Stock selection using machine learning techniques

Author. Eugenio Martínez, Jose Manuel[*]
Director. Ferreiro Castilla, Albert[†]

04 septiembre 2017

Abstract

The main objective of this research is to challenge classic portfolio management theories such as Markovitz's Portfolio Selection Theory [1] or Sharpe Diagonal Model [2], where after making a small presentation of them and foundations on which they are based, we will discuss depthly about what other factors affect when reasoning about the paradigm "the greater risk, the greater profitability" or "high risk, high return". We review quantitative investment strategies or factors that are commonly used in practice, including quantitative factors which are classified into four categories; low volatility, value, quality, momemtum. In this research, we explore the fundamental underpinnings of these factors, along with the relevant academic literature. All these previous arguments are based on the empirical irregularity known as low volatility anomaly, where it is documented how the assets with low implicit risk overperform in terms of profitability those with an implicit higher risk. In the process, we introduce the concept, Smart Beta, whose basis is how to combine these factors in a multi-factor [17] investment process in order to improve profitability or neutralize exposure to stressors and therefore extract a cleaner factor premium. For this we will create a factor that tries to gather all these positive characteristics but in an elementary way. Next, we will introduce advanced techniques of statistics known as machine learning which will construct a new factor under different variables trying to optimize the best results which is the main objective of the research.

Index

1. Key words **4**

2. Introduction to invesment **4**
2.1 Asset Management . . . 5
2.1.1 Risk of an asset . . . 5
2.1.2 Return of an asset . . . 6
2.2 Portfolio Management . . . 6
2.2.1 Portfolio return . . . 6
2.2.2 Portfolio risk . . . 7
2.2.3 Systematic risk . . . 8
2.2.4 Non-systematic risk . . . 8
2.2.5 Reducing Non-systematic (Unique) Risk by Diversification . . . 8
2.3 Beta . . . 10
2.3.1 Definition . . . 10
2.3.2 Beta features . . . 10
2.3.1 Disadvantages of beta's coefficient . . . 10

3. Classical Theories of Portfolio Investment **11**
3.1 H. Markowitz portfolio selection . . . 11
3.1.1 Model approach . . . 11
3.1.2 Review of Markowitz's Mean-Variance Model . . . 12
3.2 Willian F. Sharpe: A simplified model for portfolio analysis. . . . 12
3.2.1 Model approach . . . 12
3.3 Capital Asset Pricing Model (CAPM) . . . 13

[*]Universitat Autònoma de Barcelona
[†]Universitat Autònoma de Barcelona

3.3.1 CAPM model assumptions . . . 13
3.3.2 Capital Market Line (CML) . . . 14
3.3.3 Security Market Line (SML) . . . 16
3.3.4 Comparison of the CML with the SML . . . 16

4. Data processing **17**

5. Challenge to classic investment theories **20**
5.1 Low-Risk Investing . . . 21
5.1.1 Time-varying betas . . . 22

6. Exposure to fundamental factors **23**
6.1 Rationale behind risk factors . . . 23
6.1.1 Value . . . 23
6.1.2 Momemtum . . . 26
6.1.3 Low risk . . . 27
6.1.4 Quality . . . 28
6.2 Fundamental factor comparison . . . 30

7. Smart Beta **30**
7.1 Elementary Smart Beta . . . 31
7.1.1 Construction of the elemental smart beta factor . . . 31
7.1.2 Results obtained by the elemental smart beta factor . . . 31

8. Machine learning approach: next horizont to smart beta **32**
8.1 Preliminaries . . . 32
8.2 Machine Learning Techniques . . . 32
8.2.1 Decission trees . . . 33
8.2.2 Random Forest . . . 33
8.2.2.1 Random forest: Influence measures . . . 34
8.2.2.2 Random forest: Applications in our research . . . 34
8.2.2.3 First and second term variable importances computation . . . 35
8.3 Smart beta factor results . . . 36
8.3.1 One month training portfolio results . . . 36
8.3.2 Three months training portfolio results . . . 37

9. Conclusions **39**

9. Bibliografy **40**

10. Appendix **41**

Resumen

El principal objetivo de este trabajo es desafiar las teorías clásicas de gestión de cartera más conocidas como la Teoría de la Selección de Cartera de Markovitz [1] o Sharpe Diagonal Model [2], donde después de exponer los fundamentos en los que se basan, discutiremos en profundidad sobre qué otros factores afectan al razonar sobre el paradigma *a mayor riesgo, mayor rentabilidad* o *alto riesgo, alto rendimiento*. Revisaremos estrategias de inversión cuantitativa y los factores que se utilizan comúnmente en inversiones reales. En este trabajo, los factores que utilizamos se pueden agrupar en cuatro categorías: low risk, value, quality y momemtum, en los que profunduzaremos apoyándonos en la literatura académica más relevante. Todos estos argumentos expuestos hasta este punto son la base para introducir la irregularidad empírica conocida como *anomalía de baja volatilidad*, donde el concepto principal trata de demostrar cómo los activos con bajo riesgo implícito obtienen un rendimiento superior en términos de rentabilidad, que aquellos con un riesgo implícito mayor. A continuación, también presentaremos el concepto *Smart Beta*, cuya idea principal es cómo conseguir una combinación óptima de factores fundamentales a través de un proceso de inversión multi-factorial [17] con el objetivo de conseguir mejorar la rentabilidad y/o neutralizar la exposición a factores de estrés y por tanto extraer una prima de factor más eficiente. Para conseguirlo, crearemos un factor que intente reunir todas estas fortalezas donde en primer lugar, de una manera elemental, utilizaremos ponderaciones bajo nuestro criterio basadas en el comportamiento de cada uno de los factores de manera individual, y por otro lado, crearemos un nuevo factor basado en técnicas avanzadas de estadística conocidas como *machine learning techniques*, con el objetivo de optimizar los resultados obtenidos bajo el criterio elemental, y tratando de demostrar el objetivo principal del trabajo.

Resum

El principal objectiu d'aquest treball és desafiar les teories clàssiques de gestió de cartera més conegudes com pot ser la Teoria de la Selecció de Cartera de Markovitz [1] o Sharpe Diagonal Model [2], on després d'exposar els fonaments en què es basen, discutirem en profunditat sobre quins altres factors afecten al raonar sobre el paradigma *a major risc, major rendibilitat* o *alt risc, alt rendiment*. Revisarem estratègies d'inversió quantitativa i els factors que s'utilitzen comunment en inversions reals. En aquest treball els factors que utilitzem es poden agrupar en quatre categories: low risk, value, quality i momemtum, agafant com a suport la literatura acadèmica més rellevant. Tots aquests arguments exposats fins aquest punt són la base per introduir la irregularitat empírica coneguda com anomalia de baixa volatilitat, on el concepte principal tracta de demostrar com els actius amb baix risc implícit obtenen un rendiment superior en termes de rendibilitat, és a dir, aquells amb un risc implícit més gran. A continuació, també introduïrem el concepte *Smart Beta*, la idea principal és com aconseguir una combinanació òptima de factors fonamentals a través d'un procés d'inversió multi-factorial [17] amb l'objectiu d'aconseguir millorar la rendibilitat i/o neutralitzar l'exposició a factors d'estrès, i per tant, extraure una prima de factor més eficient. Per aconseguir-ho, farem servir un nou factor que intenti reunir, en primer lloc, totes aquestes fortaleses d'una manera elemental, utilitzant ponderacions mitjantçant el nostre criteri. Un cop visualitzat el comportament de cadascun dels factors de manera individual, crearem un nou factor basat en tècniques avançades d'estadística conegudes com *machine learning*, amb l'objectiu d'optimitzar els resultats obtinguts sota el criteri elemental, i tractant de demostrar l'objectiu principal del treball.

1. Key words

Portfolio management, risk, profiability, fundamental factors, smart beta, machine learning.

2. Introduction to invesment

We begin by presenting the concept of a financial asset which refers to a type of intangible asset that represents a legal right over a future monetary amount. That is, it awards at the buyer the right to receive some future income from the seller, a right over the issuer's real assets and the cash it generates. They can be issued by any economic unit, be it a private company like Government, etc. but the current study will focus on equity assets exclusively.

The acquisition of financial assets is called financial investment, and through it, investors build a portfolio of securities. Thanks to these instruments the entities with a debt can be financed and in turn, people who want to invest their savings get some returns by investing in that debt. Financial assets are represented by physical certificates or book entries ie. bank's account.

Any investor who wants to manage their portfolio efficiently must have a portfolio built under a clearly identified structure as it is the basis for profit. In addition, all rational investors and financial economists agree on the need to diversify the risk to optimize these benefits;risk management is no less important than the management of the profitability of the same.

The main objective of portfolio management is to achieve performance against a benchmark index, either any traditional market index such as the SP500, NASDAQ, etc., or referenced to any interest rate curve such as LIBOR or EURIBOR, given a risk ratio. Although the 4 keys that every quantitative investor should take into account are:

1. The alpha model, which forescast what performance you will get from the portfolio based on the investment made, ie. how much money will I earn.
2. The risk and the correlation of each asset that forms the portfolio.
3. How to deal with performance and risk predictions to maximize profits optimally.
4. The entire portfolio management process should be optimized to be able to market how and when it is needed.

In this sense, Markowitz contributed a model in which he affirmed that *"every rational investor seeks to maximize the profitability of the portfolio while minimizing risk"*. Therefore, each investor should optimize the performance-risk binomial in order to measure the mathematical expectation of such returns and risk through variance, a model known as the Mean Variance Model (MV) [3] that was extended in all areas of quantitative finance such as asset allocation, equity and fixed income portfolio management.

As a result of the above statement, the investors could have two possible objectives faced, risk and performance, so you should look for a diversification within your portfolio, becasuse achieving optimal diversification reduces risk.

2.1 Asset Management

Over the past century, risk measurement has undergone the greatest development as far as the theory of investment is concerned. Galitz [4] defines the concept of risk as any alteration either positive or negative when it comes to achieving the objectives.

Knight (1921) [5] provided a rather trivial definition that distinguished uncertainty with risk: *"decision makers crudely operate in a world of random uncertainty, and risk is a condition in which the decision maker assigns formal mathematical probabilities to specify the uncertainty."*

It was Markowitz (1952) [1] who under his theory exposed the idea we discussed above, any rational investor seeks to maximize it is expected utility function by assuming a certain measure of risk which will be measured by it is variance.

After trying to define the risk under the citations to the different classical authors we can give our own definition of the same as possible circumstances that are not under the control of the investor itself, and a way to mitigate or reduce that risk is to associate a probability to each of these circumstances but it is important to emphasize that it will never be possible to reduce the risk completely, since it is one of the characteristics implicit in the equity securities.

Mathematically, what expresses risk or dispersion in the data is the expected value in terms of variance or standard deviation, and all assets will differentiate to each other by that level of risk and on the other hand, from the profitability obtained.

2.1.1 Risk of an asset

$$Variance: \quad \sigma^2(R_i) = \frac{(R_i - E(R_i))^2}{n} \tag{1}$$

$$Standard \;\; deviation: \sigma(R_i) = \sqrt{\sigma^2(R_i)} \tag{2}$$

where,

R_i: return of asset i.

$E(R_i)$: expected retunr of asset i.

2.1.2 Return of an asset

On the other hand, we must also define the profitability of a financial asset, which it is understood by us as the ability of an asset to offer an income. We also denote the absolute return (equation 3) and relative return (equation 4) of an asset that we will use in this research:

$$Absolute \quad Return: \quad R_{initial} - R_{final} \tag{3}$$

$$Relative \quad Return: \quad \frac{R_{initial} - R_{final}}{R_{initial}} \tag{4}$$

where,

$R_{initial}$: initial return

R_{final}: final return

The expected return on an asset is defined as the return obtained by the sum of each asset i multiplied by the probability that scenario appears for title i and is denoted as follows:

$$E[R_i] = \sum_{i=1}^{N} R_i P_i \tag{5}$$

where,

$E(R_i)$: expected retunr of asset i.

R_i: expected return of asset i.

P_i: event probability of asset i.

2.2 Portfolio Management

2.2.1 Portfolio return

The combination of the expected returns, or averages of probability distributions of possible returns, of all assets in an investment portfolio. Hence portfolio return will be given by the weighted average of the random returns of the securities that compose it:

$$R_p = w_1 R_1 + w_2 R_2 + \cdots + w_N R_N = \sum_{i=1}^{N} w_i R_i \tag{6}$$

$$\forall w_1, w_2, ..., w_N$$

where,

R_p: expected portfolio return

w_i: fraction of the investment budget allocated to the asset i

N: number of assets

R_i: expected return of asset i

If we apply the mathematical expectation operator a we can conclude that the mathematical expectation as the sum of random variables weighted by their respective constants is the weighted average of the expected returns of the securities that make up the portfolio:

$$E(R_p) = \sum_{i=1}^{N} w_i E(R_i) \tag{7}$$

2.2.2 Portfolio risk

Exactly the same happens with the risk, an investor wants to know the risk that assumes with the portfolio that has formed and for this will apply the operator variance as follows:

$$\sigma^2(R_p) = \sum_{i=1}^{N} w_1^2\sigma^2(R_1)+...+w_N^2\sigma^2(R_N)+2w_1w_2Cov(R_1,R_2)+2w_1w_2Cov(R_1,R_3)+\cdots+2w_{n-1}w_nCov(R_{n-1},R_n) \tag{8}$$

From this expression above we can obtain a general risk equation expression:

$$\sigma^2(R_p) = \sum_{i=1}^{N} w_i^2\sigma^2(R_i) + \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j Cov(R_i,R_j) \tag{9}$$

And now we consider two extreme cases:

$$\sigma^2(R_p) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j Cov(R_i,R_j) \begin{cases} CaseA: & si\ i \neq j \iff Cov(R_i,R_j) = 0 \\ CaseB: & si\ i = j \iff Cov(R_i,R_j) = \sigma^2(R_i) \end{cases}$$

Where,

R_p: expected portfolio return

σ_p^2: expected portfolio variance

w_i: fraction of the investment budget allocated to the asset i

w_j: fraction of the investment budget allocated to the asset j

$Cov(R_i, R_j)$: covariance between the performance of asset i and asset j

The covariance is:

$$Cov(R_i,R_j) = \sum_{i=1}^{N}(R_{ij} - E(R_i))(R_{kj} - E(R_k)) \tag{10}$$

where,

R_{ij}: return of asset i in time j

R_i: return of asset i

Let is denote correlation coefficient as follow:

$$\rho_{ij} = \frac{Cov(R_i,R_j)}{\sigma(R_i)\sigma(R_j)} \tag{11}$$

The variance of the return on the portfolio also depends on the covariances between the returns of the different securities and oblige the investor to treat the securities jointly, since the negative or reduced covariances reduce the variance of the performance of the portfolio as a whole. Therefore, a positive covariance warns that yields will tend to move in the same direction, while a negative covariance indicates that yields will move in opposite directions.

Once both the risk of an asset and the profiability have been defined, the main components of risk must be defined as systematic and non-systematic risk.

2.2.3 Systematic risk

Fundamentally, systematic risk is associated with a market. In other words, the risk that does not affect a particular stock or sector, but rather the entire market. To give an example, when a stock market crash in Japan the share price of Japonnesse stocks falls together simultaneously, but it may not affect Spanish stocks at all. It is an unpredictable risk but also impossible to avoid completely. Commonly, this systematic risk is also known as non-diversificable risk.

