highest return on the whole period (July 2018 to June 2021)*

| Machine Learning methods | Mean squared error | Root mean square error | R Squared | Mean Absolute Error |
|---|---|---|---|---|
| OLS | 0,03% | 1,71% | 86,84% | 1,30% |
| LASSO | 0,03% | 1,64% | 86,83% | 1,24% |
| Random Forest | 0,03% | 1,75% | 86,99% | 1,42% |
| XG Boost | 0,00% | 0,52% | 98,88% | 0,37% |
| DNN | 0,20% | 4,44% | 20,02% | 3,42% |
| Ridge | 0,03% | 1,70% | 86,48% | 1,31% |
| Elastic Net | 0,03% | 1,67% | 86,97% | 1,27% |

XG Boost is the machine learning method that is the most performing in Mean Squared Error, Root Mean Square Error, R Squared and Mean Absolute Error presented in Table 16. The R squared of XG Boost has 98.88%, which is significant since closer the value of 1, the better the model will predict the mutual fund. For the Root Mean Square Error, XG boost achieves 0.52%, meaning that at 0.52% level of error, the system will emit in its predictions. According to the Mean Squared Error, the model also is at approximatively 0% far away from the predicted value from the true value. The Mean Absolute Error of XG Boost has an absolute difference from predicted values and an observed equivalent of 0.37%. Of all the seven machine learning methods, the most underperformed is the Deep Neural Network with an R Square of 20.02% and 4.44% for Root Mean Square Error.

Table 17 demonstrate the mutual fund Fidelity Sel Energy Service, which has the lowest return on the whole period in our set. This mutual fund invests with the objective of the Energy sector. Similar to before, XG Boost is the machine learning technique outperforming the others. Compared to all the seven machine learning methods, the Deep neural network performs the least on the measures demonstrated in Table 17.

*Table 16: Summary of statistics error measures of the mutual fund FSESX US Equity with the lowest return on the whole period (July 2018 to June 2021)*

| Machine Learning methods | Mean squared error | Root mean square error | R Squared | Mean Absolute Error |
|---|---|---|---|---|
| OLS | 0,76% | 8,73% | 21,81% | 6,16% |
| LASSO | 0,78% | 8,85% | 19,53% | 6,23% |
| Random Forest | 0,81% | 9,02% | 16,96% | 6,48% |
| XG Boost | 0,02% | 1,43% | 97,89% | 0,73% |
| DNN | 0,98% | 9,88% | 9,88% | 7,35% |
| Ridge | 0,81% | 8,97% | 17,51% | 6,39% |
| Elastic Net | 0,80% | 8,96% | 17,63% | 6,33% |

On the Mean Absolute Error values, apart from XG Boost and Deep Neural Network, the other machine learning techniques have an absolute difference from predicted values and are observed around 6.3%. The level of error of the model apart from XG Boost that will emit in its predictions is about 9% which is significant. Besides, these machine learning methods (OLS, Lasso, Random Forest, Deep Neural Network, Ridge and Elastic Net) are approximately 1% far away from the predicted values from the actual values. We can observe that depending on the returns of the mutual funds, the error measures of each machine learning method differentiate highly. When the mutual return is low, the machine learning methods tend to have bad results on accuracy measures such as Mean Squared Error, Root Mean Square Error, R Square and Mean Absolute Error.

To conclude, XG Boost was the technique standing out on the study of Mean Squared Error, Root Mean Square Error, R Squared and Mean Absolute Error. We reckon that these measures have been positively or negatively affected by the best and worst monthly mutual funds returns. Indeed, machine learning methods like Deep Neural Network had a drastic increase of errors with the worst monthly returns, for instance, in the mean absolute error (3.42% with the best return compared to 7.35% with the worst return). Thus, negative returns increase outliers and prediction errors which could adversely affect the model's performance on selecting mutual funds.

# 6. Discussion

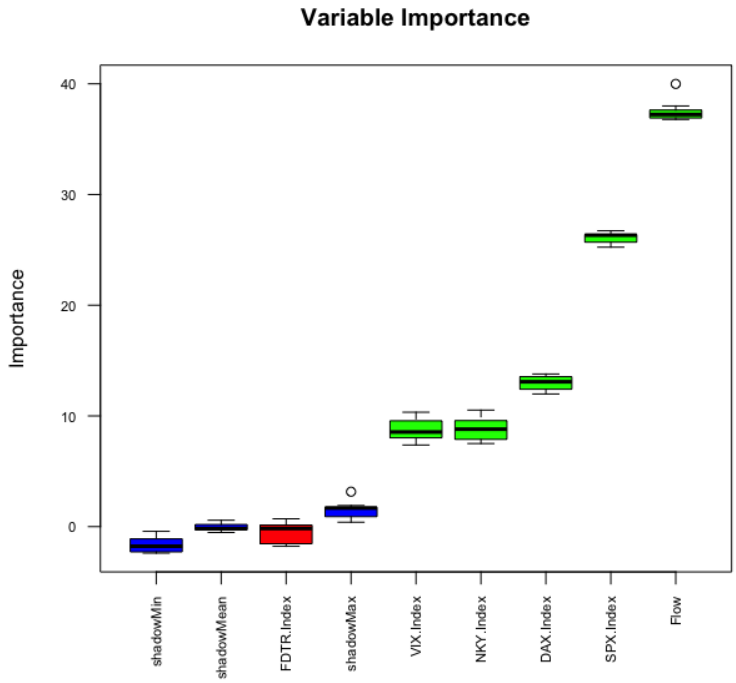
In this chapter, we want to focus on the main concerns when using machine learning models. Chapter 5 reckon that our seven machine learning methods can select a mutual fund and achieve superior performance, as we have seen with the Figure 11 of cumulative returns. However, as we have nuanced through the analysis, the results are not satisfactory to ultimately assume that machine learning algorithms are an accurate tool for mutual funds selection. Indeed, apart from the Random Forest, all six other machine learning methods failed to achieve positive average annual returns in the first year. On the contrary, in year 3, all methods achieved it and even outperformed the returns predicted. One of the explanations is that machine learning methods do have not the ability to predict unexpected events from historical data. Another explanation could be due to the feature chosen for the model to predict and select mutual funds.

The selection of features for the model is crucial as it impacts the prediction results and the selection of mutual funds. Empirical research on machine learning methods uses a different number of factors. Keywan Christian Rasekhschaffe & Robert C. Jones (2019) used 194 factors for selecting stocks from 1994 through 2016. However, interpretability is a primary concern as it the difficult to see which factors rely on the model. Besides, adding inputs that are indeed associated with the output leads to a reduced test error. However, adding noisy inputs that are not associated with the output leads to an increased test set error because of overfitting. This is the curse of dimensionality. Even if the inputs added are relevant, the variance incurred in fitting their coefficients may outweigh the reduction in bias that they bring. Thereby, choosing the features that will rely on the model is critical and needs to be carefully done.

In our set, according to the stepwise regression with the strategy of forward selection, the features that are the most important for predicting mutual funds returns are the SPX Index, Flow, NKY Index and the VIX Index. By doing the stepwise regression on Fama French, the following features are the most important: Alpha and Value added. These results were also found with the stepwise regression with stepwise selection strategy (strategy composed of forwarding selection and backward selection as explained in chapter 4). The variable Flow and SPX Index appear to be the essential characteristic for forecasting mutual fund returns with the multivariate adaptive regression

splines. Based on generalized cross-validation of 100 % (GCV), the variable flow occurs within the number of 4 subset models (n subsets). While the Generalized cross-validation of 13.5 % (GCV), the variable SPX Index occurs within the number of 3 subset models (n subsets). It means that the variable flow has more importance than the SPX Index variable, which is more important than all other variables. This is again seen with the plot of Boruta, as shown below in Figure 12. Indeed, the variable flow is vital. However, the flow variable possesses outliers which could increase the prediction error on the set and thus tarnish the relevance of the feature in the model. The FDTR index appears to be the least important in the prediction of mutual funds returns. As its importance was low, we could have removed it from our model.

*Figure 12: Summary of Boruta analysis on the importance of variables*




By computing the correlation between the features of our model, exhibiting in Table 18, the FDTR Index is the only variable that is positively associated with all other variables. Yet, the FDTR has a very low correlation with the other variables, even with the variable of the Flow or Median. It reflects the unimportance of this variable in the model again. The flow is only negatively correlated with the VIX Index supposing that the increase of the volatility is negatively impacting the flow of the mutual funds in our set and thus impacting the net asset value of the mutual funds. The same

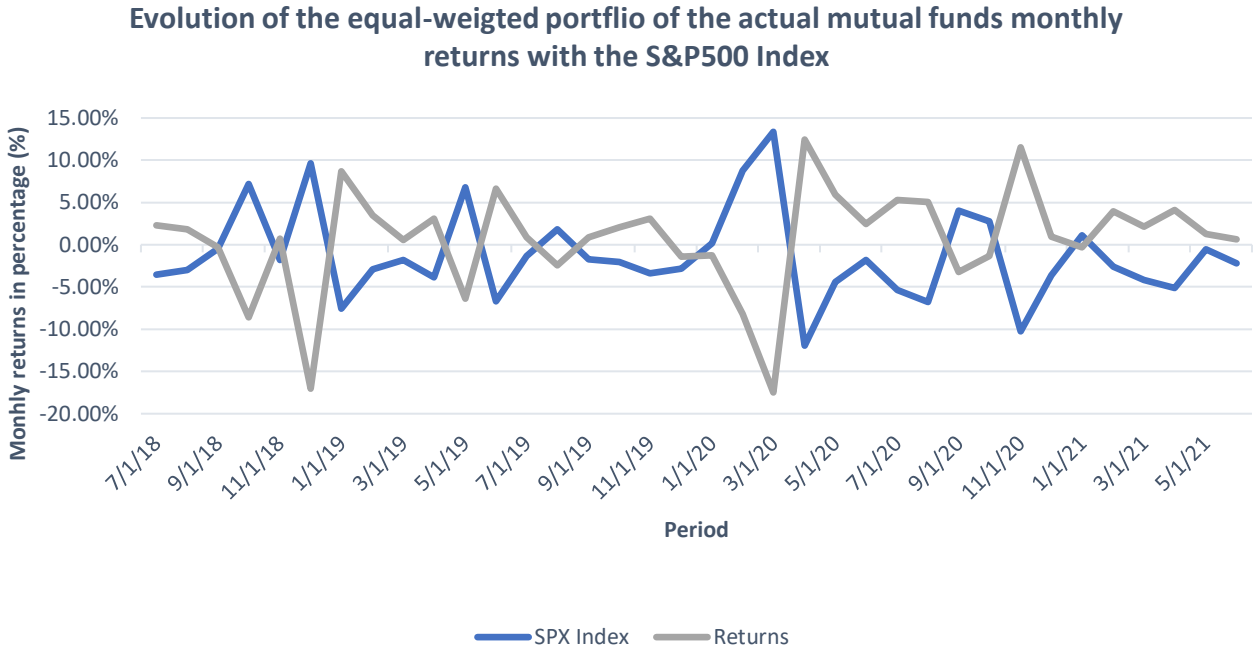
interpretation can be made for the median of the mutual funds in our set with the stock market's expectation of volatility based on the S&P 500 index. Hence, the correlation of the median on the VIX index is at -0.6463. Further, the DAX Index is highly correlated with the SPX Index with a correlation of 0.7327. The S&P 500 is significantly associated with our mutual funds' flow and median returns with 0.8315 and 0.9401, respectively.

*Table 17: Summary of correlation between features*

| | VIX.Index | SPX.Index | NKY.Index | DAX.Index | FDTR.Index | Flow | Median |
|---|---|---|---|---|---|---|---|
| **VIX.Index** | 1,0000 | −0,6368 | −0,3694 | −0,4984 | 0,0509 | −0,5916 | −0,6463 |
| **SPX.Index** | −0,6368 | 1,0000 | 0,5239 | 0,7327 | 0,0494 | 0,8315 | 0,9401 |
| **NKY.Index** | −0,3694 | 0,5239 | 1,0000 | 0,4968 | 0,0647 | 0,4775 | 0,5614 |
| **DAX.Index** | −0,4984 | 0,7327 | 0,4968 | 1,0000 | 0,0520 | 0,6275 | 0,7139 |
| **FDTR.Index** | 0,0509 | 0,0494 | 0,0647 | 0,0520 | 1,0000 | 0,0213 | 0,0309 |
| **Flow** | −0,5916 | 0,8315 | 0,4775 | 0,6275 | 0,0213 | 1,0000 | 0,9054 |
| **Median** | -0,6463 | 0,9401 | 0,5614 | 0,7139 | 0,0309 | 0,9054 | 1,0000 |

In opposite, by plotting the evolution of the equal-weighted portfolio of the actual funds' monthly returns with the S&P 500 Index, we remark a negative association between them on the whole period from July 2018 to June 2021. Indeed, as soon, the S&P500 Index increase, the actual returns of the mutual fund's decrease. For instance, in March 2020, the S&P 500 achieved 13.37%, while mutual funds had a drop attaining -17.49%. In April 2020, mutual funds have an increase realized a monthly return of 12.45%, while the S&P 500 realized a negative performance with a -11.94% of return on this month. In Appendix 7, the details of the equal-weighted portfolio of the 288 mutual funds are described as well as the difference of annual statistics between the S&P 500 and this equal-weighted portfolio. This negative correlation can be interpreted as the willingness of mutual funds to hedge the equity risk. Thus, as the mutual funds have the asset-class focus of equity, ideally, to hedge the equity risk of the mutual funds, the mutual funds need to have a negative correlation with the S&P 500 Index as possible.

*Figure 13: Evolution of the actual monthly returns of the equal-weighted portfolio composed of 288 mutual funds with the S&P 500 index from July 2018 to June 2021.*

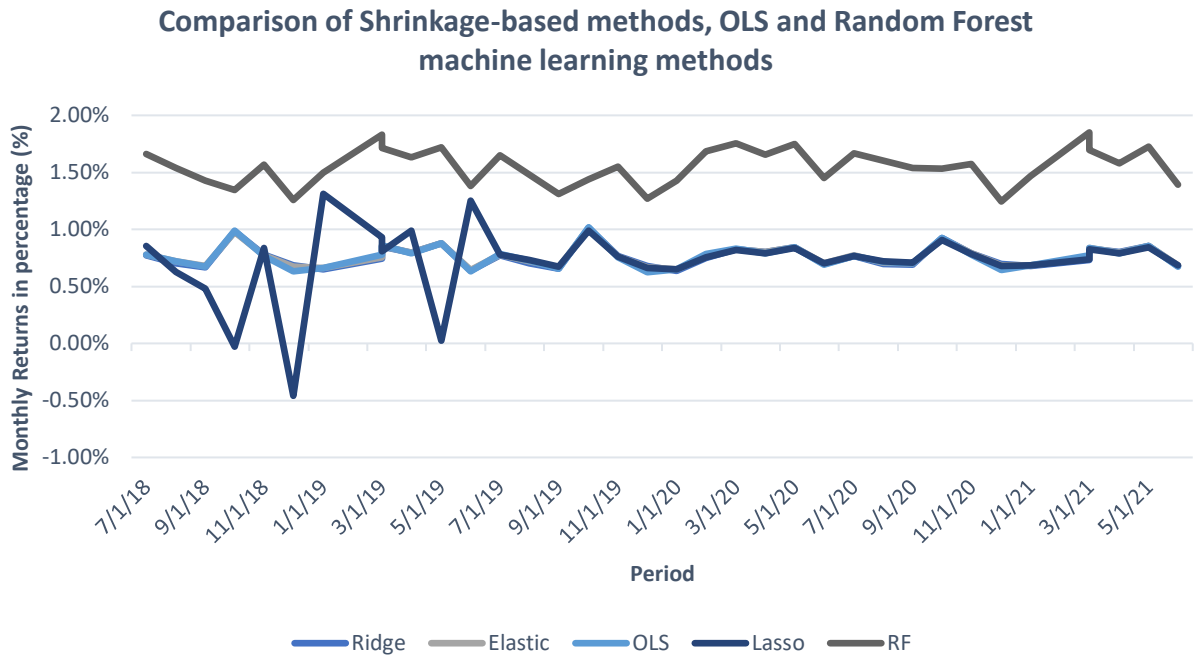



Machine Learning models are very good at picking patterns from the data. Even if a small part of the data leaks information about the future, the model will most likely find it and exploit it. Thus, the quality of data is important, but most of all, the machine learning technique chosen can also have an impact. We have seen that specific machine learning methods realize superior performance in our sample. This is the case for Deep Neural Network, Random Forest and XG Boost. With the equally- weighted portfolio, XG Boost followed slighted the reality, characterized by the number of positive months and the drawdown period. Random Forest was the technique that achieves the highest cumulative return in selecting the portfolio (group 8) and performs distinctively during bad periods compared to the other methods. Deep Neural Network also achieved a high score on the cumulative returns by selecting the portfolio group 8 with a third out of seven. The shrinkage-based method and OLS select almost the same mutual funds and follow the same path (Figure 14).

Thus, using these methods simultaneously will not help to advise on the proper selection of mutual funds. One method of this kind could have been enough in our model. As Figure 14 is illustrating, Random Forest also closely follows the movement and velocity of OLS and Shrinkage-based methods. Indeed, we can perceive that in January 2020, OLS and the Shrinkage-based methods had

a decline monthly return of around -0.84%, followed by another drop of around -7.23% in February 2020 and -16.9% in March 2020. In April 2020, these methods had a positive monthly return of 12.8%. This drift was similar to the Random Forest tendency (-8.73% in February 2020, -21.72% in March 2020 and 13.47% in April 2020). However, it was on higher levels, meaning the monthly returns declined steeper, and the increase of monthly returns was higher.

*Figure 14: Comparison of Shrinkage-based methods, OLS and Random Forest machine learning methods*




Nevertheless, the turnover for Deep Neural Network and Random Forest is slightly higher than the XG boost, implying high transaction costs. However, OLS and shrinkage-based methods had a much lower turnover ratio which could be a more feasible investment strategy in transaction costs. Further, predicted returns, for instance, in year 1 with the Deep Neural Network technique, were highly overestimated, which could impact the actual results if this particular stock's investment had been made. Indeed, Deep Neural Network predicted 49.57%, while the true result was at -8.51% average annual return. This significant difference, equivalent to 41.06%, could have repercussions on the investment profit tremendously.