2.2.4 Non-systematic risk

On the contrary, there is the non-systematic risk, which is the particular risk of each value or investment. That is, the result of the specific factors of each of the investment securities. In the case of stocks, for example, it may be linked to the merge of the company, the discovery of a new product or service, report of negative benefits unexpected. This type of risk is also known as diversificable risk because it can be reduced or controlled adequately through a diversification strategy. To do this, the correlation coefficient of the asset itself should be used against the market index. By combining different assets whose correlation coefficient is positive which indicates that it behaves like the market, and other with negative ones, the meaning of which is that it behaves contrary to the market. Investment strategy is obtained by maximize the expected return on the asset portfolio and reduce it is risk. The risk of a security can be expressed as follows, from which will be extracted each of the parts that correspond to the risks seen in this section:

$$Unsystematic \;\; Risk = [R_i - E(R_i)] - [R_M - E(R_M)]\beta_M \tag{12}$$

where:

R_i is the actual return on the asset i

$E(R_i)$ is the expected return on the asset i

R_M is the actual return on the market

$E(R_M)$ is the expected return on the market

β_M is beta's market coeffient. See chapter 2.3 (Equation 16).

2.2.5 Reducing Non-systematic (Unique) Risk by Diversification

The aim of this section is to reduce portfolio risk through diversification, ie. the standard deviation of returns on the portfolio of assets may be less than the sum of the standard deviations from the assets. We can put an example: we know that when the economy is booming demand for new cars is high, and the automotive industry's returns are great, but as economic growth tends to lower, people will not be able to easily change their car and will keep it with spare parts. Then the spare parts industry, in this period, will obtain high yields. Because of the cyclical behavior of the automotive industry and the anti-cyclical nature of the spare parts industry, an investor with securities in the two industries may have more stable returns for diversification than if it invested only in one industry.

Systematic risk can not be reduced by diversification, but the unique component (non systematic risk) can be reduced only by diversification. If we consider two above cases above from equation 9:

Case A $Cov(R_i, R_j) = 0$

The returns of the assets are pairwise uncorrelated. Therefore,

$$\sigma^2(R_p) = \sum_{i=1}^{N} w_i^2 \sigma^2(R_i) \tag{13}$$

If we assume further that the portfolio is equally weighted, that is $w_i = \frac{1}{N} \ \ \forall i$. Then,

$$\sigma^2(R_p) = \frac{1}{N^2} \sum_{i=1}^{N} \sigma_i^2 \leq \frac{\sigma_M^2}{N} \tag{14}$$

where,

$$\sigma_M = max[\ \sigma_i \mid i = 1, 2, \ldots, N].$$

As N grows to infinity, $\sigma(R_p)$ goes to zero. This means, *the greater the number of pairwise uncorrelated assets, the smaller the risk of the portfolio.*

Case B: $Cov(R_i, R_j) = \sigma^2(R_i)$

The returns in the portfolio assets are similarly correlated. This can be realized if all assets have the same variance σ^2 and the correlation between all assets is constant $0 \leq C \leq 1$. Consider also the case of and equally weighted portfolio. Then,

$$\begin{aligned} \sigma^2(R_P) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \ \sigma_{ij} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} = \\ = \frac{1}{N^2} (\sum_{1 \leq i \leq N} \sigma_{ij} + \sum_{1 \leq i < j < N} \sigma_{ij}) = \frac{\sigma^2}{N} + (1 - \frac{1}{N}) C\sigma^2 = \\ = \frac{1 - C}{N} \sigma^2 + C\sigma^2 \end{aligned} \tag{15}$$

This shows that no matter how large N is made, it is impossible to reduce the variance of the portfolio below σ^2. The conclusion is that, *the greater the presence of similarly correlated assets, the closer is the risk of the portfolio to a risk common to all assets.* Therefore, diversification in a mean-variance world is acomplished by considering highly uncorrelated assets in some reasonable number.

2.3 Beta

2.3.1 Definition

Beta coefficient is denoted as a measure for the volatility of an asset with respect to the variability of the market itself. And it should be expressed as follows:

$$\beta_i = \frac{Cov(R_i, R_m)}{\sigma_m^2} \tag{16}$$

The beta coefficient indicates how the price of an asset reacts to market fluctuations. The more sensitive the price of an asset to these fluctuations, the greater its beta coefficient.

It follows that β_j is the slope of the best linear regression predictor of the i_{th} stock's return using the returns of the tangent portfolio[1] as predictor variable, ie:

$$\hat{R}_i = \alpha + \beta_{0,i} + \beta_i R_M + \epsilon_i \tag{17}$$

where,

$\hat{R}_i$: predicted returns of asset i

α: intercept of regression

R_M: market returns

ϵ_i: zero mean random error term

2.3.2 Beta features

Depending on the values of the parameter β, four types of titles can be distinguished:

$\beta \approx 1$: similar fluctuations between market and securities

$\beta \approx -1$: rare values but really wanted by investors

$\beta > 1$: extremely volatile or "aggressive"

$0 < \beta < 1$: low volatile or "defensive"

2.3.1 Disadvantages of beta's coefficient

The following is a series of drawbacks related to the parameter β:

- The value that this parameter takes depends on the assets chosen to do the linear regression. (See equation 17)
- It is known as an unstable parameter, that is, its stability increases in function of decreasing the period of observations.

[1] The tangent portfolio gives us the tangent point that unites the risk free interest rate with the efficient frontier, providing the highest profitability with the minimum risk. See section 3.3.2 "Capital Market Line"

3. Classical Theories of Portfolio Investment

3.1 H. Markowitz portfolio selection

The basis of the model developed by Markowitz [1] is based on the fact that the behavior of every rational investor is to seek profitability by rejecting risk, therefore, for the investor, an efficient portfolio is which achieves the highest possible return for a given level of risk. This model is based on a series of hypotheses which we specify below:

- The performance of any portfolio or asset is considered a random variable whose probability distribution is known to the investor, and where it is performance will be calculated through it is mathematical expectation.
- All the investments analyzed must have the same period of time, ie cover an instant T of time.
- All assets that make up the portfolio are known.
- All the assets that make up the portfolio will be risky assets.
- The measure of risk is the standard deviation of the returns.
- Short positions are not allowed, short selling, that is, credit sales or short positions are not allowed, implying that all proportions invested in each asset, or better known as portfolio weights, w , will be positive or null.
- Each investor with rational behavior will prefer to invest in assets or a portfolio that will return greater profitability to a certain level of risk.

3.1.1 Model approach

$$Minimize: \quad \sigma^2 = X^T \Sigma X$$

$$Maximize: \quad E_p = \sum_{i=1}^{n} X^T E_p$$

$$constraints: \quad \sum_{i=1}^{n} X_k = 1$$

$$\forall i \epsilon [1, 2, .., n]$$

$$X_i > 0$$

where:

σ_p: portfolio's variance

X_i: proportion of investor budget allocated to the asset i

X: weight proportions vector in each asset i

X^T: weight proportions transposed vector in each asset i

Σ: matrix of variances and covariances of annualized returns

E_P: expected portfolio p returns

E_i: expected asset i returns

But the most important virtue of Markowitz's model was to show that it is not the risk of an asset what should be matter to the investor but the contribution that such asset makes the portfolio risk. The risk of a portfolio depends on the covariance between the assets and not the average risk of them. Therefore, their precise estimation is fundamental in determining the efficient portfolio in the mean-variance model, since it contains the information about the volatility of the assets, as well as between them.

3.1.2 Review of Markowitz's Mean-Variance Model

In spite of the great advance of Markowitz's model supposed to consider the portfolio of investment as a whole, academic community was aware of the revolution caused by the idea exposed by Markowitz began to analyze it in detail looking for it is weaknesses.

The main criticisms of the model are:

- Are investors as rational as the model implies? Is it possible that despite being so rational, this rationality is not able to shape the model?
- Is the variance the appropriate measure of risk ?
- Another criticisms of the model is that, it does not consider the volatility of a financial series, assuming that the variance is constant over time, that is to say, homocedastic, but on the contrary is very frequent the heterogneity, that is to say, the variance has systematic changes in the time.

Initial development of model there had a serious technical and economical problem: calculating the efficient frontier was very costly and timeconsuming. Nevertheless, as we will begin to see in the following section, this difficulty ended giving rise to a much more famous and useful model.

3.2 Willian F. Sharpe: A simplified model for portfolio analysis.

The main drawback of Markowitz's Mean-Variance model lay in it is practical performance, since it housed a large amount of calculations to be carried it out. Seeking to simplify the Markowitz's model, Dr. William F. Sharpe (1970) [2] published a new model proposing the referred simplification known as *"Sharpe Diagonal Model"*[2]. This model did not propose not only this simplification, but also an important innovation, the ability to simplify the total risk of any asset and any given portfolio. The model is also able to classify these securities according to the impact they have on their expected return, in addition, Sharpe observed that the securities that make up a portfolio were subject to certain common influences between them, reasoning that yields between portfolio assets would be related to a general index, and that the relation between the assets derives from it. This model is expressed as follows:

3.2.1 Model approach

$$Mazimize: \quad E[R_p] = \sum_{i=1}^{N} X_i E[\alpha_i + \beta_i R_{mt}] = \sum_{i=1}^{N} X_i \alpha_i + \beta_i E[R_{mt}] + \mu_{it}$$

$$constraint: \quad \sigma_p^2 = \beta_i^2 \sigma_{mt}^2 + \sum_{i=1}^{N} X_i^2 \sigma^2(\mu_{it})$$

$$\sum_{i=1}^{N} X_i = 1$$

[2]Sharpe was 26 years old when he wrote this article which by itself would have made him one Of the financial academics who have most influenced the stock market by enabling the practical development of portfolio selection theory, but it was his subsequent development of the CAPM that led him to share his Nobel Prize in 1990. The name "Diagonal model" refers to that in the variance-covariance matrix, the latter are equal to zero, in the Sharpe model, leaving only the variances that form the diagonal of that matrix.

where,

$$X_i > 0 \ \ \forall \ i = 1, 2..., N$$

R_p: returns of portfolio p

X_i: proportion of investor budget allocated to the asset i

α_i: equity's return of asset i not explained by the market

$\beta_i = \frac{Cov(R_i, R_m)}{\sigma_m^2}$: the ratio of the covariance of the return of asset i and the return of the market index to the variance of the return of the market index

R_{mt}: return on the market index m over the period t

R_i: asset i returns

μ_{it}: a zero mean random error term

σ_p^2: variance of portfolio p

σ_m^2: variance of index market p

3.3 Capital Asset Pricing Model (CAPM)

As a result of the research of Markowitz and Sharpe, among others, a theoretical model arises that relates the problems of profitability and risk assets through the concept of diversification, known as *Capital Assets Pricing Model (CAPM)*.

CAPM model tries to determine the intrinsic value of the assets, thus help to determine the expected return of them as well as their standard deviation. The model assumes that the performance of an asset will depend on the standard deviations (risk), both concepts will be related between them through beta's coefficient (β).

At least, two inputs are required in order to use mean-variance optimization for portfolio construction. They are *expected return forecast* and *return covariance matrix*. Additional inputs are practical constraints that are required for realistics portfolios, i.e. limits on stock holdings and/or sector weights.

3.3.1 CAPM model assumptions

1. Investors are risk averse and try to maximize their expectaction useful.
2. All investors have homogenous subjective expectations for the probability distributions of future returns. Homogeneous expectations regarding covariances as well.
3. The market rewards people for assuming unavoidable risk, but there is no reward for needless risks due to inefcient portfolio selection. Therefore, the risk premium on a single security is not due to it is "standalone" risk, but rather to it is contribution to the risk of the tangency portfolio.
4. Assets are perfectly divisible.
5. There are no taxes or transaction costs.
6. The market is efficient, which means stock prices reflect all available information by quickly adjusting to new information. See chapter 5, section 5.1

3.3.2 Capital Market Line (CML)

The *capital market line (CML)* [6] relates the excess expected return on an efficient portfolio to it is risk. Excess of expected return is the expected return minus the risk-free rate and is also called the *risk premium*. Given this new situation, where risk-free investment comes into play, the set of efficient portfolios are defined by the tangent line (CML) from the risk-free interest rate (Rf), and the tangent point (tangent portfolio) with the efficient frontier (plot in green above), and not by the curve of the efficient frontier itself. Therefore, all investors will choose to be located at tangent portfolio because the utility is maximized, that is, there will be no other point that governs their choice.

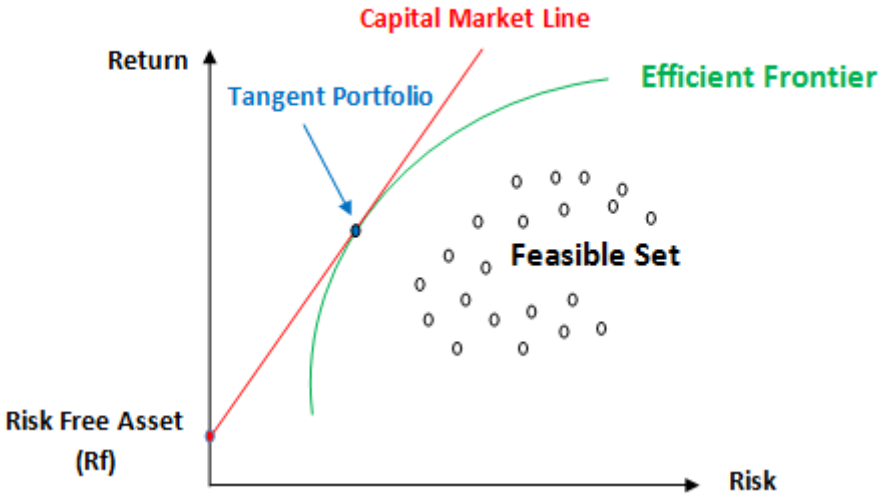


Figure 1: Shows an auto-ilustration with all concepts defined in this chapter

The concepts shown in Figure 1 are denoted as:

1. The region under the hyperbola is known as a **feasible set** that collects all possible combinations of N risky assets for all existing portfolios.
2. The green hyperbola is the **efficient frontier**, given a set of assets it is possible to combine them for a certain level of risk obtaining the greatest profitability.
3. **Tangible portfolio**, is the combination of assets such that maximize profitability with the lowest posible risk.
4. **Risk-free asset** it refers to the risk free asset rate.

This means that everyone will devote part of their budget to the title without risk, and another to the risky portfolio. The latter will be called *tangent portfolio*, since, by investing all a part of it is wealth in it (namely, the portion invested in risky securities), it will represent the totality of the securities with risk existing in the market, and the weight that each one has in it.

The CML equation is:

$$E_p = R_f + r\sigma_p \tag{18}$$

where,

E_p: portfolio expected return

R_f: risk free rate

r: slope of CML

σ_p: risk of portfolio

The slope of the CML is, of course:

$$\frac{\mu_M - \mu_f}{\sigma_M} \tag{19}$$

which can be interpreted as the ratio of the risk premium to the standard desviation of the market portfolio.

If we change the above equation into CML equation we obtain:

$$\mu_R = \mu_f + \frac{\mu_m - \mu_f}{\sigma_m}\sigma_R \tag{20}$$

where,

$\mu_R = E(R)$: is the return on a given efficient portfolio

μ_f: is the risk-free rate

$\mu_m = E(RM)$: return on the market portfolio

σ_M: is the standar desviation of R_M

σ_R: is the standar desviation of R and R is the return on a given effcient portfolio

μ_R-μ_f: is the risk premium of R

μ_M-μ_f: is the risk premium of the market portfolio

The CAPM says that the optimal way to invest is to:

1. decide on the risk σ_R that you can tolerate, $0 \leq \sigma_R \leq \sigma_M$.[3]
2. calculate $w = \frac{\sigma_R}{\sigma_M}$
3. invest w proportion of your investment in an index fund, that is, a fund that tracks the market as a whole.
4. invest $1 - w$ proportion of your investment in risk-free Treasury bills, or a money-market fund.

Alternatively,

1. choose the reward $\mu_R - \mu_f$ that you want; the only constraint is that $\mu_f \leq \mu_R \leq \mu_f$ so that $0 \leq w \leq 1$[4]
2. calculate:

$$w = \frac{\mu_R - \mu_f}{\mu_M - \mu_f} \tag{21}$$

3. do steps 3 and 4 as above.