## 7. Conclusion

According to a survey of The Bank of England (BoE) and Financial Conduct Authority (FCA) on the utilization of machine learning by financial institutions in the United Kingdom, they discover that machine learning is increasingly being adopted in the United Kingdom financial services (Bank of England, Financial Conduct Authority, 2019). Of 106 respondents, 66.67% report they already use machine learning daily in their business. The primary usage of machine learning lies within the support of client interaction, business decisions, investment strategy or transactions.
In this thesis, we studied machine learning as a tool for investment strategy by predicting and selecting mutual funds. Our sample was open-ended mutual funds with the asset-class focus of equity from 1990 until 2021. Using a training window of 28.5 years, seven machine learning methods predicted and selected the best mutual funds for three years. Our model's seven machine learning methods were OLS, Lasso, Ridge, Elastic, Random Forest, Deep Neural Network and XG boost.

On investing at a weight of 0.35% with an equally-weight portfolio, from July 2018 to June 2021, XG Boost was the machine learning which realized the best performance in terms of realistically of the drawdown period and number of positive months. Yet, Deep Neural Network receives the highest cumulative returns with 30.97% from July 2018 to June 2021. However, Random Forest predicted the closest with 20.14% from the actual cumulative returns equivalent to 21.94%.

On the selection strategy, according to the predicting years, the results were contrasting. In the first year, Random Forest achieves superior returns in terms of average annual returns. In opposite, in the second year, Deep Neural Network and XG boost realize the highest average annual returns. In the third year, all machine learning techniques outperformed the selection strategy realizing around 30.49% on average. Nonetheless, the Elastic machine learning method reached the highest average annual return with 31.94%. In terms of cumulative returns, Random Forest gains the first position with 37.21%, followed by XG Boost with 30.76%, and closely by Deep Neural Network with 29.43%.

By regressing the mutual funds' returns with the Fama French data, we reckon that these seven machine learning techniques are more likely to choose low-book to market and small-cap mutual funds. It will also select mutual funds that are negatively associated with the value-weighted excess return on the market portfolio. The efficiency of machine learning was also perceived on the ability to change the strategy according to the market. It will prefer a conservative strategy with an overall upward trend in broad market indices and an aggressive strategy during a downturn. Moreover, measures like the Mean Squared Error, Root Mean Square Error, R Squared and Mean Absolute Error are positively or negatively affected according to the monthly returns of mutual funds.

Yet, the performance of these seven machine learning methods should be nuanced as they had high turnover, which could affect the performance when taking into account transactions costs. Besides, negative returns increase outliers and prediction errors, affecting the model's performance in selecting mutual funds. Thus, this thesis has proven that machine learning methods are effective yet not perfect. The application of machine learning methods has the potential to improve outcomes for both businesses and consumers in the financial services industry. However, in terms of selecting mutual funds, machine learning could be a tool to verify theories or ideas. It shouldn't be taken for granted. It should be questioned, asked if it is a realistic and profitable investment strategy. Nonetheless, the promise of machine learning to make financial services and markets more efficient raises more broadly profound questions around the use of data, the complexity of techniques and the automation of processes, systems and decision-making.

## 8. Bibliography

Ahmad, Yaqoob; Sun, Guangguo; Khan, Wahab (2018) The Impact Of Fund's Characteristics On Performance: Evidence From Equity Mutual Funds In Pakistan. City University Research Journal; Peshawar Vol. 8, N° 2, (Jul 2018): 209-218. https://www.proquest.com/openview/8e3e3e0c1b61517ec26456103408389c/1?pq-origsite=gscholar&cbl=2068970

Alberg, John & Lipton, Zachary. (2017). Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals. Euclidean Technologies and Carnegie Mellon University. https://www.researchgate.net/publication/321095166_Improving_Factor-Based_Quantitative_Investing_by_Forecasting_Company_Fundamentals

Amihud, Y., & Goyenko, R. (2012). Mutual Fund's R2 as Predictor of Performance. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.1319786

Arik Ben Dor & Ravi Jagannathan & Iwan Meier, 2005. "Understanding Mutual Fund And Hedge Fund Styles Using Return-Based Style Analysis," World Scientific Book Chapters, in: H Gifford Fong (ed.), The World Of Hedge Funds Characteristics and Analysis, chapter 4, pages 63-108, World Scientific Publishing Co. Pte. Ltd. https://ideas.repec.org/h/wsi/wschap/9789812569448_0004.html

Babalos, V., Mamatzakis, E. C., & Matousek, R. (2015). The performance of US equity mutual funds. Journal of Banking & Finance, 52, 217–229. https://doi.org/10.1016/j.jbankfin.2014.12.008

Babbar, S., & Sehgal, S. (2018). Mutual Fund Characteristics and Investment Performance in India. Management and Labour Studies, 43(1–2), 1–30. https://doi.org/10.1177/0258042x17745183

Bank of England, Financial Conduct Authority (FCA).(2019).Machine learning in UK financial services.Joint suvey published as PDF files, as of published date 16.10.2019, accessed on 13.08.2021

Barber, B. M., Huang, X., & Odean, T. (2016). Which Factors Matter to Investors? Evidence from Mutual Fund Flows. Review of Financial Studies, 29(10), 2600–2642. https://doi.org/10.1093/rfs/hhw054

Barber, B., Odean, T., & Zheng, L. (2005). Out of Sight, Out of Mind: The Effects of Expenses on Mutual Fund Flows*. The Journal of Business, 78(6), 2095–2120. https://doi.org/10.1086/497042

Berk, J. B., & Green, R. C. (2002). Mutual Fund Flows and Performance in Rational Markets. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.383061

Berk, J. B., & Green, R. C. (2002b). Mutual Fund Flows and Performance in Rational Markets. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.383061

Berk, J., & Green, R. (2004). Mutual Fund Flows and Performance in Rational Markets. Journal of Political Economy, 112(6), 1269–1295. https://doi.org/10.1086/424739

Berk, Jonathan & van Binsbergen, Jules. (2012). Measuring Skill in the Mutual Fund Industry. Journal of Financial Economics. 118. 10.2139/ssrn.2038108.

Berk, J. B., & van Binsbergen, J. H. (2015). Measuring skill in the mutual fund industry. *Journal of Financial Economics*, *118*(1), 1–20. https://doi.org/10.1016/j.jfineco.2015.05.002

Berkowitz, J. P., & Schorno, P. J. (2016). Characteristics of Mutual Funds with Extreme Performance. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2784403

Berkowitz, J. P., Schorno, P. J., & Shapiro, D. A. (2017). Characteristics of mutual funds with extreme performance. Review of Financial Economics, 34, 50–60. https://doi.org/10.1016/j.rfe.2017.04.003

Bers, Martina & Madura, Jeff. (2000). Why Does Performance Persistence Vary Among Closed-End Funds?. Journal of Financial Services Research. 17. 127-147. 10.1023/A:1008110529949.

Białkowski Jędrzej, Roger Otten, (2011) Emerging market mutual fund performance: Evidence for Poland,The North American Journal of Economics and Finance,Volume 22, Issue 2,Pages 118-130,ISSN 1062-9408,https://doi.org/10.1016/j.najef.2010.11.001.

Bollen, Nicolas & Busse, Jeffrey. (2005). Short-Term Persistence in Mutual Fund Performance. Review of Financial Studies. 18. 10.1093/rfs/hhi007.

Breiman, L. (2001) Random Forests, *Machine Learning* 45, 5–32. Also see the bibliography at: http://www.stat.berkeley.edu/˜breiman/RandomForests/cc papers.htm

Busse, J. A., Chordia, T., Jiang, L., & Tang, Y. (2014). How Does Size Affect Mutual Fund Performance? Evidence from Mutual Fund Trades. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2350583

Carhart Mark M. (1997) On Persistence in Mutual Fund Performance. The Journal of Finance, Vol. 52, No. 1. (Mar., 1997), pp. 57-82. http://links.jstor.org/sici?sici=0022-1082%28199703%2952%3A1%3C57%3AOPIMFP%3E2.0.CO%3B2-G

Chen, J. S., Hong, H. G., Huang, M., & Kubik, J. D. (2004). Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.372721

Chiang, W. C., Urban, T., & Baldridge, G. (1996). A neural network approach to mutual fund net asset value forecasting. Omega, 24(2), 205–215. https://doi.org/10.1016/0305-0483(95)00059-3

Choi, H. S. (2015). Seasonality In Mutual Fund Flows. Journal of Applied Business Research (JABR), 31(2), 715. https://doi.org/10.19030/jabr.v31i2.9162

Chou, W. H., & Hardin, W. G. (2014). Performance chasing, fund flows and fund size in real estate mutual funds. Journal of Real Estate Finance and Economics, 49(3), 379–412. https://doi.org/10.1007/s11146-013-9444-x

Cuthbertson, K., Nitzsche, D. and O'Sullivan, N. (2010), Mutual Fund Performance: Measurement and Evidence1. Financial Markets, Institutions & Instruments, 19 : 95-187. https://doi.org/10.1111/j.1468-0416.2010.00156.x

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & A. P. Santos, A. (2021). Can Machine Learning Help to Select Portfolios of Mutual Funds? Version March 24, 2021. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3768753

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & A. P. Santos, A. (2021). Can Machine Learning Help to Select Portfolios of Mutual Funds? Version June 3, 2021. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3768753

Dhagash Mehta and Dhruv Desai and Jithin Pradeep (2020), Machine Learning Fund Categorizations. The Vanguard Group Malvern, PA, USA. https://arxiv.org/abs/2006.00123