[3] In fact, $\sigma_R \geq \sigma_M$ is possible by borrowing money to buy risky assets on margin

[4] This constraint can be relaxed if one is permitted to buy assets on margin

One can view $w = \frac{\sigma_R}{\sigma_M}$ as an index of the risk aversion of the investor. The smaller the value of w the more risk-averse the investor. If an investor has $w = 0$, then that investor is 100% in risk-free assets. Similarly, an investor with $w = 1$ is totally invested in the tangency portfolio of risky assets.[5]

3.3.3 Security Market Line (SML)

The *security market line (SML)* [6] relates the excess return on an asset to the slope of its regression on the market portfolio. The SML differs from the CML in that the SML applies to all assets while the CML applies only to efficient portfolios.

SML consequences are:

1. The expected return on an asset does not depend on it is total risk but depends only on the risk portfolio correlated with the portfolio of the market.
2. The beta measures the risk of an asset that can not be diversified.
3. To estimate the expected rate of return on an asset, we only need to estimate it is beta it.

Suppose that there are many securities indexed by j.

Let denotes, σ_{jM} as the covariance between the returns on the j_{th} security and the market portfolio and, $\beta_j = \frac{\sigma_{jM}}{\sigma_M^2}$ as the slope of the best linear predictor of the j_{th} security's returns using returns of the market portfolio as the predictor variable.

Using the CAPM, *security market line (SML)* can be written as follow:

$$\mu_j - \mu_M = \beta_j(\mu_M + \mu_f) \tag{22}$$

SML impplies some limitations which explain below:

1. The CAPM predicts the expected value of the profitability of the title, if the market is in balance.
2. We must estimate how risk-free is at the present time.
3. It is necessary to calculate the market expectation as the beta value of the title.
4. The CAPM rests on very restrictive hypotheses: absence of taxes, investment in a single period, homogeneity of expectations, aversion to the risk of all investors, economic rationality...

Consider what would happen if an asset exists and it is find below the SML. Investors would not want to buy it because it is risk premium is too low for the risk given by its beta. They would invest less in this asset and more in other securities. Therefore, the price would decline and after this decline it is expected return would increase. After that increase, the asset would be on the SML, or so the theory predicts.

3.3.4 Comparison of the CML with the SML

The *CML* applies only to the return R of an efficient portfolio. It can be arranged so as to relate the excess expected return of that portfolio to the excess expected return of the market portfolio:

$$\mu_R - \mu_f = (\frac{\sigma_R}{\sigma_M})(\sigma_M - \sigma_f) \tag{23}$$

[5] An investor with $w > 1$ is buying the market portfolio on margin, that is, borrowing money to buy the market portfolio

The *SML* applies to any asset and like the CML relates it is excess expected return to the excess expected return of the market portfolio:

$$\mu_j - \mu_f = \beta_j(\mu_M - \mu_f) \tag{24}$$

If we take an efficient portfolio and consider it as an asset, then R and j both denote the expected return on that portfolio/asset. Both equation below hold so that:

$$\frac{\sigma_R}{\sigma_M} = \beta_R \tag{25}$$

4. Data processing

In this chapter, and before starting with the challenge, we will detail the entire process of collecting, and processing used database in this research, which has been one of the most difficult part of the studys. First, getting free financial data is not easy, since there are many platforms that serve financial data by paying significant amounts of money to financial institutions themselves, traders, brokers or people who are dedicated to the quantitative study of markets. On the other hand, one of the most used free website platforms[6] for the collection of data by users of R-Studio software has stopped serving this kind of data through the usual package that did it. To achieve it, we have used several libraries of different statistical packages of the software mentioned above, in addition, we have had to establish links with different websites to download data through them directly.

The first statistical package used has been a very recent library, integrated in the software a few months ago between April 17 and July 17. *Tidyquant package*[7] integrates the best resources for collecting and analyzing financial data, zoo, xts, quantmod, TTR, and PerformanceAnalytics, with the tidy data infrastructure of the tidyverse allowing for seamless interaction between each. You can now perform complete financial analyzes in the tidyverse.

In particular, we have used two of it is more specific functions that we show below:

- Key.ratios: Get 89 historical growth, profitablity, financial health, efficiency, and valuation ratios that span 10-years from Morningstar.
- Key.stats: Get 55 real-time key statistics such as Ask, Bid, Day's High, Day's Low, Last Trade Price, Current P / E Ratio, EPS, Market Cap, EPS Projected Current Year, EPS Projected Next Year and Many more from Yahoo Finance.

In addition, and in order to complete a good database in terms of fundamental factors that can provide greater precision, we have created a connection to a website known as "www.sistematicportfolio.com"[8] where through a function we managed to download as many fundamental factors as "EBIT", "NetEBITDDA", "Price To Earning", "Dividend Yield", "ROE", etc. which will be explain below.

Finally, we have been able to create a database with 101.550 operations from STOXX600[9]under a time horizon of 16 years, specifically, from 29/12/2000 to 29/07/2016 with monthly observations, and classified by the type of sector to which they belong. Below we will explain in detail what factors we have achieved and the procedure to achieve this.

A brief description of each variable in dataset is show in table 1 below:

[6]https://www.yahoo-finance.com

[7]https://cran.r-project.org/web/packages/tidyquant/tidyquant.pdf

[8]http://www.systematicportfolio.com/sit.gz

[9]The STOXX Europe 600 or STOXX 600 is a stock index composed of the top 600 companies by European market capitalization (90% total capitalization of the European equity market). The countries that compose the index are Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Holland, Norway, Portugal, Spain, Switzerland, Sweden and United Kingdom. This index is the benchmark for Exchange-Traded Funds (ETF) exchange.

Variable	Description
Date	Date on which the operation occurred
Symbol	ISIN NUMBER for each observation
Sector	Industrial sector
Universe Returns	Return generated by each security in next period
Benchmark Weight	Weight that has the security on the reference index
Country	Original country of the symbol
Market Cap	Market capitalization of the company
Adjusted price	Adjusted price to dividends, splits...referring to the stock
Earning Yield	Finantial ratio that allows us to give a more informative reading on the relationship between profitability, income and time, for some market value
Dividend Yield	Financial ratio that shows, in percentage, the ratio between dividends per share distributed by a company in the last year and the equity price
Momemtum	Finantial indicator that shows difference between today's price close and N days ago
NetDebt EBITDA	Finantial debt ratio that shows how many years it would take to pay back its debt if net debt and EBITDA are held constant
ROE	Finantial ratio measures the return that the shareholders obtain from the funds invested in the company, that is, it tries to measure the capacity that the company has to pay its shareholders
Beta	Measures the degree of variability of the profitability of an action with respect to the average profitability of the "market" in which it is negotiated
Volatility	Variability of the profitability of an action with respect to its average in a determined period of time. When that volatility is compared to market volatility it is called beta

Table 1: Brief description of each variable in dataset.

Next, the first step of data set processing has been how to perform a classification of the operations depending firstly, on the sector to which they belong but always taking into account the periods existing within them, ie. for each sector have been classified all operations in each period. And second, and always depending on previous classification, we have calculated for each of the 8 factors previously explained, the probabilities of belonging to the 5 quintiles we have decided to create in this research. The classification was as follows:

- Top quintile: value that leaves to it is left the 20% of the best cases.
- Quintile 2: reference the value that leaves to its left the 40% of the highest values.
- Quintile 3: refers to the value left to the left of 60% of the highest values.
- Quintile 4: reference the value that leaves to its left the 80% of the highest values.
- Bottom quintile: one that leaves the lowest 20% values classified.

To clarify, on the top quintile will fall the highest values of each factors, the best ones. It should be noted that there are 3 factors whose highest values do not refer to a better behavior, the opposite, and *Beta, Volatility and Net Debt EBITDA*, in the next step we will make a specific treatment of these three factors.

Finally, before showing a small example of how the dataset would remain after such classification, we need to invert the values of the 3 factors named above, since, this gives the proper meaning to research, because the lower values to volatility, beta and netdebt factors, the higher yield.

Therefore, once all these steps are ejected, the database remains such that in the following example:

Date	Symbol	Sector	Uni.Ret	MarketCap	q	EY	q	DY	q	Momtm	q	NetD	q	ROE	q	Beta	q	Vol	q
29/12/2000	FR0000121261	Auto	6.35	5193.3	4	0.1	3	2.2	3	0.84	3	1.66	5	12.37	3	0.3	2	0.017	1
31/01/2001	DE0005439004	Auto	1.9	2358.61	5	0.1	2	2.74	2	1.03	3	1.2	4	11.84	4	0.631	3	0.019	1
28/02/2001	IT0000088457	Auto	-5.79	7554.43	3	0.04	5	2.30	3	1.6	1	-1.75	1	11.042	4	0.724	4	0.018	1
30/03/2001	FR0000131906	Auto	1.13	13740.45	2	0.11	2	1.86	5	1.35	1	0.5	3	13.83	2	0.75	3	0.024	5
30/04/2001	SE0000382335	Auto	-3.072	2201.6	5	0.079	3	2.177	3	0.63	5	1.37	4	7.73	5	0.74	3	0.02	4
31/05/2001	DE0007664005	Auto	-2.67	18155.8	2	0.104	2	2.12	4	1.3	2	1.1	4	17.1	2	1.01	5	0.02	3
29/06/2001	SE0000382335	Auto	16.86	1984.53	5	0.09	3	2.49	3	0.938	5	1.35	4	7.7	4	0.83	4	0.024	4
31/07/2001	DE0007664005	Auto	-7.94	16526.6	2	0.12	1	2.4	3	1.2	2	1.01	4	18.1	2	1.01	5	0.02	3
31/08/2001	DE0007664005	Auto	-19.732	15213.3	2	0.13	1	2.6	2	1.07	2	1.09	4	18.22	2	1.05	5	0.02	2
30/09/2002	DE0007100000	Auto	18.53	33180.34	1	0.07	5	6.96	1	1.004	3	0.14	3	6.34	5	1.19	5	0.02	3
31/10/2002	FR0000121261	Auto	-5.81	4745.31	3	0.11	3	2.78	3	0.75	4	2.08	5	12.27	3	0.86	2	0.02	1
29/11/2002	FR0000121261	Auto	-9.61	4659.31	3	0.12	3	2.96	3	0.9	3	2.10	5	12.09	3	0.88	2	0.02	1
31/12/2002	FR0000121261	Auto	1.346	4211.2	3	0.136	3	3.35	3	0.7	3	2.10	5	12.17	3	0.89	2	0.02	2
31/01/2003	FR0000121261	Auto	-16.27	4267.961	3	0.13	3	3.376	3	0.65	3	1.81	5	12.35	4	0.90	2	0.02	3
28/02/2003	FR0000121261	Auto	31.54	3573.17	3	0.16	3	4.102	3	0.69	3	1.76	5	12.46	3	0.97	2	0.02	3
31/03/2003	FR0000121261	Auto	-3.6	4700.43	3	0.12	3	3.11	4	0.618	2	1.72	5	12.14	4	1.08	3	0.02	3
30/04/2003	FR0000121261	Auto	9.67	4395.575	3	0.13	3	3.35	4	0.7	2	1.73	5	11.95	3	1.09	3	0.02	3
30/05/2003	FR0000121261	Auto	-5.235	4820.95	3	0.12	3	3.08	3	0.75	2	1.68	5	12.06	3	1.06	3	0.02	3
30/06/2003	FR0000121261	Auto	14.77	4619.93	3	0.1	2	3.28	3	0.84	3	1.7	5	12.05	3	1.05	3	0.02	3
31/07/2003	FR0000121261	Auto	-13.65	5302.45	3	0.11	2	2.78	3	0.88	2	1.72	5	12.65	3	1.03	3	0.028	3
29/08/2003	FR0000121261	Auto	5.69	4578.348	3	0.13	2	3.15	3	1.30	1	1.70	5	12.81	3	1.06	3	0.02	3
30/09/2003	FR0000121261	Auto	-1.92	4839.312	3	0.12	2	2.98	3	1.08	2	1.72	5	12.76	3	1.0	3	0.02	4
31/10/2003	FR0000121261	Auto	9.90	4746.11	3	0.12	2	3.05	3	0.98	3	1.70	5	13.00	3	1.27	3	0.02	3
28/11/2003	FR0000121261	Auto	6.12	5216.42	3	0.119	2	2.80	3	1.00	5	1.71	5	13.14	3	1.1	3	0.023	4
31/12/2003	FR0000121261	Auto	0.49	5536.173	3	0.1	2	2.71	3	1.25	3	1.66	5	12.91	3	1.021	3	0.022	3
30/01/2004	FR0000121261	Auto	-5.00	5563.417	3	0.11	2	2.72	3	1.31	4	1.565	5	12.68	3	1.011	3	0.02	4
27/02/2004	GB0030646508	Auto	-2.74	2643.323	5	0.08	4	4.83	1	1.70	3	1.65	5	14.91	2	1.15	3	0.02	4
31/03/2004	GB0030646508	Auto	1.714	2488.36	5	0.08	4	5.108	1	1.17	3	1.63	5	14.62	2	1.31	4	0.019	4
30/04/2004	GB0030646508	Auto	8.36	2531.385	5	0.09	4	5.09	1	1.11	5	1.776	5	15.83	2	1.37	5	0.019	5

Table 2: Data set example.

5. Challenge to classic investment theories

Having presented the fundamentals of investment, portfolio management and the most influential investment portfolio theories, in this section we are going to challenge them, focusing especially on how the equity markets premium can be achieved investing in securities with low volatility, that is, how to optimize the exposure of our portfolio of equity assets in periods with great uncertainty.

The Principles of Classical Finance as Markowitz, *Modern Portfolio Theory (MPT) (1952)* [1], *Capital Asset Price Model (CAPM-1964)* [2] and Lintner (1965) [7] argue that assets with high risk deliver, on average, higher performance as compensation for maintaining that risk. Surprisingly, low-risk ($\beta < 1$) portfolios seem to contradict this more basic conclusion of traditional finance. Low volatility investing provides a way to invest more defensive naturally in the market crashes while being able to capture most, if not all, market rises. In other words, the expected returns of all portfolio securities are aligned with their market betas, with the slope of the straight line being equal to the equity risk premium (ERP), as shown in the following figure.