Dhume, Pournima. (2014). Fund Size & its Impact on Fund Performance: An Empirical Evidence from Selected Indian Mutual Fund Companies. Department of Commerce, Goa University. https://www.researchgate.net/publication/299452742_Fund_Size_its_Impact_on_Fund_Performance_An_Empirical_Evidence_from_Selected_Indian_Mutual_Fund_Companies

Ding, H., Zheng, H., & Zhu, C. (2015). Equity funds in emerging Asia: Does size matter? International Review of Economics & Finance, 35, 149–165. https://doi.org/10.1016/j.iref.2014.09.012

El Mohandiz, Mehdi. Fund Performance and Fund Characteristics: Evidence from Belgian Equity Mutual Funds. Louvain School of Management, Université catholique de Louvain, 2019. Prom. : Grégoire, Philippe. http://hdl.handle.net/2078.1/thesis:20235

Droms, W. G., & Walker, D. A. (2001). Persistence of mutual fund operating characteristics: returns, turnover rates, and expense ratios. Applied Financial Economics, 11(4), 457–466. https://doi.org/10.1080/096031001300314018

Elton, e. j., Gruber, m. j., & Blake, c. r. (1995). Fundamental Economic Variables, Expected Returns, and Bond Fund Performance. The Journal of Finance, 50(4), 1229–1256. https://doi.org/10.1111/j.1540-6261.1995.tb04056.x

El Naqa I., Murphy M.J. (2015) What Is Machine Learning?. In: El Naqa I., Li R., Murphy M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1

Fabregat-Aibar, L., Sorrosal-Forradellas, M. T., Barberà-Mariné, G., & Terceño, A. (2021). Can Artificial Neural Networks Predict the Survival Capacity of Mutual Funds? Evidence from Spain. Mathematics, 9(6), 695. https://doi.org/10.3390/math9060695

Fama, E.F. & French, K.R. (1992). The cross-section of expected returns. Journal of Finance 47(2): 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (2015). Choosing Factors. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2668236

Ferreira, M. A., Keswani, A., Miguel, A. F., & Ramos, S. B. (2012). The flow-performance relationship around the world. Journal of Banking & Finance, 36(6), 1759–1780. https://doi.org/10.1016/j.jbankfin.2012.01.019

Friesen, G. C., & Sapp, T. R. (2007). Mutual fund flows and investor returns: An empirical examination of fund investor timing ability. Journal of Banking & Finance, 31(9), 2796–2816. https://doi.org/10.1016/j.jbankfin.2007.01.024

Fu XingYu, JinHong Du, YiFeng Guo, MingWen Liu, Tao Dong, XiuWen Duan(2018). A Machine Learning Framework for Stock Selection. ArXiv.Org. https://arxiv.org/abs/1806.01743v2. Consulted on 13 August 2021

Gamble, J. (2020). Can Expense Ratios Signal Performance? An Analysis of Equity ETFs & Mutual Funds. American Journal of Undergraduate Research, 16(4), 23–40. https://doi.org/10.33697/ajur.2020.004

Gil-Bazo, Javier & Ruiz-Verdú, Pablo. (2009). The Relation Between Price and Performance in the Mutual Fund Industry. Journal of Finance. 64. 2153-2183. 10.2139/ssrn.947448.

Gilberto Batres-Estrada (2015) Deep Learning for Multivariate Financial Time Series. KTH Stockholm, Sweden. https://www.math.kth.se/matstat/seminarier/reports/M-exjobb15/150612a.pdf

Goldstein, I., Jiang, H., & Ng, D. T. (2017). Investor flows and fragility in corporate bond funds. Journal of Financial Economics, 126(3), 592–613. https://doi.org/10.1016/j.jfineco.2016.11.007

Grinblatt, M., & Titman, S. (1989). Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings. The Journal of Business, 62(3), 393. https://doi.org/10.1086/296468

Grinblatt, M., & Titman, S. (1992). The Persistence of Mutual Fund Performance. The Journal of Finance, 47(5), 1977–1984. https://doi.org/10.1111/j.1540-6261.1992.tb04692.x

Grinblatt, M., & Titman, S. (1993). Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns. The Journal of Business, 66(1), 47. https://doi.org/10.1086/296593

Grinblatt, M., & Titman, S. (1994). A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques. The Journal of Financial and Quantitative Analysis, 29(3), 419. https://doi.org/10.2307/2331338

Gruber, m. j. (1996). Another Puzzle: The Growth in Actively Managed Mutual Funds. The Journal of Finance, 51(3), 783–810. https://doi.org/10.1111/j.1540-6261.1996.tb02707.x

Gu, S., Kelly, B. T., & Xiu, D. (2019). Empirical Asset Pricing Via Machine Learning. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3159577

Harris, M., & Raviv, A. (1993). Differences of Opinion Make a Horse Race. Review of Financial Studies, 6(3), 473–506. https://doi.org/10.1093/rfs/5.3.473

Hastie, T. (2021). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics) 2nd ed. 2009. Corr. 3rd printing 5th Printing. edition. Springer.

Heaton, J., & Polson, N. (2016). Deep Learning for Finance: Deep Portfolios. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2838013

Hendricks, d., Patel, j., & Zeckhauser, r. (1993). Hot Hands in Mutual Funds: Short-Run Persistence of Relative Performance, 1974–1988. The Journal of Finance, 48(1), 93–130. https://doi.org/10.1111/j.1540-6261.1993.tb04703.x

Hunter, D., E. Kandel, S. Kandel, and R. Wermers (2014). Mutual fund performance evaluation with active peer benchmarks. Journal of Financial Economics 112(1), 1–29. 10.1016/j.jfineco.2013.12.006

Guo, H. (2004). Stock prices, firm size, and changes in the federal funds rate target. The Quarterly Review of Economics and Finance, 44(4), 487–507. https://doi.org/10.1016/j.qref.2003.11.001
ICI (2020) A Review of Trends and Activities in the Investment Company Industry. 60th Edition Investment Company Fact Book, Investment Company Institute.
https://www.ici.org/system/files/attachments/pdf/2020_factbook.pdf

Indro, Daniel & Jiang, Christine & Hu, Michael & Lee, Wayne. (1999). Mutual Fund Performance: Does Fund Size Matter?. Financial Analysts Journal. 55. 74-87. 10.2469/faj.v55.n3.2274.

J. Edward Graham, Carlos Lassala & Belén Ribeiro Navarrete (2020) Influences on mutual fund performance: comparing US and Europe using qualitative comparative analysis, Economic Research-Ekonomska Istraživanja, 33:1, 3049-3070, DOI: 10.1080/1331677X.2019.1696695

Jeffrey A. Busse  Tarun Chordia  Lei Jiang  Yuehua Tang (2016) Mutual Fund Transaction Costs.
https://www.ilpost.it/wp-content/uploads/2016/12/mutual-fund-transaction-costs-SSRN-id2350583-1.pdf

Jensen, M. (1968). The Performance of Mutual Funds in the Period 1945-1964. The Journal of Finance, 23(2), 389-416. doi:10.2307/2325404

Friedman Jerome H. (1991) Multivariate Adaptive Regression Splines. The Annals of Statistics , Mar., 1991, Vol. 19, No. 1 (Mar., 1991), pp. 1-67. Institute of Mathematical Statistics. https://www.jstor.org/stable/2241837

Jiaqi Chen & Michael Tindall & Wenbo Wu, 2016. "Hedge Fund Return Prediction and Fund Selection: A Machine-Learning Approach," Occasional Papers 16-4, Federal Reserve Bank of Dallas. https://ideas.repec.org/p/fip/feddop/2016_004.html

José Diogo Seca (2019) Forecasting stock trends through Machine Learning. Faculdade de engenharia da universidade do Porto. https://repositorio-aberto.up.pt/bitstream/10216/121271/2/343586.pdf

Kaminsky G., Lyons R., Schmukler S. (2001) Mutual Fund Investment in Emerging Markets: An Overview. In: Claessens S., Forbes K.J. (eds) International Financial Contagion. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-3314-3_7

Kandel, E., & Pearson, N. D. (1995). Differential Interpretation of Public Signals and Trade in Speculative Markets. Journal of Political Economy, 103(4), 831–872. https://doi.org/10.1086/262005

Karpoff, J. M. (1987). The Relation Between Price Changes and Trading Volume: A Survey. The Journal of Financial and Quantitative Analysis, 22(1), 109. https://doi.org/10.2307/2330874

Kaur Inderjit (2018) Effect of mutual funds characteristics on their performance and trading strategy: A dynamic panel approach, Cogent Economics & Finance, 6:1, 1493019, DOI: 10.1080/23322039.2018.1493019

Keim, D., & Madhavan, A. (1997). Transactions costs and investment style: an inter-exchange analysis of institutional equity trades. Journal of Financial Economics, 46(3), 265–292. https://doi.org/10.1016/s0304-405x(97)00031-7

Keywan Christian Rasekhschaffe & Robert C. Jones (2019) Machine Learning for Stock Selection, Financial Analysts Journal, 75:3, 70-88, DOI: 10.1080/0015198X.2019.1596678

Kursa, Miron & Jankowski, Aleksander & Rudnicki, Witold. (2010). Boruta - A System for Feature Selection. Fundamenta. Informaticae. 101. 271-285. 10.3233/FI-2010-288.

Li, B., & Rossi, A. G. (2021). Selecting Mutual Funds from the Stocks They Hold: A Machine Learning Approach. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3737667

Linnainmaa, J. T., & Roberts, M. R. (2016). The History of the Cross Section of Stock Returns. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2897719

Livingston, M., Yao, P., & Zhou, L. (2019). The volatility of mutual fund performance. Journal of Economics and Business, 104, 105835. https://doi.org/10.1016/j.jeconbus.2019.02.001

MA, Linlin; Yuehua Tang; and GOMEZ, Juan-Pedro (2016) Portfolio manager compensation and mutual fund performance. Finance Down Under 2014 Building on the Best from the Cellars of Finance, Melbourne, Australia, 2016 May 13. Research Collection Lee Kong Chian School Of Business. Available at: https://ink.library.smu.edu.sg/lkcsb_research/5169

Malaquias Rodrigo Fernandes, Lucas Alves Duarte de Sarvas & Lynnea Naves Souza Oliveira (2016) Size and Performance: Evidence on Brazilian Multimarket Funds. Global Journal of Management and Business Research: C Finance , Volume 16 Issue 9 Version 1.0 Year 2016. Publisher: Global Journals Inc. (USA). https://globaljournals.org/item/6697-size-and-performance-evidence-on-brazilian-multimarket-funds

Malkiel, B. G. (1995). Returns from Investing in Equity Mutual Funds 1971 to 1991. The Journal of Finance, 50(2), 549–572. https://doi.org/10.1111/j.1540-6261.1995.tb04795.x

Mamaysky, Harry & Spiegel, Matthew & Zhang, Hong. (2006). Improved Forecasting of Mutual Fund Alphas and Betas. Review of Finance. 11. 359-400. 10.1093/rof/rfm018.

Mark Grinblatt and Sheridan Titman (1992) The Persistence of Mutual Fund Performance. The Journal of Finance , Dec., 1992, Vol. 47, No. 5 (Dec., 1992), pp. 1977-1984. Wiley for the American Finance Association. https://www.jstor.org/stable/2329005

Mark Grinblatt, Sheridan Titman and Russ Wermers (1995) Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior.

Masaya Abe, Hideki Nakayama (2018) Deep Learning for Forecasting Stock Returns in the Cross-Section. Nomura Asset Management Co., Ltd., Tokyo, Japan and The University of Tokyo, Tokyo, Japan , https://arxiv.org/abs/1801.01777

Mateus, Irina & Mateus, Cesario & Todorovic, Natasa. (2017). The Impact of Benchmark Choice on US Mutual Fund Benchmark-Adjusted Performance and Ranking. SSRN Electronic Journal. 10.2139/ssrn.3014010.

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2020). What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. Journal of Theoretical and Applied Electronic Commerce Research, 16(3), 466–490. https://doi.org/10.3390/jtaer16030029

Miles Jeremy (2014) R Squared, Adjusted R Squared. Encyclopedia of Statistics in Behavioral Science, https://doi.org/10.1002/9781118445112.stat06627

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2740751

Murcia, I. (2011). Spanish mutual fund performance: An analysis of the determinants (Working Paper No. 48). Spain: National Securities Market Commission. https://www.cnmv.es/DocPortal/Publicaciones/MONOGRAFIAS/DT48_weben.pdf

Odean, T. (1998). Volume, Volatility, Price, and Profit When All Traders Are Above Average. The Journal of Finance, 53(6), 1887–1934. https://doi.org/10.1111/0022-1082.00078

Otten, R., & Bams, D. (2002). European Mutual Fund Performance. European Financial Management, 8(1), 75–101. https://doi.org/10.1111/1468-036x.00177

Peterson, James & Petranico, Paul & Riepe, Mark & Xu, Fran. (2001). Explaining the Performance of Domestic Equity Mutual Funds. The Journal of Investing. 10. 81-91. 10.3905/joi.2001.319477

Porter, G., & Trifts, J. (1998). Performance persistence of experienced mutual fund managers. Financial Services Review, 7(1), 57–68. https://doi.org/10.1016/s1057-0810(99)80013-4

Prather, L., Bertin, W. J., & Henker, T. (2004). Mutual fund characteristics, managerial attributes, and fund performance. Review of Financial Economics, 13(4), 305–326. https://doi.org/10.1016/j.rfe.2003.11.002

Raphaëlle BELLANDO & Laura-Dona CAPOTA & Sébastien GALANTI (2019). "The impact of return shocks on mutual funds' flows: an empirical study of French bond mutual funds," LEO Working Papers / DR

LEO 2730, Economic Research Department of the University of Orleans (LEO° / Laboratoire d'Economie d'Orleans (LEO), University of Orleans.

Sammut, C., Webb G.I. (2011). “Mean Squared Error” in the Encyclopedia of Machine Learning and Data Mining. Boston, Springer.

Schapire, R. E. and Y. Freund (2012). Boosting: Foundations and Algorithms. Kybernetes, Vol. 42 No. 1, pp. 164-166. https://doi.org/10.1108/03684921311295547

Sirri, E. R., & Tufano, P. (1998). Costly Search and Mutual Fund Flows. The Journal of Finance, 53(5), 1589–1622. https://doi.org/10.1111/0022-1082.00066

Soleymani, F., & Paquet, E. (2020). Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—DeepBreath. Expert Systems with Applications, 156, 113456. https://doi.org/10.1016/j.eswa.2020.113456

Stoll, H. R., & Whaley, R. E. (1983). Transaction costs and the small firm effect. Journal of Financial Economics, 12(1), 57–79. https://doi.org/10.1016/0304-405x(83)90027-2

Swinkels, L. (2019). Treasury Bond Return Data Starting in 1962. Data, 4(3), 91. https://doi.org/10.3390/data4030091

Tang, K., Wang, W., & Xu, R. (2012). Size and performance of Chinese mutual funds: The role of economy of scale and liquidity. Pacific-Basin Finance Journal, 20(2), 228–246. https://doi.org/10.1016/j.pacfin.2011.09.002

Taniş, Arif & Erzurumlu, Yaman. (2021). ANN Based Holistic Performance Valuation and Mutual Fund Selection: Evidence from Turkey. Progress in Intelligent Decision Science: Proceeding of IDS 2020 (Advances in Intelligent Systems and Computing, 1301) (1st ed. 2021 ed.) (pp. 379-398). Istanbul, Turkey, Springer.

The American Economic Review , Dec., 1995, Vol. 85, No. 5 (Dec., 1995), pp. 1088- 1105. https://www.jstor.org/stable/2950976

Tian, Chenzhe (2020) Applications of Machine Learning and Reinforcement Learning in Investment and Trading. Dissertation, University of California, Irvine. https://escholarship.org/uc/item/7gc5m8f5

Trichilo, G. (2020). An Explainable Machine Learning Framework for Cross-Sectional Forecast-Based Fund Selection. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3707595

Valeria Fonti (2017) Feature Selection using LASSO. VU Amsterdam, https://scholar.google.fr/scholar_url?url=https://www.researchgate.net/profile/David-Booth-7/post/Regression-of-pairwise-trait-similarity-on-similarity-in-personal-attributes/attachment/5b18368d4cde260d15e3a4e3/AS%253A634606906785793%2540152813485788/download/werkstuk-fonti_tcm235-836234.pdf&hl=fr&sa=X&ei=rJ4WYe3iEcqgmQHFlrVA&scisig=AAGBfm1jcTRjZNfs4Ie_PkGxMFLtNj7BwA&oi=scholarr

Wijaya, Liliana & Irawan, Randy & Mahadwartha, Putu. (2018). Test of Fama a French five factor-model on Indonesian stock market. Advances in Social Science, Education and Humanities Research (ASSEHR), volume 186, 15th International Symposium on Management (INSYMA 2018) 10.2991/insyma-18.2018.12.

William G. Droms & David A. Walker (2001) Persistence of mutual fund operating characteristics: returns, turnover rates, and expense ratios, Applied Financial Economics, 11:4, 457-466, DOI: 10.1080/096031001300314018

Wolff, D., & Echterling, F. (2020). Stock Picking with Machine Learning. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3607845

Wu, W., Chen, J., Yang, Z. B., & Tindall, M. L. (2019). A Cross-Sectional Machine Learning Approach for Hedge Fund Return Prediction and Fund Selection. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3238466

Yakov Amihud, Ruslan Goyenko (2013) Mutual Fund's $R^2$ as Predictor of Performance, *The Review of Financial Studies*, Volume 26, Issue 3, March 2013, Pages 667–694, https://doi.org/10.1093/rfs/hhs182

Yap, C. J., & Pierce, R. M. (2008). Managed Equity Fund Attributes and Performance: Australian Evidence. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.1296987

Zheng, L. (1999). Is Money Smart? A Study of Mutual Fund Investors' Fund Selection Ability. The Journal of Finance, 54(3), 901–933. https://doi.org/10.1111/0022-1082.00131

# 9. Appendix

*Appendix 1: Bloomberg fund classification system*

The principal investment strategy is the following :

- Value: A value fund primarily holds stocks that are deemed to be undervalued in price and that are likely to pay dividends.
- Growth: A growth fund invests in stocks with growth as its primary goal. The portfolio will mainly consist of companies with above-average growth in earnings that reinvest their earnings into expansion, acquisitions, and/or research and development.
- Blend: A blend fund invests in a combination of value and growth stocks.
- Bear Market: Bear-market funds invest in short positions and derivatives in order to profit from stocks that drop in price. Because these portfolios often have extensive holdings in shorts or puts, their returns generally move in the opposite direction of the benchmark index. The Fund invests in out-of-favor stocks.
- Long/Short: Long/short Funds employ both long and short equity positions. The fund's goal is not to be market neutral, but rather to switch back and forth between an overall net long or net short portfolio, depending on market conditions and opportunities where a manager sees differences between a security's value and market price. Investments will usually be regional or sector specific.
- Market Neutral: A market neutral fund invests predominately in equity securities. The funds will long equities perceived as undervalued and short equities perceived as overvalued. This strategy will seek positive/absolute returns regardless of the overall market direction.
- Small Cap: The definition of a small cap will vary between fund companies. A general guide to this selection is a fund that targets investments in companies with a market cap of less than 1.5 billion.
- Mid Cap: The definition of a mid-cap will vary between fund companies. A general guide to this selection is a fund that targets investments in companies with a market cap of greater than 1.5 billion but less than 8 billion.
- Large: The definition of a large cap will vary between fund companies. A general guide to this selection is a fund that targets investments in companies with a market cap of greater than 8 billion.
- Broad Market (otherwise known as multi-cap): Funds investing in small, mid or large cap companies will be updated as Broad Market.