However, empirical studies, initiated in *Black, Jensen and Scholes (1972)* [8], show that the actual SML is, on the other hand, flatter than the CAPM SML predicts, sometimes even being an inverted line with a slope descendant. In other words, the low-beta portfolios exhibit empirically higher risk-adjusted returns than the high-beta portfolios. For this reason, the empirical flatness of the stock market line (SML) is commonly known as the **"low risk anomaly".**[10]

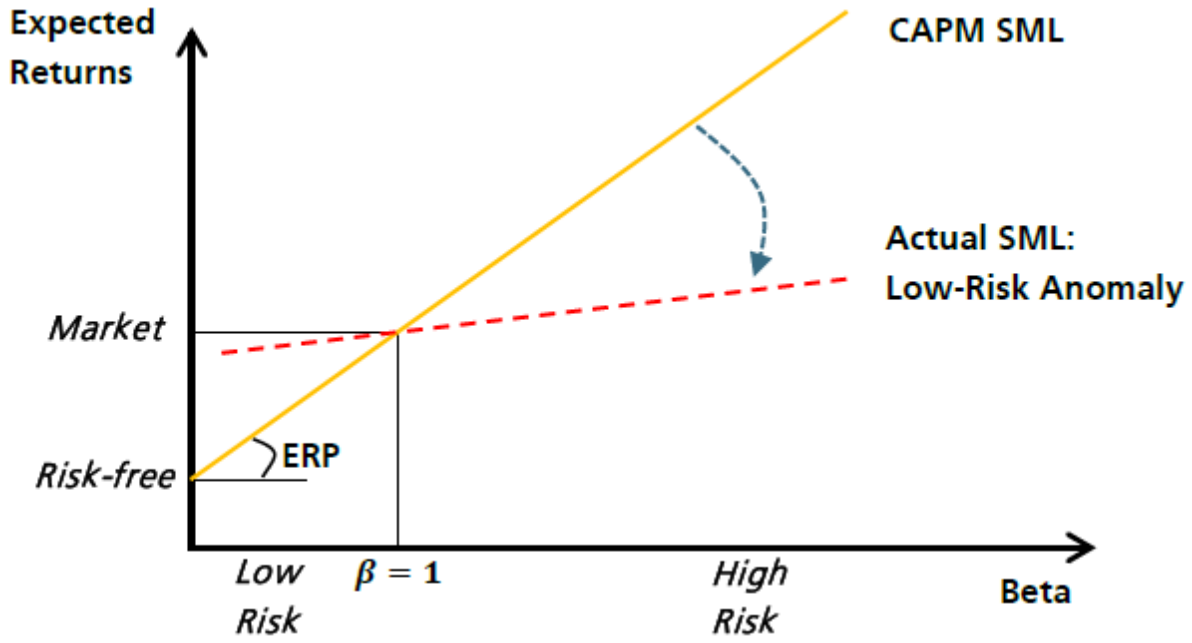


Figure 2: CAPM vs Low-Risk Anomaly. SOURCE: UBS Quantitative research, "Low-risk investing: perhaps not everywhere (2016)"

[10]The terms *low-risk,low-volatily* are often used interchangeably

5.1 Low-Risk Investing

The outperformance of low beta/volatility stocks over high beta/volatility stocks is one of the biggest challenges to CAPM, which would argue that the opposite would hold. The anomaly was first identified in 1972 when Black, Jensen and Scholes [8] reported that a portfolio constructed by being long low beta stocks and short high beta stocks generated positive returns. Fama and MacBeth (1973) [9] extended this work and showed that the relationship between risk and return is generally too flat compared to theoretical expectations. The debate surrounding this anomaly has continued unabated, with a number of explanations proposed. Unsurprisingly, all explanations offered for one of the biggest challenges to the efficient markets hypothesis (EMH) are behavioural-based. [15]

The efficient markets hypothesis (EMH) affirm that finantial market asset prices reflect all existing information and are fully and rapidly adjusted to new emerging data. For information it understand any news that may determine or affect the price of shares. In 1970, Fama published a paper dealing with both theory and evidence of these hypothesis. The paper complemented the theory giving complete and refined approach including definitions of three forms of market efficiency: weak, semi-strong and strong.

- Weak form efficiency: market has weak efficiency when stock prices incorporate only the information contained in past prices.
- Semi-strong efficiency: market has semi-strong efficiency when current stock prices incorporate past prices and information about the company, accounting information basically.
- Strong efficiency: market presents a strong level of efficiency when the current stockholders incorporate all public and private information, and also generates future expectations.

The lower level of demand in these stocks will by extension depress their price, increasing the upside potential for investors. Related to this is the argument that stocks that are consistently in the news (typically high volatility stocks) are generally overbought and thus deliver a lower average return. The mechanism by which this anomaly operates appears to be associated with the cross-sectional dispersion of beta across cycles [11]. In bear markets[11], markets are more volatile and betas more disperse. As a result, investing in low beta stocks forms a buffer against falling markets. Conversely, during bull markets[12],beta spread is tighter and, as a result, while high beta stocks outperform, the outperformance is in aggregate limited.

Consistent with the research mentioned above, Figure 3 shows the return of a theoretical portfolio in which we are consistently long the top decile of low-volatily stocks (rebalanced monthly) and short the highest beta stocks, from 2001 until 2016. The performance of the portfolio highlights an interesting result from low-vol investing which will be shows in following section.

Next, we introduce "time varying betas", the low-risk portfolios are affected to quantitative factors such value, quality and momemtum, which are known to fluctuate intensely over time. We will see how companies with high risk are largely affected by negative market returns and otherwise companies with low risk overperformance the own benchmark. The aim of our study is how to balance the exposure to those fundamental factors to achieved it.

[11] A bear market is a financial market of a group of securities in which prices are falling or are expected to fall
[12] A bull market is a financial market of a group of securities in which prices are rising or are expected to rise

5.1.1 Time-varying betas

To introduce the time varying betas [18] it is necessary to understand that low risk portfolios are directly related to quantitative factors such as value, size and momemtum, which fluctuate intensely. In other words, we will see how high risk companies are greatly affected in bear markets, and on the other hand, low risk companies are greatly favored. It is the object of our research in this section to show this relationship with the factors previously mentioned and show how to balance the exposure to them in order to achieve the best results.

It should be noted that there are many reasons why the beta coefficient of the market fluctuates over time even when the price of the underlying asset forming the portfolio remains unchanged. In particular, and related to what was presented in the previous section, first, we can confirm that a sudden fall in the stock market will increase the leverage (debt to equity capital ratio) in the balance sheet situation of the companies. Therefore, in periods of crisis where the market falls, it could be expected that the betas of the most leveraged companies will increase more than the non-leveraged ones.

On the other hand, these betas can also undergo a change or fluctuation depending on the news about the cash flows or discount rates of companies. That is, when the market is destressed volatility does not fluctuate as much as in those stressed markets where news of this type predominates.

Figure 3: High and low beta portfolios weighted about benchmark

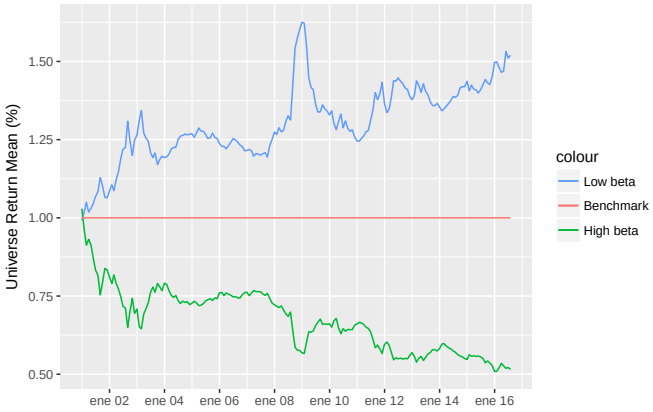


Figure 4:Shows all beta portfolios and benchmark performance

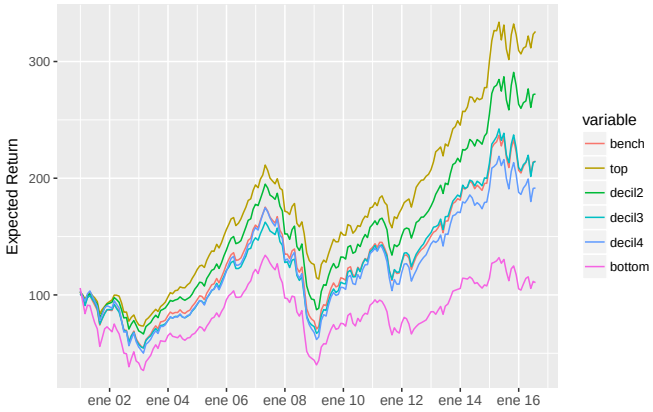


Figure 3 shows the low volatility portfolios (named as Top in blue) and high volatility (named as Bottom in green) weighted against the benchmark. We see first period between 2001 and 2003 and just in the period before the crash of the global stock markets in March 2009 due to the bankruptcy of Lehman Brothers, while the cycle remained bearish, portfolio low volatility have a better performance than the benchmark. The meaning of this behavior is that since the beta of this portfolio is low, this means that it moves less than the benchmark so when the benchmark falls, this portfolio performs much better. Exactly, if we look at table 3 where we see the results obtained, it makes 1.80% better than the market, while it obtains a volatility of 12.63%, obtaining a final variation of 1.51%. However, the portfolio of high volatility we see as worse than the benchmark due to the same reason as before but in reverse, moving more than the benchmark when the market falls, this portfolio does even more. Figure 4 shows the different portfolios under different levels of volatility against the benchmark, and how they have behaved throughout the period.

	Low Beta	High Beta
Average Excess Return	1.805153	-2.107953
Annualized Volatility (%)	12.63714	25.09011
Shape's Ratio	0.1428451	-0.0840153
Variation over time	1.519812	0.5156497

Table 3: Performance statistics of Low-beta and High-beta portfolios

6. Exposure to fundamental factors

As we have been saying in previous sections, portfolio profitability has always been related to portfolio exposure with beta market and alpha source, but then we will see that it is not just the relationship, it is also directly related to the exposure to other fundamental factors [18]. This relationship is known as *"risk factor premium".*

The existence of the risk factor premium can be explained in two ways: *explanations based on risk, and behavourial vias explanation.* When we refer to risk-based explanations, it is asserted that a premium for those investors who take on additional risks by exposing themselves to a particular factor. However, such explanations based on behavioral bias assert that the factor premium is due to systematic errors made by the investor on the basis of their biases, that is, they may have on reactions and under reactions.

For such explanations to be truly relevant, it would be necessary to accept that they include those known as arbitrage limits, ie. market characteristics such as short selling restrictions or liquidity constraints, which generally prevent investors from taking advantage of opportunities arising from irrational behavior from another ones. Therefore, a risk factor based on a strong assumption is more likely to be consistent in the future, therefore, it is much more reliable for an investor to have an explanation based on risk than on behavior.

6.1 Rationale behind risk factors

There are 4 well known equity factors driven by different characteristics.

Fundamental Factors		
Factor	Rationale	Characteristics
Value	Value stocks carry some risks Risk premium is interpreted as compensation for the greater risk they assume. Investors often over-estimate this risk	Performance highly dependent on cycle Low correlation with defensive factors Better performance under wide dispersion of estimates
Momentum	Past outperformers tend to keep doing it Difficult implementation due to possible extra charge generated Slow reaction of some investors to new information	Consistent positive performance but severe drawdowns near cycle turning points Reward to winning investors High turnover in factor values over very short time periods
Quality	High profitability and low leveraged companies tend to outperform in a defensive environment More economic than finantial factor Slow reaction of some investors to new information	Outperformance in tougher markets Negative correlation with value Low correlation with other risk premia
Low Risk	Overperformance of portfolios with low volatility Trend for investing in securities with high volatility increase low volatility performance Cross-sectional dispersion across economic cycles	Low volatility Beta coefficent fluctuates overtime even underlying price keeps unchanged High low vol portfolio take advantage of negative market cycles to overperformance

Table 4: Fundamental characteristics and rationale foundamentation factors.

6.1.1 Value

Zhang (2005) [10] argues that the risk premium for value compamies is based on the difficulty implicit in an investment in reversing it is trend, that is, is not easy for a value company to change the cycle of their investment. The great value of value companies is usually their portfolio of tangible assets, which are difficult to liquidate compared to intangible assets as equity securities. Therefore, values companies are closely related to the economic cycle in which the market is, when the economic cycle is good, value factor works very well but when the cycle is bearish, they behave badly.

6.1.1.1 Dividend yield factor

The dividend yield is a financial ratio that measures the amount of cash dividends distributed to common shareholders relative to the market value per share. Dividend yield is a way to measure how much cash flow you are getting for each dollar invested in an equity position.

The dividend yield is used by investors to show how their investment in stock is generating either cash flows in the form of dividends or increases in asset value by stock appreciation. Investors invest their money in stocks to earn a return either by dividends or stock appreciation. Some companies choose to pay dividends on a regular basis to spur investors' interest. These shares are often called income stocks. Other companies choose not to issue dividends and instead reinvest this money in the business. These shares are often called growth stocks.

The dividend yield formula is calculated by dividing the cash dividends per share by the market value per share:

$$DY = \frac{Cash \;\; Dividends \;\; per \;\; share}{Price \;\; per \;\; share}$$

In Figure 5 we can see the high and low dividend yield portfolios weighted against the benchmark where, after the 90's crash (not observable in the graph) of the technology companies, the economic cycle changed and began the upward cycle that made highest dividend portfolio had a very visible outperformance during the years prior to the economic crisis, however when the market suffered the fall of 2009 this portfolio was also very affected. After this period it has behaved very similar to the benchmark itself, which can be interpreted by the lack of interest that this factor contributed to the investors. On the other hand, the lowest dividend yield portfolio only gives us the impression of a constant low performance throughout the time horizon, exactly obtaining an expected average performance of 3.67% below the benchmark, while the high dividend yield portfolio did it 2.04% better than the benchmark itself. If we look at the volatility of both portfolios we can see that there is no big difference, both are around 19.50%. See table 5.

Figure 5: High and low dividend portfolios weighted about benchmark

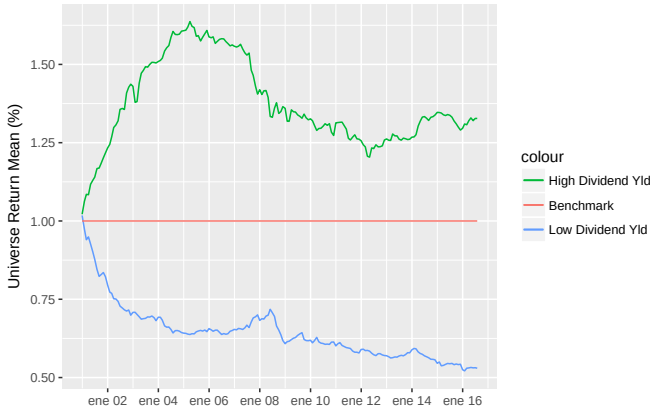
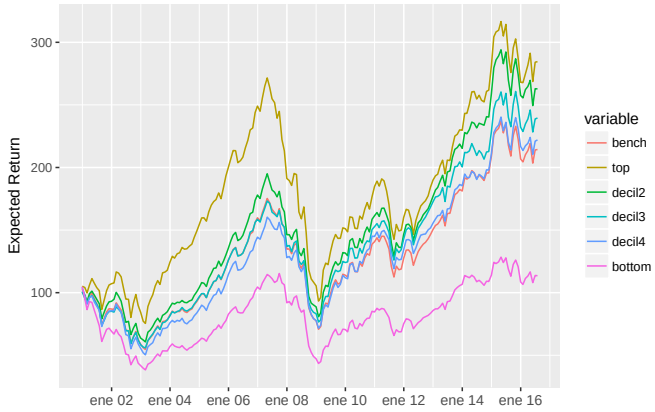


Figure 6:Shows all dividend portfolios and benchmark performance



	High Dividend	Low Dividend
Average Excess Return	2.048019	-3.671595
Annualized Volatility (%)	19.19938	19.89995
Shape's Ratio	0.1066711	-0.1845027
Variation over time	1.327507	0.5297714

Table 5: Performance statistics of dividend yield

6.1.1.2 Earning yield factor

Earnings yield are the earnings per share for the most recent 12-month period divided by the current market price per share. The earnings yield (which is the inverse of the P/E ratio) shows the percentage of each dollar invested in the stock that was earned by the company. The earnings yield is used by many investment managers to determine optimal asset allocations.

It is important to note that earnings yield does not always represent cash available to the investor, because companies may choose to reinvest earnings rather than pay dividends to shareholders. Unlike the dividend yield, earnings yield is not dependent on management's capital-allocation decisions.

The formula for calculating earning yield may be represented as follows:

$$EY = \frac{Earnings \ \ per \ \ share}{Stock \ \ price}$$

The explanations of the figures below, figures 7 and 8, will be very similar to the yield of the dividends both in results and in the behavior of the portfolios whose construction is similar. Prior to the crisis, during the economic cycle the highest earning yield portfolio has the same overperformance as the highest dividend yield portfolio, then when the market fell over 2009 it fell with it, whereas after the crisis it behaved just like the benchmark itself, no need to talk about the low earning yield portfolio that is similar to the low dividend yield portfolio with it is expected average yield of 3.83% below the benchmark. The results attached to the high earning yield portfolio obtained an average expected return of 3.21% better than the benchmark itself, with a volatility of 21.23% and a sharpe ratio of 0.16. Results visible in table 6.

Figure 7: High and low earning yld portfolios weighted about benchmark

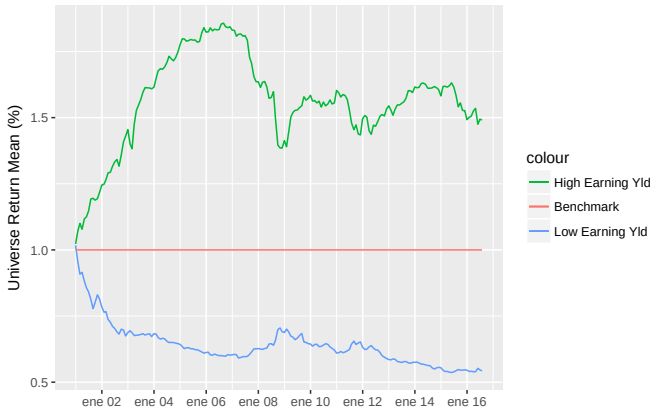
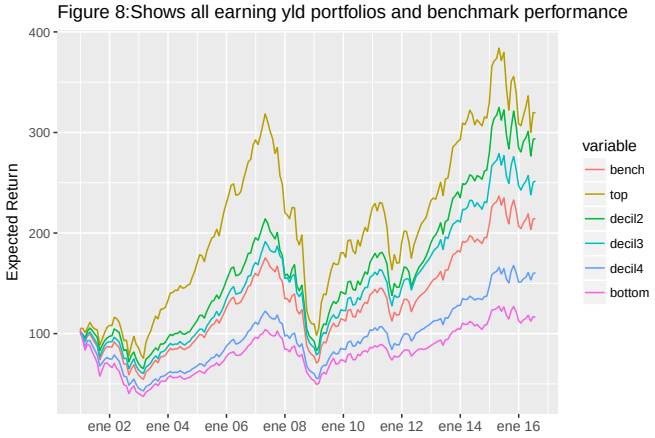


Figure 8:Shows all earning yld portfolios and benchmark performance



	High Earning Yld	Low Earning Yld
Average Excess Return	3.219673	-3.834897
Annualized Volatility (%)	21.23554	18.32467
Shape's Ratio	0.1676967	-0.2092751
Variation over time	1.4907	0.5427367

Table 6: Performance statistics of earning yield

6.1.2 Momemtum

The momemtum efect explains the continuity of patterns in the past in future yields of stocks. As a general rule, the most recent period should be excluded when deciding asset selection as it could lead to erroneous conclusions about the effects of the short-term reversal, Momentum is typically defined as the cumulative stock return over some prior time frame ignoring the most recent period of performance. Momemtum can be defined as stocks with high returns over the past 12 months omitting the last month versus stocks with low returns:

$$Momemtum = \frac{Stock \;\; Price \;\; 1 \; Month \;\; Ago}{Price \;\; 1 \; Year \;\; Ago}$$

In Figure 9 again we show low and high mometum portfolio weighted with the benchmark, we see how the highest momemtum portfolio is best one behaves in terms of expected profitability, this is due to the previously stated that the momemtum factor rewards the winners in addition to the subject on the behavior of the investors when gradually receiving the information that causes them to react by making this factor work very well, although it is also affected by the market crashes sspecially in the 2009 crash. This high-momemtum portfolio achieved an expected average return of 4.05% on the benchmark, with associated annualized volatility lower than the value factor of 15.21% and a sharpe ratio of 0.26 also higher than the factors previously shown.

On the other hand, the portfolio of low momemtum is very punished during the negative cycles of the market, we can see how it falls, however during the bullish cycles although it seems to remain bearish does it relatively to a lesser extent, suffering a variation at the end of time horizon of 0.38%, generating an expected average return of 4.53% below the benchmark and with an annualized volatility upper to none of the others only comparable to the high beta portfolio of 25.04%. These results can be checked in Table 7 below.

Figure 9: High and low momemtum portfolios weighted about benchmark

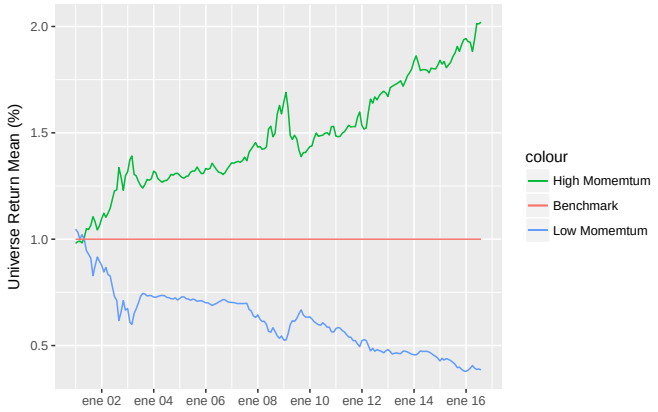
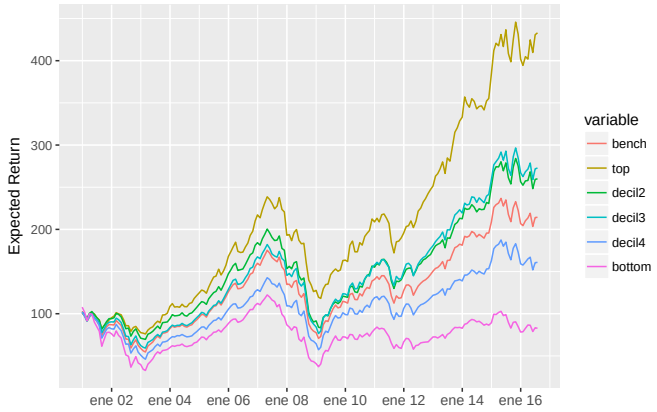


Figure 10:Shows all momemtum portfolios and benchmark performance



	High Momemtum	Low Momemtum
Average Excess Return	4.054444	-4.531791
Annualized Volatility (%)	15.21412	25.0414
Shape's Ratio	0.2664921	-0.180972
Variation over time	2.020327	0.3858493

Table 7: Performance statistics of Momemtum factor

The Momemtum factor is the most influential factor in what we see both visually and statistically in the ratios in the table above. But, what is this due to?. The momemtum factor has to do with the statistics of the securities that form the portfolio, which have a high turnover of their values even in really short (monthly) periods, and therefore, can be very difficult to implement because could generate an excess of costs. In this research we have not taken into account the cost, since it is not the object of it to build a real portfolio of assets, but if the following studies we want to take into account the cost, surely this momemtum factor would not do so well, since it is not comparable to having, in each period, to build a new portfolio because the securities have ceased to be profitable or influential due to this factor, that if for example we have to change a third of the portfolio if we look at another factor like the beta factor whose rotation is less usual in short periods. It is a good point to take into account, as I say, in the following studies.

6.1.3 Low risk

In this section we explain the low risk factor, it should be noted that in our study we have two low risk factors with beta and volatility, but previously we have exposed the beta factor since it came from an explanation about the time varying betas, so that in this section we will denote the volatility factor in detail, it is behavior and the results obtained, which on the other hand, follow the beta factor line.

Frazzini and Pedersen (2014) [11] provide a model where investors with liquidity constraints can invest in leveraged positions of low volatility stocks, but in turn are forced to liquidate those assets in bad times when their constraints no longer allow them to stay longer with said open position. For this reason, low risk assets are exposed to a liquidity crisis risk and are offset by this risk. If we rely on the behavioral explanations of the low risk premium, they rely on high risk stocks tending to have low returns because irrational investors raise their value beyond the rational, ie. they overvalue the price.

6.1.3.1 Volatility

In Figures 11 and 12 below we observe a very similar behavior that the beta factor, we see as when the economic cycle was good in the previous years of the crisis the portfolio of low volatility overperformance the benchmark because they move less than the market and when the market drops due to a cycle change gets the higher return. The variation suffered by the low volatility portfolio during the time horizon is 1.51%, generating an expected average return of 1.85% above the market, while volatility is the lowest of all factors, obviously since it is the definition itself of the factor, 12.57% and a sharpe ratio of 0.14. The high volatility portfolio, green line in figure 11 (pink line in figure 12) gives an average expected return of 2.53% below the benchmark, thus reaffirming all the challenge presented so far. See table 8

Figure 11: High and low volatility portfolios weighted about benchmark

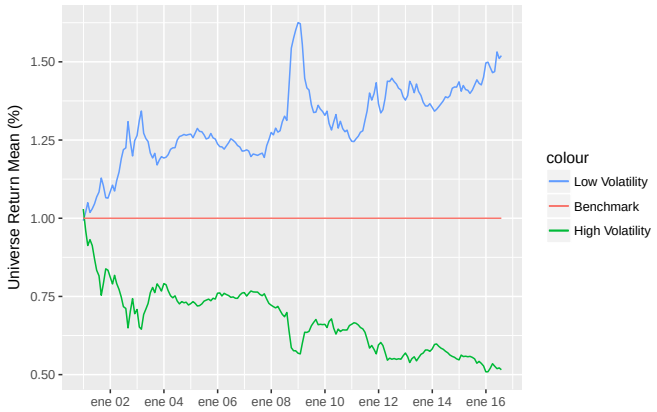
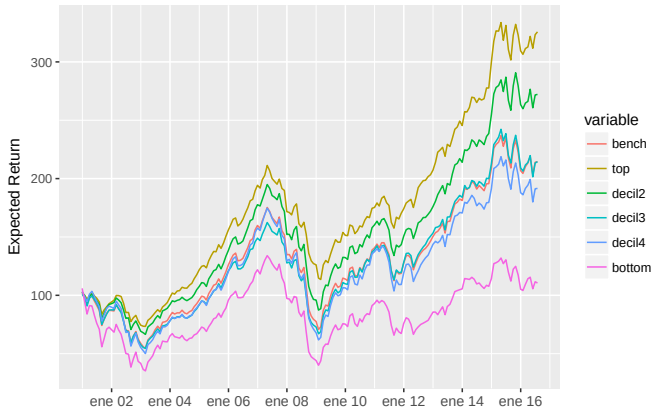


Figure 12:Shows all volatility portfolios and benchmark performance



	Low Volatility	High Volatility
Average Excess Return	1.857483	-2.538464
Annualized Volatility (%)	12.57218	25.56583
Shape's Ratio	0.1477455	-0.09929126
Variation over time	1.519812	0.5156497

Table 8: Performance statistics of volatility

6.1.4 Quality

Before starting with the analysis of factor quality we must denote the following point. The quality factor is a factor more economic than financial, although historically it has always been taken into account when investing although it is not one of the factors that behaves better. In addition there are some typologies of companies that by their very nature of business do not even have this factor. For example, any construction company has a debt that may be due to the purchase of materials, etc. But for example the financial sector, a bank, has net debt? A bank's business is to lend money, this factor does not make much sense then. For this reason, the quality factor is a very difficult factor to study, since this type of ratio does not work the same in all typologies of companies, in particular, neither financial nor insurance.

6.1.4.1 Net debt factor

Net debt shows a business's overall financial situation by subtracting the total value of a company's liabilities and debts from the total value of its cash, cash equivalents and other liquid assets, a process called netting. All the information necessary to determine a company's net debt can be found on its balance sheet.

Comprehensive Debt Analysis

The concept of net debt is particularly important in investing and is one of the most commonly used metrics in technical analysis. Stocks that perform well over time tend to be issued by companies that are financially healthy and able to afford to meet their obligations with ease. While the net debt figure is a great place to start, a prudent investor must also investigate the company's debt level in more detail. Important factors to consider are the actual debt figures - both short-term and long-term - and what percentage of the total debt needs to be paid off within the coming year.

The reason behind the debt is also important. A business can take on new debt financing to fund an expansion project, or it can use those funds to repay or refinance an older loan that it has not yet paid off, which may be a signal of deeper troubles. If the majority of the company's debts are short-term - meaning they must be repaid within 12 months - consider whether the business could afford to cover those obligations if its sales took a dive. If the company's current revenue stream is the only thing keeping it afloat, its long-term prospects may be in peril.

In Figure 13 we can see the behavior of the high and low net debt portfolios. Regarding the high net debt portfolio, we see how when the market gets complicated during a negative cycle, companies with greater net debt begin to have less income and chances are that even can not even deal with these debts, even entering into a competition of creditors, which causes them to fall below the market, while after the crisis we see how it has been behaving just like the benchmark itself. The results associated with this portfolio, as seen in Table 9, obtained an expected average yield of 0.51% below the benchmark, with a volatility of 18.91% including a negative sharpe ratio.

On the other hand, we have the portfolio of low net debt, marked in blue in figure 13 (in yellow in figure 14) which we can explain it is behavior as it is formed with companies with lower debt, are a safe value, and therefore, those investors who went into the stock market, did it buying these same ones. In the previous cycle of the crisis, even in the later one, it was a bit the same since interest rates are even now negative, so that all the debt, both large and small, can be refinanced without the greatest problem, each even you owe less.

However, when the economic cycle changes to bearish, and due to the previous explanation, the low met debt portfolios suffer an overperformance on the notorious market, specifically, the expected average yield of this portfolio is 1.86%, with an annualized volatility Of 16.59% and a sharpe ratio of 0.11 that can be seen in Table 9.

Figure 13: High and low net debt portfolios weighted about benchmark

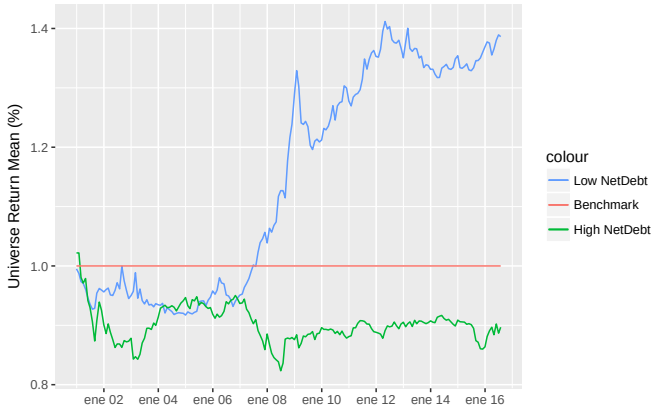
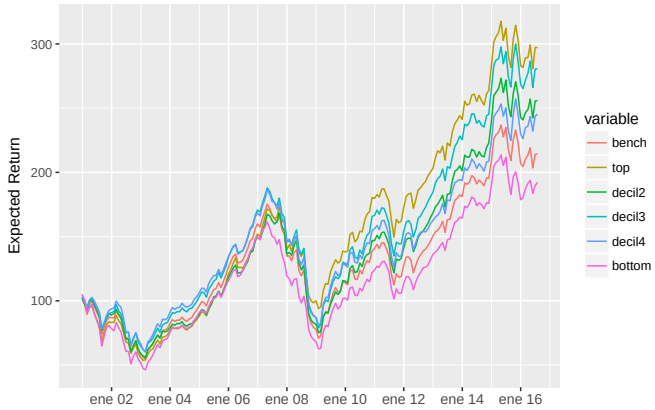


Figure 14:Shows all net debt portfolios and benchmark performance



	Low Net Debt	High Net Debt
Average Excess Return	1.864642	-0.5170789
Annualized Volatility (%)	16.59497	18.91296
Shape's Ratio	0.1123619	-0.02733993
Variation over time	1.386278	0.8967437

Table 9: Performance statistics of Net debt factor

6.1.4.2 ROE factor

Return on equity (ROE) is the amount of net income returned as a percentage of shareholders equity. Return on equity measures a corporation's profitability by revealing how much profit a company generates with the money shareholders have invested.