Source: Bloomberg (2013). Bloomberg Fund Classifications: Guide to our new funds Classification System. PDF Document available on the website : https://cpb-us-e1.wpmucdn.com/wp.txstate.edu/dist/2/916/files/2020/01/tmp14.pdf. Consulted on 13 August 2021

*Appendix 2 : Monthly returns of the second and third predicted years of the equal-weighted portfolio*

**Year 2 : 07/2019-06/2020**

| Date | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic | Actual | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|
| 29/07/2019 | 0,33% | 0,33% | 0,59% | 1,00% | 0,88% | 0,33% | 0,33% | 0,86% | -1,30% |
| 29/08/2019 | 0,30% | 0,29% | 0,53% | 0,97% | -0,08% | 0,27% | 0,28% | -2,44% | 1,83% |
| 29/09/2019 | 0,23% | 0,23% | 0,47% | 0,78% | -1,30% | 0,22% | 0,23% | 0,87% | -1,70% |
| 29/10/2019 | 0,54% | 0,53% | 0,75% | 0,84% | 5,15% | 0,55% | 0,54% | 2,03% | -2,02% |
| 29/11/2019 | 0,31% | 0,31% | 0,55% | 0,71% | 0,29% | 0,32% | 0,32% | 3,09% | -3,35% |
| 29/12/2019 | 0,18% | 0,22% | 0,38% | 0,76% | -3,08% | 0,23% | 0,22% | -1,41% | -2,82% |
| 29/01/2020 | 0,21% | 0,21% | 0,47% | 0,90% | -1,65% | 0,20% | 0,21% | -1,25% | 0,16% |
| 29/02/2020 | 0,32% | 0,31% | 0,60% | 1,14% | 0,22% | 0,30% | 0,31% | -8,20% | 8,79% |
| 29/03/2020 | 0,38% | 0,37% | 0,63% | 0,80% | 2,00% | 0,38% | 0,38% | -17,49% | 13,37% |
| 29/04/2020 | 0,35% | 0,35% | 0,60% | 0,73% | 1,26% | 0,35% | 0,35% | 12,45% | -11,94% |
| 29/05/2020 | 0,39% | 0,38% | 0,64% | 0,85% | 2,24% | 0,39% | 0,39% | 5,89% | -4,43% |
| 29/06/2020 | 0,26% | 0,26% | 0,49% | 0,79% | -0,85% | 0,26% | 0,26% | 2,43% | -1,82% |

**Year 3: 07/2020 - 06/2021**

| Date | OLS | Lasso | RF | DNN | XG Boost | Ridge | Elastic | Actual | S&P 500 |
|---|---|---|---|---|---|---|---|---|---|
| 29/07/2020 | 0,32% | 0,32% | 0,57% | 0,99% | 0,63% | 0,32% | 0,32% | 5,33% | -5,36% |
| 29/08/2020 | 0,28% | 0,28% | 0,53% | 1,00% | -0,34% | 0,26% | 0,27% | 5,05% | -6,77% |
| 29/09/2020 | 0,27% | 0,27% | 0,51% | 0,94% | -0,63% | 0,25% | 0,26% | -3,26% | 4,00% |
| 29/10/2020 | 0,46% | 0,45% | 0,69% | 0,80% | 3,43% | 0,47% | 0,46% | -1,34% | 2,81% |
| 29/11/2020 | 0,34% | 0,34% | 0,59% | 0,69% | 0,80% | 0,35% | 0,34% | 11,54% | -10,21% |
| 29/12/2020 | 0,20% | 0,23% | 0,41% | 0,68% | -2,56% | 0,25% | 0,24% | 0,99% | -3,64% |
| 29/01/2021 | 0,24% | 0,24% | 0,49% | 0,86% | -0,96% | 0,24% | 0,24% | -0,29% | 1,12% |
| 01/03/2021 | 0,31% | 0,29% | 0,60% | 1,27% | -0,13% | 0,28% | 0,30% | 3,94% | -2,58% |
| 29/03/2021 | 0,39% | 0,38% | 0,64% | 0,81% | 2,18% | 0,39% | 0,38% | 2,11% | -4,16% |
| 29/04/2021 | 0,34% | 0,34% | 0,60% | 0,75% | 1,21% | 0,35% | 0,35% | 4,12% | -5,11% |
| 29/05/2021 | 0,40% | 0,39% | 0,64% | 0,86% | 2,43% | 0,40% | 0,40% | 1,24% | -0,55% |
| 29/06/2021 | 0,24% | 0,24% | 0,47% | 0,81% | -1,27% | 0,24% | 0,24% | 0,61% | -2,20% |

*Appendix 3: Monthly returns of the predicted years depending on the seven machine learning methods with the selection of Group 8 from July 2018 to June 2021*

Year 1 : 07/2018- 06/2019

| Returns | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OLS** | **Actual** | **Lasso** | **Actual** | **RF** | **Actual** | **DNN** | **Actual** | **XG Boost** | **Actual** | | **Ridge** | **Actual** | **Elastic** | **Actual** |
| 0,78% | 2,39% | 0,86% | 2,39% | 1,66% | 1,72% | 4,51% | 2,34% | 1,55% | | 1,91% | 0,77% | 2,29% | 0,78% | 2,30% |
| 0,72% | 4,52% | 0,62% | 4,52% | 1,54% | 1,57% | 4,67% | 1,60% | 0,96% | | 4,56% | 0,70% | 4,43% | 0,72% | 4,42% |
| 0,67% | -1,02% | 0,48% | -1,02% | 1,43% | 1,23% | 4,45% | -0,52% | -1,29% | | -0,57% | 0,67% | -0,85% | 0,68% | -1,08% |
| 0,99% | -10,44% | -0,03% | -10,44% | 1,35% | -0,07% | 2,88% | -8,37% | 6,30% | | -10,75% | 0,98% | -10,26% | 0,98% | -10,42% |
| 0,76% | 0,51% | 0,84% | 0,51% | 1,57% | 1,43% | 4,16% | -1,07% | 1,03% | | 1,68% | 0,78% | 0,38% | 0,77% | 0,46% |
| 0,63% | -17,14% | -0,46% | -17,14% | 1,26% | -0,60% | 4,06% | -18,30% | -2,31% | | -18,12% | 0,68% | -17,20% | 0,67% | -17,04% |
| 0,66% | 9,46% | 1,31% | 9,46% | 1,50% | 2,53% | 4,58% | 8,52% | -0,34% | | 9,46% | 0,65% | 9,54% | 0,66% | 9,49% |
| 0,78% | 4,11% | 0,93% | 4,11% | 1,83% | 1,93% | 4,84% | 3,52% | 1,34% | | 4,47% | 0,74% | 3,99% | 0,76% | 4,14% |
| 0,86% | 0,58% | 0,81% | 0,58% | 1,72% | 1,67% | 3,66% | 0,29% | 4,62% | | 0,80% | 0,85% | 0,72% | 0,85% | 0,75% |
| 0,79% | 2,51% | 0,99% | 2,51% | 1,63% | 1,81% | 3,98% | 3,40% | 0,52% | | 3,36% | 0,79% | 2,83% | 0,79% | 2,77% |
| 0,88% | -6,96% | 0,02% | -6,96% | 1,72% | 0,57% | 3,43% | -6,67% | 3,49% | | -6,34% | 0,88% | -6,96% | 0,88% | -7,03% |
| 0,63% | 7,17% | 1,25% | 7,17% | 1,38% | 2,05% | 4,36% | 6,76% | -1,86% | | 6,74% | 0,64% | 6,94% | 0,64% | 7,15% |