ROE is expressed as a percentage and calculated as:

$$ROE = \frac{Net \;\; Income}{Shareholder's \;\; Equity}$$

The results associated with this ROE factor are shown below in Figures 15 and 16. In Figure 15 we can see how the low ROE portfolio has a much more negative cycle changes in the market, whereas in the previous bullish cycle the crisis behaved like the bechmark itself. High ROE portfolio is that prior to the crisis investors did not really pay too much attention to this factor so it remained the same as The benchmark, but during the years after the crisis of 2009 when the cycle changed for the better, investors began to look at this factor more strongly and those companies with greater return of equity (ROE), captured an overperformance that reminds of the own momemtum factor. More specifically, in Table 10 we show the statistical ratios obtained, with respect to the expected average yield obtained by the high ROE portfolio was 1.30% higher than the benchmark, an annualized volatility of 16.35% and a very low Sharpe's ratio of 0.07, on the other hand, the low ROE portfolio obtained an average expected return of 2.22% below the benchmark and the annualized volatility of 21.5%.

Figure 15: High and low ROE portfolios weighted about benchmark

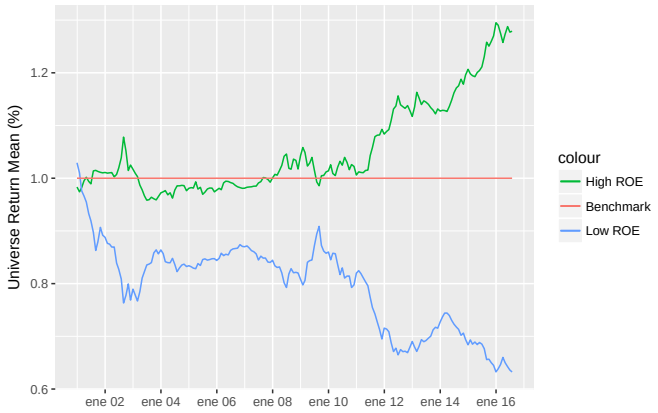
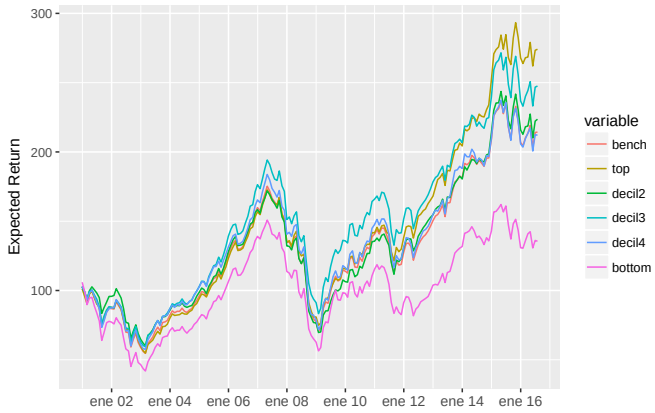


Figure 16:Shows all ROE portfolios and benchmark performance



	High ROE	Low ROE
Average Excess Return	1.308806	-2.222208
Annualized Volatility (%)	16.36549	21.50903
Shape's Ratio	0.07997352	-0.1033151
Variation over time	1.27892	0.6323498

Table 10: Performance statistics of ROE factor

6.2 Fundamental factor comparison

In this section, we will make a final comparison between all the fundamental factors explained in the previous section to give visibility to the strengths that are the ones we want to take to elaborate the new factor below. The highest expected average return obtained by the above factors is the high momemtum portfolio of 4.05% higher than the benchmark. Regarding the lowest annualized volatility obtained, the low volatility factor with 12.57% (similar to the low beta portfolio of 12.63%) obviously does honor it is own definition. Regarding the best ratio of Sharpe also refers to the high momemtum portfolio with a 0.26.

But in addition to taking into account the statistics of each factor, we will also try to capture the economical strengths in different cycles ocurrend on.

7. Smart Beta

After carefully analyzing the risk factors in the previous section, their weaknesses and strengths, the relevant question is, therefore, how to better extract the premium of a factor efficiently. Amenc et al. (2014a) present how the Smart Beta approach, whose main idea is to apply an intelligent weighting scheme to a selection of stocks, allows the construction of indexes of factors that are not only exposed to factors risk, but also avoid exposure to unrewarded risks. This approach, called **"smart factor indices"** [17] can be summarized as follows.

The explicit selection of stocks provides the desired slope, the beta, while the intelligent weighting scheme addresses the issues of concentration and it diversifies specific and unrewarded risks. Therefore, "Smart Beta approach" builds indexes of factors that explicitly seek exposures to rewarded risk factors, while unrewarded diversification of distance risks. The results we obtain suggest that factor indices lead to significant improvements in risk adjusted performance.

The flexible index build process used in second generation intelligent beta indexes enables you to take advantage of all the benefits of the smart beta, where the stock slection defines exposure to right (rewarded) risk factors and the intelligent weighting scheme allow unrewarded risks to be reduced.

7.1 Elementary Smart Beta

After presenting the concept of smart beta, we wanted in this research to make a first approximation that consists of constructing a new factor to gather the strengths of all the previous ones under a weighting done by us but in an "elemental" way, trying to leave aside tt is weaknesses with the objective of taking advantage of the upward cycles and avoiding or being able to cushion in bad times during a bearish economic cycle, for this reason we decided to rename it as "elemental smart beta".

7.1.1 Construction of the elemental smart beta factor

The process to construct this new factor is similar to the one used and explained in the chapter of data processing to find the classification in quintiles of the other factors that compose the data set. To achieve this, we first need to calculate the z-score values, normalize the values of each of the factors since it is not appropriate to use the original data since each was on a different scale, therefore, once normalized we can already compare them. We must denote that, in order to carry out this computation and following the structure we have been maintaining since the beginning, the computation has been made in function of the sector to which they belong and within each factor of the period, in other words, each value has been normalized by the period to which it belongs and within the sector to which it belongs.

Once we have the normalized values, and here comes one of the reasons why this new factor we will consider "elemental smart beta" is that we will weigh the values of each factor according to our criterion according to the average yield expected from factors presented in the previous section. As we can see in the following table 11, we present all the previous factors, their average expected performance, and the weighting that we have decided to assign to it according to our "elementary" criterion.

Factor	Expected Return	Weighted Coefficient
Momemtum	4.05%	30%
Earning Yield	3.21%	20%
Dividend Yield	2.04%	15%
NetDebt	1.86%	12%
Volatility	1.85%	10%
Beta	1.80%	8%
ROE	1.30%	5%

Table 11: Expected average yields of each factor and the weights associated with them.

7.1.2 Results obtained by the elemental smart beta factor

In this section we show the results achieved by the new elementary smart beta factor, which we will henceforth denote as ESM. If we look at the chart below, we get a pleasant surprise because at first glance we see how the portfolio built under the high ESM values gives us some rather pronounced profitability and gives us the feeling of having reduced the volatility. Specifically, if we look at the statistical results in Table 12, we see how the high ESM portfolio achieves an expected average return of 5.04% above the benchmark itself, an annualized volatility of 15.02% and a Sharpe ratio of 0.336, While this has generated a variation at the end of the time horizon of 2.36%. On the other hand, the low ESM portfolio has performed quite badly, obtaining an expected return of 6.19% below the benchmark itself, a volatility of more than 23.7% and a Sharpe ratio of -0.26.

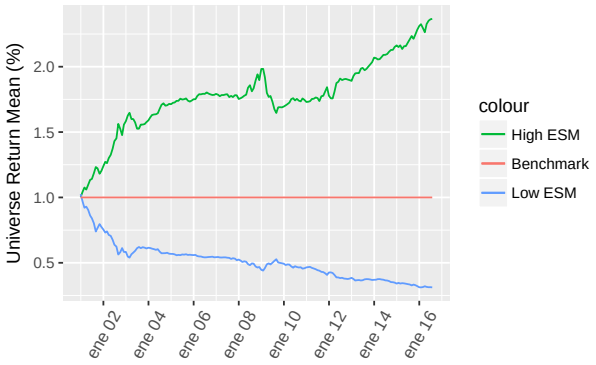
If we try to compare the results with the previous ones in order to check how close we are to achieving the ESM's own objective, we see how this ESM factor has achieved a higher performance than any of the other factors mentioned above. But how does he seem to have collected such fortresses? If we look at the graph we could say how in an earlier period of the crisis of 2009 seems to have captured the overperformance of the value factor (dividend yield and earning yield) that as we explained in the previous chapter tend to have a overperformce to the benchmark in bullish cycles. At the same time, when the cycle changed to bearish due to the crisis itself, this new ESM factor seems to have taken advantage of the strengths of the low risk factor (beta and volatility) to take advantage of the fall of the market itself and get a positive overperformance. Afterwards it suffered a fall next to the market but perhaps due to the quality factor (Netdebt EBITDA and ROE), ending up capturing the upward trend of the momemtum factor whose superior performance to the market is notorious and we have already reasoned previously.

Therefore, we can say that this ESM factor has obtained an increase in the expected average yield higher than the momemtum factor that was the one that previously improved yield, and has reduced annualized volatility compared to other factors, obviously without achieving low volatility factor that by it is own definition would not even make sense. Regarding Sharpe's ratio, a significant improvement has also been achieved in relation to the larger factor. We can say that we have achieved our goal under the construction of this ESM factor, by increasing the expected average yield, the volatility is the smaller, then we have achieved somewhat the objective of to obtain greater profitability with a lower volatility, so that is the objective.

Table 12: statistics performance of ESM factor

	High ESM	Low ESM
Excess Return	5.049712	-6.190875
Volatility	15.02863	23.76965
Ratio Sharpe	0.3360062	-0.260453
Portfolio Variation	2.365957	-0.3122333

Figure 17: Elementary Smart Beta Factor



8. Machine learning approach: next horizont to smart beta

8.1 Preliminaries

Machine learning is one of the fastest growing areas of computer science, with far-reaching applications. The aim of this chapter is the application of advanced statistical techniques such as machine learning is intended to build a new factor that combines the potentially positive characteristics of each of the factors explained above in order to be able to capture or capture the strengths of these factors and to dampen or avoid its weaknesses in a dynamic portfolio capable of knowing in each period how to maximize the average expected performance and, above all, to minimize annualized volatility in order to achieve the fundamental objective of the research.

8.2 Machine Learning Techniques

The term machine learning refers to the automated detection of meaningful patterns in data. By definition, machine learning is a branch of "artificial intelligence" where it is main purpose is converting experience into expertise or knowledge where the input to a learning algorithm is training data, representing experience, and

the output is some expertise. That is, to induce learning through feedback based on information. Roughly, the dynamics could be encircled in 3 large groups: pre-processing of data or information, training and post-processing of the results obtained.

Below are defined the fundamental elements of any learning system: inputs, outputs and types of algorithms. The entries are given in the form of an attribute vector and are called "instances". These attributes can be classified into "nominal attributes" if they take values within any finite set, or "numerical attributes" if they take real values within a finite or infinite set.

There are different types of learning algorithms which are governed by the type of output of the same:

1. Supervised learning: with this algorithm we can create a function that relates or establishes a correspondence between the output variables and the input variables, that is, transforms the input data into the results we expect. The most used supervised learning algorithms are Regression, Decision Tree, Random Forest, KNN, Logistic Regression.
2. Unsupervised learning: in this case, we do not have any objective or variable variables that we have to predict, if not rather, we seek to achieve groupings of the population in different groups. Since we do not have information about the input categories, it treats the input objects as a set of random variables and constructs a density model for that set, the algorithm being able to recognize patterns in order to be able to label the new inputs. Rata the input objects as a set of random variables, a density model being constructed for the data set. The most used non-supervised learning algorithms are: Apriori algorithm, K-means.
3. Reinforcement Learning: Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process

8.2.1 Decission trees

A decision tree is a predictor, $h : X \to Y$, that predicts the label associated with an instance x by traveling from a root node of a tree to a leaf. For simplicity we focus on the binary classification setting, namely, Y=$\{0,1\}$, but decision trees can be applied for other prediction problems as well. At each node on the root-to-leaf path, the successor child is chosen on the basis of a splitting of the input space. Usually, the splitting is based on one of the features of x or on a predefined set of splitting rules. A leaf contains a specific label.

A popular splitting rule at internal nodes of the tree is based on thresholding the value of a single feature. That is, we move to the right or left child of the node on the basis of $\mathbb{1}_{[x_i<\theta]}$, where $i \epsilon [d]$ is the index of the relevant feature and $\theta \epsilon R$ is the threshold. In such cases, we can think of a decision tree as a splitting of the instance space, $\mathbb{X} = \mathbb{R}^d$, into cells, where each leaf of the tree corresponds to one cell.

8.2.2 Random Forest

In particular, in this research we have applied the technique or algorithm "Random Forest". Another way to reduce the danger of overfitting is by constructing an ensemble of trees. In particular, in the following we describe the method of random forests, introduced by Breiman (2001) [12].This is based on a combination of predictor trees such that each tree depends on the values of a random vector independently tested and with the same distribution for each of these, let's say it is a substantial modification of bagging that builds a long collection of uncorrelated trees and then average them.

The essential idea of bagging is to average many noisy but approximately unbiased models, and therefore reduce variation. But why trees are the ideal candidates for bagging? This is because they can register complex interaction structures in the data, and if they grow deep enough, they have relatively low bias. Since trees are notoriously noisy, they benefit greatly by averaging.

The objective is to create a model that predicts the value of a target variable based on various input variables, each inner node corresponds to one of the input variables and each sheet represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

The most known or used decision trees are of two main types:

- Classification trees are when the predicted result is the class to which the data belong.
- Regression trees are when the predicted outcome can be considered a real number (for example, the price of a house, or the length of a patient's stay in a hospital).

8.2.2.1 Random forest: Influence measures

There are two ratios to explain how important the variables of the dataset are when explaining the response that we seek to optimize. In our case, we obtain them through the function of the package "RandomForest" of R-study software that we have been using, and are: **"% IncMSE"** and **"IncNodePutity"**.

The first one is the most robust measure, it offers more information and can be interpreted as follows, if a predictor is important in the current model, then assigning other values to that random predictor should have a negative influence on the prediction, the same model to predict from original data should give worse predictions. Therefore, a predictive measure (MSE) is taken with the original dataset and then with the dataset exchanged, and compared in some way. In a way, especially since we expect the original MSE to always be smaller, we can make a difference. Therefore, the higher the better or more importantly basically.

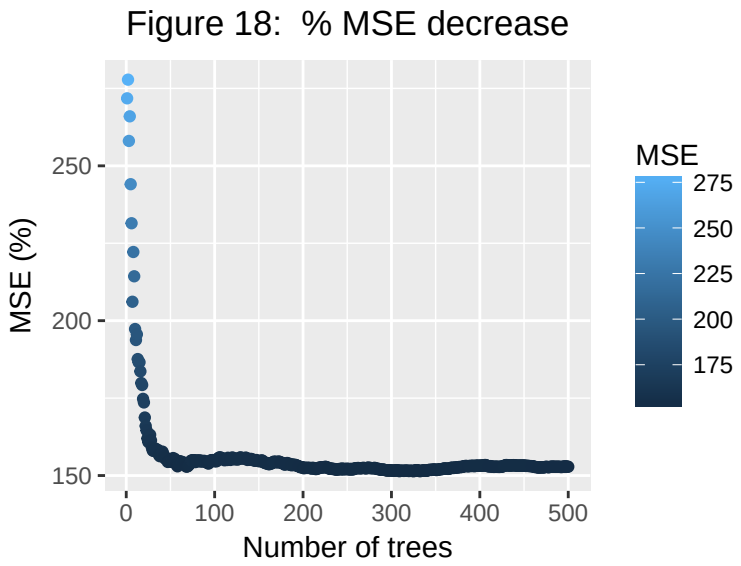
For the second measure, in each division, it is possible to calculate how much reduces the impurity of the node. The most useful variables achieve greater increases in the purity of the node, that is, find a division that has a high variance between nodes and a small intra-nodal variance. "IncNodePurity" is partial and should only be used if the extra computation time to calculate IncMSE% is unacceptable, which in our case is not really relevant since both computations provide a similar calculation time.