Year 2 : 07/2019 - 06/2020

| Returns | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OLS** | **Actual** | **Lasso** | **Actual** | **RF** | **Actual** | **DNN** | **Actual** | **XG Boost** | **Actual** | | **Ridge** | **Actual** | **Elastic** | **Actual** |
| 0,78% | 0,98% | 0,78% | 0,98% | 1,65% | 0,10% | 0,78% | 1,79% | 1,96% | | 1,79% | 0,77% | 1,19% | 0,78% | 1,02% |
| 0,72% | -3,15% | 0,73% | -3,15% | 1,48% | -2,27% | 0,94% | -2,11% | 1,38% | | -2,11% | 0,70% | -3,09% | 0,72% | -3,16% |
| 0,66% | -0,18% | 0,68% | -0,18% | 1,31% | 1,05% | 0,82% | -0,95% | -1,48% | | -0,95% | 0,65% | -0,09% | 0,66% | -0,09% |
| 1,02% | 2,35% | 0,99% | 2,35% | 1,44% | 1,38% | 0,37% | 2,80% | 6,97% | | 2,80% | 1,01% | 2,40% | 1,01% | 2,21% |
| 0,75% | 3,92% | 0,76% | 3,92% | 1,55% | 2,33% | 0,57% | 4,84% | 0,83% | | 4,84% | 0,77% | 3,92% | 0,76% | 3,86% |
| 0,62% | -1,89% | 0,66% | -1,89% | 1,27% | 0,95% | 0,57% | -2,71% | -2,81% | | -2,71% | 0,68% | -1,83% | 0,67% | -1,69% |
| 0,65% | -0,84% | 0,65% | -0,84% | 1,43% | -2,57% | 0,91% | 0,37% | -0,45% | | 0,37% | 0,64% | -0,61% | 0,65% | -0,84% |
| 0,78% | -7,23% | 0,75% | -7,23% | 1,68% | -8,73% | 0,99% | -7,01% | 1,45% | | -7,01% | 0,75% | -7,24% | 0,76% | -7,32% |
| 0,83% | -16,71% | 0,82% | -16,71% | 1,76% | -21,72% | 0,55% | -14,74% | 3,10% | | -14,74% | 0,83% | -16,93% | 0,83% | -16,98% |
| 0,79% | 12,91% | 0,79% | 12,91% | 1,66% | 13,47% | 0,50% | 13,34% | 0,95% | | 13,34% | 0,80% | 12,75% | 0,80% | 12,94% |
| 0,84% | 7,16% | 0,83% | 7,16% | 1,75% | 5,95% | 0,54% | 8,02% | 3,08% | | 8,02% | 0,84% | 7,09% | 0,84% | 7,14% |
| 0,69% | 2,98% | 0,70% | 2,98% | 1,45% | 2,81% | 0,73% | 3,57% | -0,24% | | 3,57% | 0,70% | 3,08% | 0,70% | 3,07% |

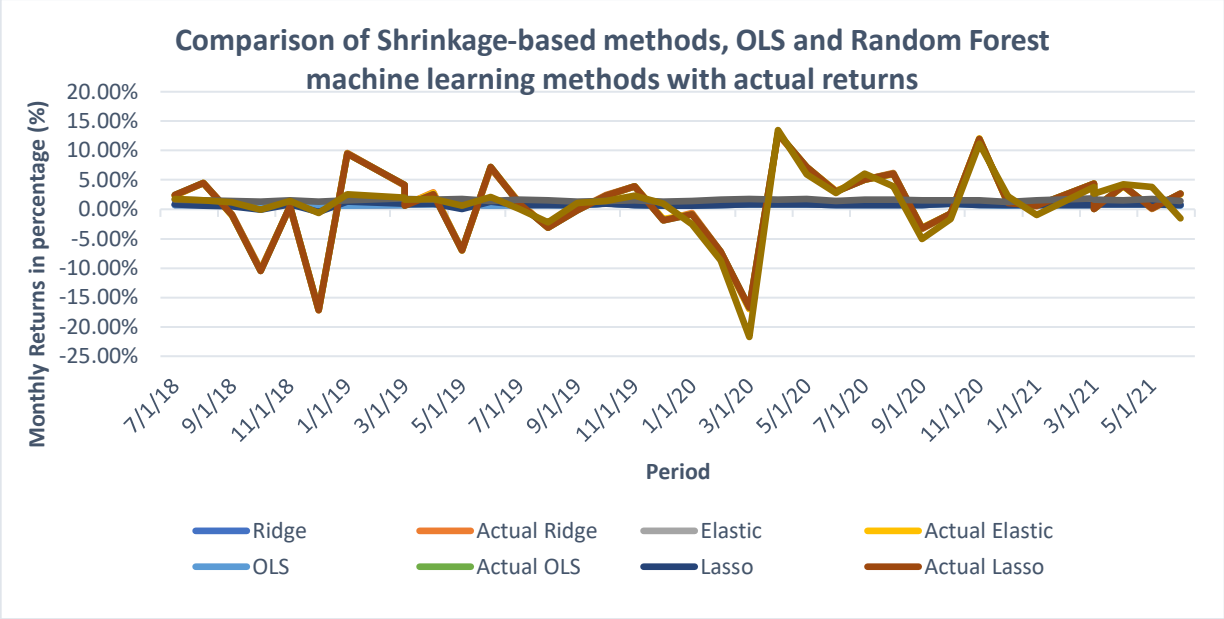
Year 3: 07/2020-06/2021

| Returns | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OLS** | **Actual** | **Lasso** | **Actual** | **RF** | **Actual** | **DNN** | **Actual** | **XG Boost** | **Actual** | | **Ridge** | **Actual** | **Elastic** | **Actual** |
| 0,77% | 5,08% | 0,77% | 5,08% | 1,67% | 6,04% | 4,56% | 6,36% | 1,25% | | 5,22% | 0,76% | 5,05% | 0,77% | 5,05% |
| 0,71% | 6,04% | 0,72% | 6,04% | 1,61% | 3,90% | 4,59% | 5,19% | 0,64% | | 5,86% | 0,70% | 6,18% | 0,71% | 6,01% |
| 0,70% | -3,23% | 0,71% | -3,23% | 1,54% | -5,01% | 4,55% | -3,25% | -0,31% | | -2,69% | 0,69% | -3,34% | 0,70% | -3,17% |
| 0,92% | -0,70% | 0,91% | -0,70% | 1,53% | -1,70% | 3,24% | -1,89% | 4,74% | | -1,43% | 0,92% | -0,76% | 0,92% | -0,78% |
| 0,78% | 11,99% | 0,78% | 11,99% | 1,58% | 11,15% | 4,01% | 11,62% | 1,46% | | 11,45% | 0,79% | 11,92% | 0,79% | 12,05% |
| 0,65% | 1,06% | 0,68% | 1,06% | 1,24% | 2,26% | 4,10% | 0,71% | -1,80% | | 0,42% | 0,69% | 1,12% | 0,68% | 0,96% |
| 0,68% | 0,60% | 0,69% | 0,60% | 1,47% | -0,94% | 4,50% | -0,83% | 0,03% | | 0,20% | 0,68% | 0,56% | 0,68% | 0,56% |
| 0,77% | 4,31% | 0,74% | 4,31% | 1,85% | 3,84% | 4,88% | 4,17% | 1,59% | | 2,96% | 0,73% | 4,35% | 0,75% | 4,41% |
| 0,84% | 0,01% | 0,83% | 0,01% | 1,69% | 2,70% | 3,85% | 2,40% | 3,74% | | -2,30% | 0,83% | 0,09% | 0,84% | 0,17% |
| 0,79% | 3,98% | 0,79% | 3,98% | 1,58% | 4,20% | 3,99% | 4,81% | 0,92% | | 4,01% | 0,80% | 4,01% | 0,80% | 3,94% |
| 0,85% | 0,11% | 0,84% | 0,11% | 1,72% | 3,76% | 3,68% | 1,62% | 3,23% | | -0,44% | 0,86% | 0,06% | 0,85% | 0,23% |
| 0,67% | 2,70% | 0,69% | 2,70% | 1,39% | -1,58% | 4,33% | -0,15% | -0,72% | | 3,07% | 0,68% | 2,65% | 0,68% | 2,54% |

*Appendix 4 : Monthly returns of the predicted years depending on the Deep Neural Network and XG Boost machine learning methods with the selection of Group 8 from July 2018 to June 2021*

| Returns | | | | |
|---|---|---|---|---|
| **Date** | **DNN** | **Actual DNN** | **XG Boost** | **Actual XG Boost** |
| 29/07/2018 | 4,51% | 2,34% | 1,55% | 1,91% |
| 29/08/2018 | 4,67% | 1,60% | 0,96% | 4,56% |
| 29/09/2018 | 4,45% | -0,52% | -1,29% | -0,57% |
| 29/10/2018 | 2,88% | -8,37% | 6,30% | -10,75% |
| 29/11/2018 | 4,16% | -1,07% | 1,03% | 1,68% |
| 29/12/2018 | 4,06% | -18,30% | -2,31% | -18,12% |
| 29/01/2019 | 4,58% | 8,52% | -0,34% | 9,46% |
| 01/03/2019 | 4,84% | 3,52% | 1,34% | 4,47% |
| 29/03/2019 | 3,66% | 0,29% | 4,62% | 0,80% |
| 29/04/2019 | 3,98% | 3,40% | 0,52% | 3,36% |
| 29/05/2019 | 3,43% | -6,67% | 3,49% | -6,34% |
| 29/06/2019 | 4,36% | 6,76% | -1,86% | 6,74% |
| 29/07/2019 | 0,78% | 1,79% | 1,96% | 1,79% |
| 29/08/2019 | 0,94% | -2,11% | 1,38% | -2,11% |
| 29/09/2019 | 0,82% | -0,95% | -1,48% | -0,95% |
| 29/10/2019 | 0,37% | 2,80% | 6,97% | 2,80% |
| 29/11/2019 | 0,57% | 4,84% | 0,83% | 4,84% |
| 29/12/2019 | 0,57% | -2,71% | -2,81% | -2,71% |
| 29/01/2020 | 0,91% | 0,37% | -0,45% | 0,37% |
| 29/02/2020 | 0,99% | -7,01% | 1,45% | -7,01% |
| 29/03/2020 | 0,55% | -14,74% | 3,10% | -14,74% |
| 29/04/2020 | 0,50% | 13,34% | 0,95% | 13,34% |
| 29/05/2020 | 0,54% | 8,02% | 3,08% | 8,02% |
| 29/06/2020 | 0,73% | 3,57% | -0,24% | 3,57% |
| 29/07/2020 | 4,56% | 6,36% | 1,25% | 5,22% |
| 29/08/2020 | 4,59% | 5,19% | 0,64% | 5,86% |
| 29/09/2020 | 4,55% | -3,25% | -0,31% | -2,69% |
| 29/10/2020 | 3,24% | -1,89% | 4,74% | -1,43% |
| 29/11/2020 | 4,01% | 11,62% | 1,46% | 11,45% |
| 29/12/2020 | 4,10% | 0,71% | -1,80% | 0,42% |
| 29/01/2021 | 4,50% | -0,83% | 0,03% | 0,20% |
| 01/03/2021 | 4,88% | 4,17% | 1,59% | 2,96% |
| 29/03/2021 | 3,85% | 2,40% | 3,74% | -2,30% |
| 29/04/2021 | 3,99% | 4,81% | 0,92% | 4,01% |
| 29/05/2021 | 3,68% | 1,62% | 3,23% | -0,44% |
| 29/06/2021 | 4,33% | -0,15% | -0,72% | 3,07% |

*Appendix 5: Comparision of OLS, shrinkage-based methods and Random Forest machine learning methods with actual returns from July 2018 to June 2021*




*Appendix 6 : Cumulative returns of the predicted years depending on the seven machine learning methods with the selection of Group 8 from July 2018 to June 2021*

| Cumulative returns | | | | | | |
|---|---|---|---|---|---|---|
| Actual OLS | Actual Lasso | Actual RF | Actual DNN | Actual XG Bc | Actual ridge | Actual Elasti |
| 2,39% | 2,39% | 1,72% | 2,34% | 1,91% | 2,29% | 2,30% |
| 6,91% | 6,91% | 3,29% | 3,93% | 6,47% | 6,72% | 6,73% |
| 5,89% | 5,89% | 4,52% | 3,41% | 5,90% | 5,87% | 5,65% |
| -4,55% | -4,55% | 4,46% | -4,96% | -4,85% | -4,39% | -4,78% |
| -4,04% | -4,04% | 5,89% | -6,04% | -3,16% | -4,01% | -4,31% |
| -21,18% | -21,18% | 5,29% | -24,34% | -21,28% | -21,21% | -21,36% |
| -11,72% | -11,72% | 7,82% | -15,82% | -11,82% | -11,67% | -11,86% |
| -7,61% | -7,61% | 9,75% | -12,30% | -7,35% | -7,68% | -7,72% |
| -7,03% | -7,03% | 11,42% | -12,00% | -6,55% | -6,96% | -6,98% |
| -4,52% | -4,52% | 13,23% | -8,60% | -3,18% | -4,12% | -4,20% |
| -11,48% | -11,48% | 13,80% | -15,28% | -9,52% | -11,08% | -11,23% |
| -4,31% | -4,31% | 15,85% | -8,51% | -2,78% | -4,14% | -4,08% |
| -3,33% | -3,33% | 15,95% | -6,72% | -0,99% | -2,95% | -3,06% |
| -6,48% | -6,48% | 13,68% | -8,84% | -3,10% | -6,05% | -6,22% |
| -6,66% | -6,66% | 14,74% | -9,79% | -4,06% | -6,14% | -6,31% |
| -4,31% | -4,31% | 16,11% | -6,99% | -1,26% | -3,74% | -4,10% |
| -0,38% | -0,38% | 18,44% | -2,15% | 3,58% | 0,18% | -0,24% |
| -2,27% | -2,27% | 19,39% | -4,86% | 0,87% | -1,65% | -1,93% |
| -3,12% | -3,12% | 16,82% | -4,49% | 1,24% | -2,26% | -2,77% |
| -10,34% | -10,34% | 8,09% | -11,50% | -5,77% | -9,50% | -10,09% |
| -27,06% | -27,06% | -13,64% | -26,24% | -20,51% | -26,43% | -27,08% |
| -14,15% | -14,15% | -0,16% | -12,90% | -7,17% | -13,68% | -14,14% |
| -6,99% | -6,99% | 5,79% | -4,88% | 0,85% | -6,59% | -6,99% |
| -4,00% | -4,00% | 8,60% | -1,30% | 4,43% | -3,51% | -3,93% |
| 1,07% | 1,07% | 14,64% | 5,05% | 9,65% | 1,54% | 1,12% |
| 7,11% | 7,11% | 18,54% | 10,24% | 15,51% | 7,72% | 7,13% |
| 3,88% | 3,88% | 13,52% | 6,99% | 12,82% | 4,39% | 3,96% |
| 3,18% | 3,18% | 11,83% | 5,10% | 11,39% | 3,63% | 3,17% |
| 15,17% | 15,17% | 22,97% | 16,72% | 22,84% | 15,55% | 15,22% |
| 16,23% | 16,23% | 25,23% | 17,43% | 23,26% | 16,67% | 16,18% |
| 16,83% | 16,83% | 24,29% | 16,60% | 23,46% | 17,23% | 16,75% |
| 21,14% | 21,14% | 28,13% | 20,76% | 26,42% | 21,58% | 21,16% |
| 21,15% | 21,15% | 30,83% | 23,17% | 24,12% | 21,68% | 21,33% |
| 25,13% | 25,13% | 35,02% | 27,97% | 28,13% | 25,69% | 25,27% |
| 25,24% | 25,24% | 38,78% | 29,59% | 27,69% | 25,75% | 25,50% |
| 27,95% | 27,95% | 37,21% | 29,43% | 30,76% | 28,39% | 28,03% |

*Appendix 7 : Comparison of the monthly returns, average annual returns, standard deviation and Share ratio of the equal-weighted portfolio and a full investment of 100% on the S&P 500 index from July 2018 to June 21*

Year 1

| | Actual Results | S&P 500 Index |
|---|---|---|
| Average annual return | -4,94% | -7,90% |
| Standard deviation | 23,55% | 18,40% |
| Sharpe ratio | -0,21 | -0,43 |

Year 2

| | Actual Results | S&P 500 Index |
|---|---|---|
| Average annual return | -3,16% | -5,25% |
| Standard deviation | 24,34% | 21,22% |
| Sharpe ratio | -0,13 | -0,25 |

Year 3

| | Actual Results | S&P 500 Index |
|---|---|---|
| Average annual return | 30,03% | -32,65% |
| Standard deviation | 12,82% | 13,55% |
| Sharpe ratio | -0,21 | -0,43 |

| Predicted period | Dates | Returns | SPX Index |
|---|---|---|---|
| Year 1 | 31/07/2018 | 2,31% | -3,54% |
| | 31/08/2018 | 1,85% | -2,98% |
| | 28/09/2018 | -0,27% | -0,43% |
| | 31/10/2018 | -8,58% | 7,19% |
| | 30/11/2018 | 0,73% | -1,77% |
| | 31/12/2018 | -17,07% | 9,63% |
| | 31/01/2019 | 8,67% | -7,57% |
| | 28/02/2019 | 3,45% | -2,93% |
| | 29/03/2019 | 0,59% | -1,78% |
| | 30/04/2019 | 3,11% | -3,86% |
| | 31/05/2019 | -6,39% | 6,80% |
| | 28/06/2019 | 6,66% | -6,67% |
| Year 2 | 31/07/2019 | 0,86% | -1,30% |
| | 30/08/2019 | -2,44% | 1,83% |
| | 30/09/2019 | 0,87% | -1,70% |
| | 31/10/2019 | 2,03% | -2,02% |
| | 29/11/2019 | 3,09% | -3,35% |
| | 31/12/2019 | -1,41% | -2,82% |
| | 31/01/2020 | -1,25% | 0,16% |
| | 28/02/2020 | -8,20% | 8,79% |
| | 31/03/2020 | -17,49% | 13,37% |
| | 30/04/2020 | 12,45% | -11,94% |
| | 29/05/2020 | 5,89% | -4,43% |
| | 30/06/2020 | 2,43% | -1,82% |
| Year 3 | 31/07/2020 | 5,33% | -5,36% |
| | 31/08/2020 | 5,05% | -6,77% |
| | 30/09/2020 | -3,26% | 4,00% |
| | 30/10/2020 | -1,34% | 2,81% |
| | 30/11/2020 | 11,54% | -10,21% |
| | 31/12/2020 | 0,99% | -3,64% |
| | 29/01/2021 | -0,29% | 1,12% |
| | 26/02/2021 | 3,94% | -2,58% |
| | 31/03/2021 | 2,11% | -4,16% |
| | 30/04/2021 | 4,12% | -5,11% |
| | 31/05/2021 | 1,24% | -0,55% |
| | 30/06/2021 | 0,61% | -2,20% |