8.2.2.2 Random forest: Applications in our research

The procedure used in this research, it is worth pointing out that it will have two objectives. The first objective is to create a function that computes regression trees for each of the monthly periods in order to build a portfolio that we will denotas as *high smart beta* (the continuation of the previous high elementary smart beta ESM), based on the criterion previously explained of the importance of the variables in each period. With this, we want to develop a dynamic portfolio that, depending on it, is able of capturing the power of the most important variable in the immediately previous period, being able to recognize it and invest in the current period according to it.

The second objective or perhaps it can be denoted as a variant of the previous one, is to be able to build a portfolio high smart beta, but in this case, capturing the most important variable during the three periods immediately previous to the present one, and investing in the following according to these. Let is say that we are trying to adjust that procedure to see how well or how badly it works.

It should be noted that the adjustment of the number of trees to be computed in each period has been tested under the criterion of minimizing the mean square error obtained as a function of the number of trees calculated. To give a visual sample of the same, below we can see a figure where the% MSE obtained is plotted as a function of the number of trees computed. It is observed how a smaller number of trees greater MSE obtained, but this percentage is reduced when we increase that number until there are around 300 trees where it seems that this percentage begins to stabilize, for this reason the value of the parameter ntrees to which reference we decided to leave it in this amount.

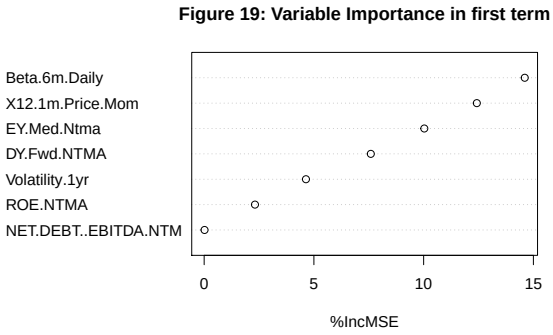


8.2.2.3 First and second term variable importances computation

In order to continue showing the development, in this section, we show the results obtained by the algorithm used during the first two periods in order to contribute the conclusions that we will make for each and every period of the entire time horizon. For period one, table 13 is shown both important measures explained in the previous sections, although as we have clarified the one used in this research has been IncMSE. Most relevant in this table is that the most important variable is the beta factor with 31.75%, which is interpreted as the increase of 31.75% of the MSE in the prediction of the average returns expected by the portfolio if we extract this variable from the model , followed by the momemtum factor with 23.27%. These results are also reflected in figure 18 attached to the table showing the classification of all variables according to their importance.

	IncMSE	IncNodePurity
Beta	31.757874	12141.248
Momemtum	23.276754	12217.477
Earn.Yld	10.919058	8297.780
Divd.Yld	8.406248	5840.536
Volatility	4.960217	8239.339
ROE	2.011630	7284.472
NET.DEBT	1.774	5972.498

Table 13: Importance measures of each factor.

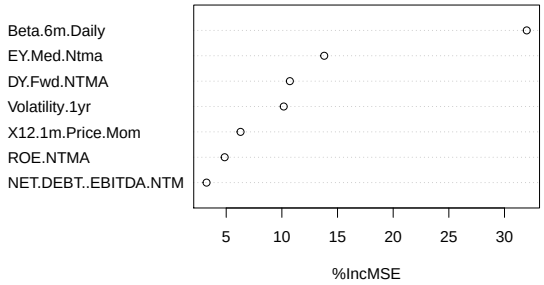


Next we show the results but for the second period as a visual explanation to get a clear the algorithm used. In table 14 we can see how for this second period agrees the most important variable as the beta factor which would contribute an increase of the MSE in predicting the expected average yields of 68.47%, being in this period almost double against the second more important variable than in this case would be the earning yield factor with a 29.51%. This can be explained, and we remember an affirmation made in a previous section, that the beta factor is not a rotational factor as other as the momemtum factor so it would explain better the variable response during periods longer than other factors that suffer more rotation. Again we present Figure 19 where the classification of the table attached to its side is graphed.

	IncMSE	IncNodePurity
Beta	68.4769871	15611.874
Earn.Yld	29.5199103	8229.745
Divd.Yld	11.2605642	5194.761
Volatility	11.1924177	8989.533
Momemtum	6.6287787	6300.367
ROE	2.1791019	3003.160
NET.DEBT	0.8122283	2932.054

Table 14: Importance measures of each factor.

Figure 20: Variable Importance in second term



Finally, we generate the algorithm that period to period stores the most important variable and we build the portfolio high smart beta and low smart beta factor.

8.3 Smart beta factor results

In this section we will show the results obtained by the portfolios built under the procedure explained above for both the portfolio built with one month of training and three months of training. We will show under the same structure that we have maintained throughout the research, a table with the statistical results obtained by both portfolios and a graph where you can clearly visualize the behavior obtained by each of these portfolios.

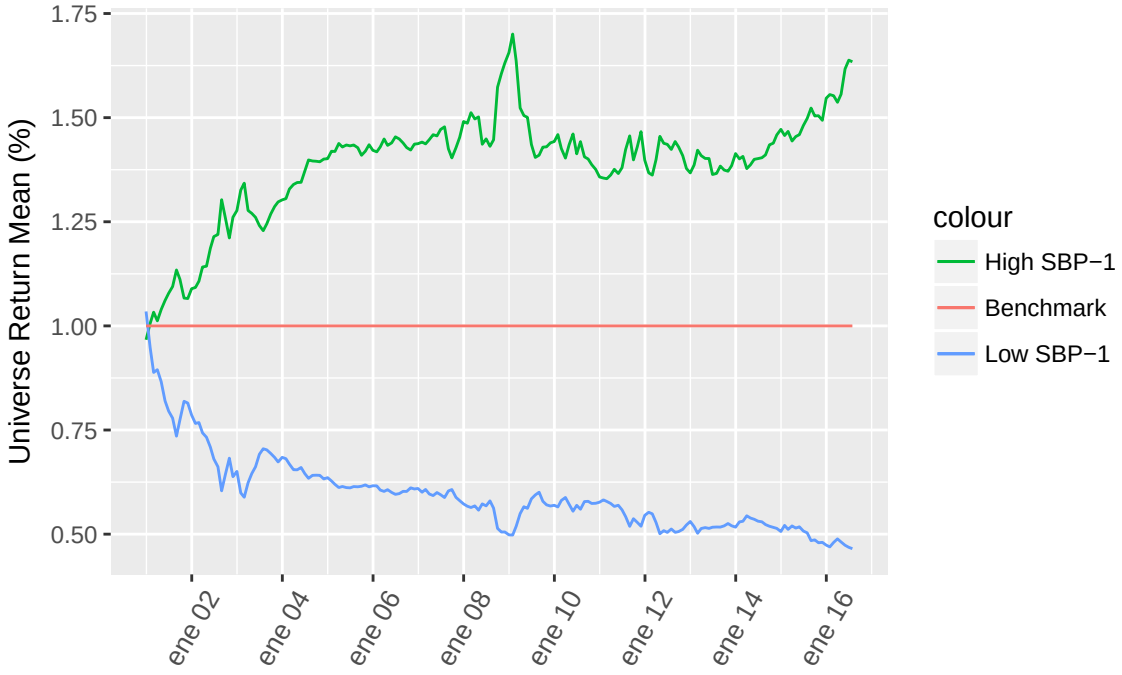
8.3.1 One month training portfolio results

Firstly, we will show the results obtained by the smart beta portfolio under the one-month training period which we will henceforth denote as SBP-1 (Smart Beta Portfolio 1 month). In table 15, the statistical results obtained by both portfolios, we see how the high SMP-1 achieves an average expected return of 2.83% higher than the benchmark, an annualized volatility of 14.15% and a Sharpe's ratio of 0.20, if we compare it with the fundamental factors we see that only the momemtum factor individually surpasses this performance. With regard to volatility, we have achieved relatively low annualized volatility, only surpassed by the low risk factor, which makes sense to us, and Sharpe's ratio could be said to be high compared to other factors, surpassed only by a few tenths earning yield factor, that is, the results at first sight appear to provide an improvement.

If we compare it with the elementary smart beta portfolio (ESB), in terms of average expected performance we are below the previously mentioned 2.83% versus 5.04% of the ESB, it means an expected return average almost half. But in terms of annualized volatility if we experience a decline of the same even below the ESB but without reaching 12.57% of the low risk portfolio. The Sharpe's ratio is also lower than the ESB.

Finally, if we try to compare the behavior of SBP-1 low depending on the strengths of the fundamental factors, we see how in the first cycle, before the crisis, it captured the overperformance of the low risk and high value factors, then the rebound of the high momemtum factor just when the cycle changed during the crisis, to later set the similiar behavior to the benchmark of the high quality factor and finally to overperformance of the high momemtum factor.

Figure 21: Shows the performance of high and low smart beta portfolios



	High SBP-1	Low SBP-1
Average Excess Return	2.831524	-3.845189
Annualized Volatility (%)	14.1567	0.2000129
Shapes Ratio	0.2000129	-0.1565171
Variation over time	1.711089	0.4377229

Table 15: Performance statistics of Smart beta portfolio one month training (SBP-1)

8.3.2 Three months training portfolio results

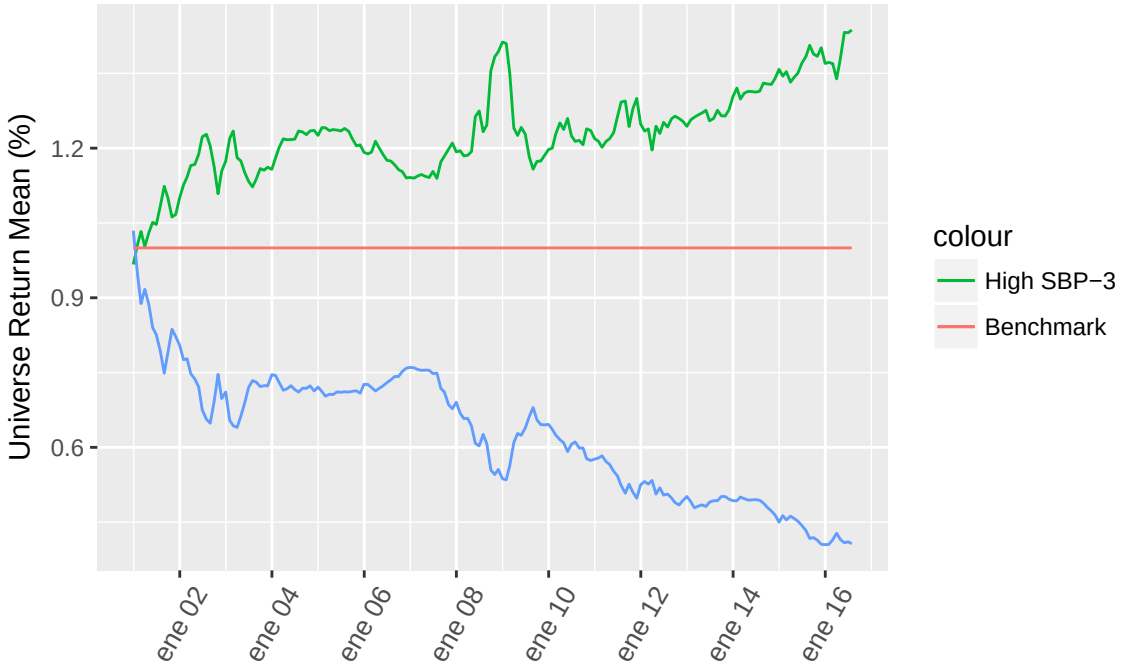
Before presenting the results obtained with the portfolio of high and low smart beta with 3 months of training denoted hereafter as SBP-3, it should be pointed out that the procedure for obtaining the most important variables for each period is the same but taking as training period 3 months instead of one, so we decided not to repeat it in order to speed up the understanding and the visual load in the research.

The results obtained by the high and low SBP-3 can be seen in Table 16 below, the expected average performance of SBP-3 high is 1.88% above the becnhmark, annualized volatility is 15.25% and Sharpe's ratio is 0.12. But how can we verify how good they are? Comparing them with the SBP-1 result. If we look at the statistical results of both tables, we can see how for this variant of the 3-month training the results have worsened at all points, both lower expected performance, higher annualized volatility and a lower Sharpe's ratio, so the training adjustment of 3 months has not given us a better performance than the one month.

However, if we compare it with the factors we see in terms of expected average profitability is not one of the best factors, but in terms of annualized volatility, we managed to reduce it only surpassed by the low risk portfolio.

Finally, trying to give some sense to it is behavior during the time horizon and the possible strengths captured by it, well, we get more or less similar results to SBP-1, where in a first period it seems to capture the overperformance of the low risk Portfolio, while when the cycle turned around due to the crisis also makes sense the high momemtum until the end of the time horizon.

Figure 22: Shows the performance of high and low smart beta portfolios



	High SBP-3	Low SBP-3
Average Excess Return	1.880496	-4.329258
Annualized Volatility (%)	15.25179	24.5535
Shapes Ratio	0.1232968	-0.1763194
Variation over time	1.436896	0.4067427

Table 16: Performance statistics of Smart beta portfolio one month training (SBP-3)

9. Conclusions

To conclude this research, we will provide a final reasoning about the behavior and results obtained under the elementary smart beta and the smart beta one month and it iss variant of three months.

In a first approach to building a factor that brings together the strengths of the fundamental factors low risk, value, momemtum and quality, which we denote as elementary smart beta because the foundation on which we rely for it is construction is to take the simple average of the factors that make it better and worse, we have obtained a very profitable result not only in terms of expected average profitability, and Sharpe's ratio, but in the reduction of annualized volatility, which was therefore successful with the fundamental objective of the research itself.

The next step has been to use more elaborate techniques of advanced machine learning statistics such as random forest. For this, it has been decided to compute a large number of regression trees so as to minimize the %MSE in each period, taking exclusively the most important variable. Under this criterion, we have obtained results that are not as good as the previous ones, say that they do improve or comply with the research's own objective of minimizing variance by maximizing expected returns.

To finish developing a small variant of the previous one, but taking as a training period a somewhat longer period of 3 months, where again even worse or not so good results have been obtained for the period of one month of training, that yes improving even those of each factor individually.

Obviously, you can, moreover, you should go much further, and as a goal for a next research or even the continuation of this should be able to get develop and improve this process even reaching beyond that elementary smart beta that we have obtained here. The calibration of the model is a point where to develop this work, that is to say, which training period is the optimal one at the time of the computation of the regression trees, what is the exact number of trees that manage to minimize the %MSE in each one of those training periods, what number of variables we should extract as important rather than just one or even the average of a number of them the most important and if we use the least significant ones to extract a number of portfolio titles belonging to those variables.

Therefore, the line of research to follow onwards is broad and motivating

9. Bibliografy

References

[1] Markowitz, Harry M. (March 1952), *"Portfolio Selection, The Journal of Finance".*

[2] Willian F. Sharpe: A simplified model for portfolio analysis, *"Portfolio Theory and Capital Markets"*, (McGraw-Hill, 1970);ISBN 0-07-135320-8.

[3] Markowitz, Harry M. (March 1987), *"Mean-variance in portfolio choice and capital Markets"*, Oxford: Blackwell.

[4] Galitz L. (1996), *"Finantial Engineering"*, Prentice Hall.

[5] Knight, Frank. (1921), *"Risk, uncertainly and profit"*, ISBN 978-0-9840614-2-6.

[6] Ruppert, David, *"Statistics and data analysis for financial engineering"*, Springer.

[7] Lintner, John (February 1965), *"The Valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets"*

[8] Jensen, Michael C., Black, Fischer and Scholes, Myron S. (1972), *"The Capital Asset Pricing Model: some empirical tests"*, Praeger Publishers Inc.

[9] Fama, Eugene F. and MacBeth, James D., *"Risk, Return and Equilibrium: Empirical Tests, Journal of Political Economy"*, Vol.81, N3 (May-Jun 1973).

[10] Zhang, Lu. (2005), *"The value premium"*, Journal of finance. Pg:67-103

[11] Frazzini A, Pedersen L H (2010), *"Betting Against Beta"*, NBER Working Paper.

[12] Breiman, L. (1996), *"Bagging Predictors: Machine Learning"*, pp. 123-140.

[13] Financial times handbook of financial engineering, *"Tools and techniques for managing derivaties, options, swaps and risk"*,Prentice Hall.

[14] Baltas, N., Jessop, D., Jones, S., Winter, P., Wu, S., Antrobus, O. and Stoltz, P. (January 2015), *Quantitative Monographs: "Stock selection using Machine Learning"*, UBS Global Research.

[15] Baltas, N., Jessop, D., Jones, C., Lancetti, S., Winter, P. and Holcroft, J (July 2015), *Quantitative Monographs. "Low-Risk Investing: perhaps not everywhere"*, UBS Global Research.

[16] Sefton, J., Jessop, D., De Rossi, G., Zhang, H., Jones, C. (March 2012), *UBS Investment Research. "A fresh look of beta puzzle"*, UBS Global Equity Research.

[17] Baltas, N., Jessop, D., Jones, C., Lancetti, S., Winter, P. and Holcroft, J., Gerken, J., Ivanova, J., Wu, S., Antrobus, O. and Stoltz, P. (September 2016), *Academic Research Monitor. "Combining Smart Beta Factors"*, UBS Global Research.

[18] Jessop, D. and Lancetti, S. (December 2013), *Global Quantitative Research Monographs. "3 Reasons why high-risk underperformed"*, UBS Global Research.

10. Appendix

```
###  Download Data Set Script


# Cargamos ticket STOXX600
stocksLst <- read.csv("E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/Datos/sp500.csv",
                      header = F, stringsAsFactors = F)
nrow(stocksLst)
stocksLst<-stocksLst[,1]

# Aplicamos el sufijo "NYSE:"
tickets<-sapply(stocksLst, function(x) paste("NYSE:", x), USE.NAMES=FALSE)



###############################################################################
# Load Systematic Investor Toolbox (SIT)
# https://systematicinvestor.wordpress.com/systematic-investor-toolbox/
############################################### ################################
con = gzcon(url('http://www.systematicportfolio.com/sit.gz', 'rb'))
source(con)
close(con)


###############################################################################
# determine date when fundamental data is available
# use 'date preliminary data loaded' when available
# otherwise lag 'quarter end date' 2 months for Q1/2/3 and 3 months for Q4
###############################################################################
date.fund.data <- function(data)
{
  # construct date
  quarter.end.date = as.Date(paste(data['quarter end date',], '/1', sep=''), '%Y/%m/%d')
  quarterly.indicator = data['quarterly indicator',]
  date.preliminary.data.loaded = as.Date(data['date preliminary data loaded',], '%Y-%m-%d') + 1

  months = seq(quarter.end.date[1], tail(quarter.end.date,1)+365, by='1 month')
  index = match(quarter.end.date, months)
  quarter.end.date = months[ iif(quarterly.indicator == '4', index+3, index+2) + 1 ] - 1

  fund.date = date.preliminary.data.loaded
  fund.date[is.na(fund.date)] = quarter.end.date[is.na(fund.date)]

  return(fund.date)
}

#****************************************************************
# Load historical fundamental data
# http://advfn.com/p.php?pid=financials&symbol=NYSE:WMT&mode=quarterly_reports
#****************************************************************
# 1436513D UN
Symbol = data
fund = fund.data(Symbol, 80)


prub<-data.frame(fund)
for(i in 1:length(aaa)){
  prub<-cbind(prub,data.frame(fund.data(aaa[i], 80)))
}
ccc<-row.names(fund)
prub<-cbind(ccc,prub)
ddd<-c("Fundamentals",rep("NYSE:AAPL",81),rep("NYSE:DDD",81),rep("NYSE:FB",81))
prub<-rbind(ddd,prub)
Symbol2 = 'NYSE:AAPL'
fund2 = fund.data(Symbol = Symbol2, 80)
Symbol3 = 'NYSE:DDD'
fund3 = fund.data(Symbol = Symbol3, 80)
Symbol4 = 'NYSE:FB'
fund4 = fund.data(Symbol = Symbol4, 80)
# construct date
fund.date = date.fund.data(fund)


#****************************************************************
# Load historical data
#****************************************************************
load.packages('quantmod')
tickers = 'WMT'
```

```
data <- new.env()
getSymbols(tickers, src = 'yahoo', from = '1980-01-01', env = data, auto.assign = T)
for(i in ls(data)) data[[i]] = adjustOHLC(data[[i]], use.Adjusted=T)

data$WMT = merge(data$WMT, EPS)
# back fill EPS
data$WMT$EPS = ifna.prev(coredata(data$WMT$EPS))

##  Almacenamos los datos en un data set llamado DATA, a partir de ahora los cargaremos directamente.
save(data, file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/datos/SortedData.RData")


## Creación de un vector donde almacenar las posiciones con el cambio de fecha, ya que a partir de ahora
##  necesitamos utilizarlo ya que trabajamos periodo a periodo

# Creamos un vector vacío
v<-c()

# Loop para almacenar dicha posición cuando encuentre diferencias en j y j+1
j=1
for(i in 1:(nrow(data)-1)){
  if(data[i,2]!=data[i+1,2]){
    v[j]<-i+1
    j=j+1
  }

}

# Comprobación
v[1:5]

##  Ahora tenemos un vector con la posición donde cambia el periodo pero necesitamos otro
##  donde se almacene la última fecha para cada periodo

v2=v-1 # Simplemente se obtiene como la resta de la posición anterior al cambio.
v2[1]=1 # indicamos que la primera posición es un uno

##  A partir de este momento almacenamos ambos vectores con el fin de cargarlos
##  cuando se necesite.

save(v, file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/v2.RData")
save(v2, file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/v2.RData")

## Función base del proyecto

##  A continuación, creamos una función que toma como parámetros de entrada cualquier factor
##  que se desee estudiar: "obj", y además cada uno de los quintiles que se desee calcular y graficar.
##  Por ejemplo, el factor volatility y dibújame el top portfolio (low volatility) contra el bottom
##  portfolio (high volatility).


simulacion<-function(obj,top=F,f2=F,f3=F,f4=F,bottom=F){

  ### Necesitamos la base de datos completa y la base de datos donde solo esta el universe returns
  load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/Datos/datos1.RData")
  load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/Datos/datos2.RData")
  attach(datos)
  attach(datos2)
  library(data.table)

  ### Amacenamos el nombre de la variable objeto de estudio que ha entrado como parámetro de entrada.
  nombre=deparse(substitute(obj))



  ###   Calcula para cada dia la media del UNIVERSE RETURN y clasifica por fechas
  uniret<-aggregate(datos2[, 2], list(Period..YYYYMMDD.), mean)
  colnames(uniret)<-c("Fecha","Media") #  Renombramos las columnas

  ###   A continuación recogemos los valores de la variable objeto de estudio
  datos3<-data.frame(datos2,obj)
  colnames(datos3)<-c("Fecha","Universe Returns","Factor") #  Renombramos columnas

  ### Problema con los NA's, tenemos que evitarlos a la hora de hacer este calculo (3 NA's)
  ### Metemos un 0 porque los fractiles se mueven en un rango de {1,5} digamos que no los cogeriamos
  ### bajo ningun caso
  datos3$Factor[is.na(datos3$Factor)] <- 0
  summary(datos3)
```

```
### DataSet Quintile 1
F1<-datos3[which(datos3$Factor==1),]

### Agrupamos por fechas y al mismo tiempo se realiza una funcion de media para cada uno de ellos.
F1<-aggregate(F1[, 2], list(F1$Fecha), mean)
colnames(F1)<-c("Fecha","Media") #  Renombramos columnas


### DataSet Quintile 2
F2<-datos3[which(datos3$Factor==2),]

### Agrupamos por fechas y al mismo tiempo se realiza una funcion de media para cada uno de ellos
F2<-aggregate(F2[, 2], list(F2$Fecha), mean)
colnames(F2)<-c("Fecha","Media")#  Renombramos columnas

### DataSet Quintile 3
F3<-datos3[which(datos3$Factor==3),]

##Agrupamos por fechas y al mismo tiempo se realiza una funcion de media para cada uno de ellos

F3<-aggregate(F3[, 2], list(F3$Fecha), mean)
colnames(F3)<-c("Fecha","Media")#  Renombramos columnas

### DataSet Quintile 4
F4<-datos3[which(datos3$Factor==4),]

##Agrupamos por fechas y al mismo tiempo se realiza una funcion de media para cada uno de ellos
 F4<-aggregate(F4[, 2], list(F4$Fecha), mean)
colnames(F4)<-c("Fecha","Media")#  Renombramos columnas

### DataSet Quintile 5
F5<-datos3[which(datos3$Factor==5),]

##Agrupamos por fechas y al mismo tiempo se realiza una funcion de media para cada uno de ellos
F5<-aggregate(F5[, 2], list(F5$Fecha), mean)
colnames(F5)<-c("Fecha","Media")#  Renombramos columnas


### Construyo un data table que reúne todo los anteriores
require(data.table)
datos4<-data.table(F1,F2[,2],F3[,2],F4[,2],F5[,2])
colnames(datos4)<-c("Date","Top","F2","F3","F4","Bottom")

### SIMULACION

### Simulación del BENCHMARK:
BNCHM<-c(100,rep(0,nrow(uniret)))

#En primer lugar la inversion inicial = 100, le debe sumar o restar a esta cantidad,
#el retorno que le corresponda que como es un porcentage...
for(i in 1:nrow(uniret)){
  BNCHM[i+1]<-BNCHM[i]+(BNCHM[i]*uniret[i,2]/100)
}

### Simulación del Quintile 1

SimF1<-c(100,rep(0,nrow(F1)))

for(i in 1:nrow(F1)){
  SimF1[i+1]<-SimF1[i]+(SimF1[i]*F1[i,2]/100)
}


### Simulación del Quintile 2

SimF2<-c(100,rep(0,nrow(F2)))

for(i in 1:nrow(F2)){
  SimF2[i+1]<-SimF2[i]+(SimF2[i]*F2[i,2]/100)
}


### Simulación del Quintile 3

SimF3<-c(100,rep(0,nrow(F3)))

for(i in 1:nrow(F3)){
  SimF3[i+1]<-SimF3[i]+(SimF3[i]*F3[i,2]/100)
}
```

```
### Simulación del Quintile 4

SimF4<-c(100,rep(0,nrow(F4)))

for(i in 1:nrow(F4)){
  SimF4[i+1]<-SimF4[i]+(SimF4[i]*F4[i,2]/100)
}


### Simulación del Quintile 5

SimF5<-c(100,rep(0,nrow(F5)))

for(i in 1:nrow(F5)){
  SimF5[i+1]<-SimF5[i]+(SimF5[i]*F5[i,2]/100)
}


### Una vez contruídas las simulaciones, vamos a graficar los resultados. Mostraremos dos gráficos,
### uno donde se muestran el comportamiento de todos los quintiles de la variable de estudio (portfolios),
### y otro donde ponderaremos contra el benchmark con el fin de observar con detalle cómo se comportan
### ante el mercado.

### Creamos un data set con los resultados de todas las simulaciones
completeGraph<-data.table(date=uniret$Fecha,bench=BNCHM[2:189],top=SimF1[2:189],decil2=SimF2[2:189],
                          decil3=SimF3[2:189],decil4=SimF4[2:189],bottom=SimF5[2:189])
attach(completeGraph)

### Creamos una ventana donde observar ambos gráficos al mismo tiempo.
par(mfrow=c(1,2),bty="n")

### Reordenamos el data set anterior
graf_CTO_long <- melt(completeGraph, id="date")  # convert to long format

### Creamos el primer gráfico

gto<-ggplot(data=graf_CTO_long,
       aes(x=date, y=value, colour=variable)) +
  ggtitle("Figure 10:Shows all momemtum portfolios and benchmark performance")+
  labs(x = "", y = "Expected Return")+
  theme(axis.text.x=element_text( size=10, vjust=0.5))+
  geom_line()+ scale_x_date(date_breaks = "2 years",date_labels = "%b %y")


### Ahora creamos el gráfico donde se ponderan contra el propio benchmark

### Calculamos los resultados para cada portfolio
bnchm<-BNCHM/BNCHM
top<-SimF1/BNCHM
resF2<-SimF2/BNCHM
resF3<-SimF3/BNCHM
resF4<-SimF4/BNCHM
bottom<-SimF5/BNCHM

### De nuevo creamos un data set con todos los resultados para cada simulación
graf<-data.table(date=uniret$Fecha,bench=bnchm[2:189],top=top[2:189],decil2=resF2[2:189],decil3=resF3[2:189],
                 decil4=resF4[2:189],bottom=bottom[2:189])
attach(graf)


### Cogemos el caso donde queremos dar como parámetros de entradas el top and bottom portfolio
if(top==T && bottom==T){

  ###   Creamos el segundo gráfico
  g<- ggplot(graf,aes(x=graf$date,y=top,color="High Momemtum"))+
    geom_line()+
    labs(x = "", y = "Universe Return Mean (%)")
    theme(axis.text.x=element_text(angle=75, size=10, vjust=0.5))

  g<-g+geom_line(data=graf, aes(y = bottom,color="Low Momemtum"))+
    ggtitle("Figure 9:  High and low momemtum portfolios weighted about benchmark")+
    geom_line(data=graf,aes(y=bench,color="Benchmark"))+
    scale_color_discrete(breaks=c("High Momemtum","Benchmark","Low Momemtum"))+
    scale_x_date(date_breaks = "2 years",date_labels = "%b %y")

  ### Devolvemos los dos resultadios de la función (ambos gráficos)
  print(g)
  return(gto)

}}
```

```
###   Como bien hemos dicho esta función es digamos el eje principal del trabajo ya que todos los resultados
###   se obtienen bajo la llamada a la misma.

###   El siguiente paso es normalizar la base de datos bajo el procedimiento explicado en el trabajo


###   Cargamos los Datos (ya ordenados por sector, periodo (mensual) y ISIN)
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/datos/SortedData.RData")

## #  Cargamos los vectores donde se recogen los cambios de fecha (para pasar de un periodo a otro)
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/v.RData")
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/v2.RData")

###   Creamos un nuevo data set pero esta vez solo con un ID, Fechas, ISIN, Sector y todos los factores
col<-c(1,2,3,4,12,14,16,18,20,22,24)
factores<-data[,col]

###   Creamos una copia de dataset para calcular el z-score de cada factor.
factores1<-factores

###   Vuelvo a crear una copia del data set (aunque no se necesita porque no hemos hecho la media)
factores2<-factores1

###   Cálculo del z-score, pero en función del periodo, es decir, la media y la desviación utilizado es solo por periodo.

##    Especificamos las columnas correspondientes a los fctores
col2<-c(5:11)

###   Loop donde se calculan los coeficientes z-score
i=1
for(i in 1:3479){
  factores2[v[i]:v2[i+1],col2]<-scale(factores2[v[i]:v2[i+1],col2])
  print(i)
}


###   Periodo de comprobación del loop
###   primer periodo:
i=1
factores1[v[i]:v2[i+1],5]
# View(factores1)
factores2[v[i]:v2[i+1],5]
##    Comprobación manual (coinciden los primeros coeficientes)
(0.109096-mean(factores1[v[i]:v2[i+1],5]))/sd(factores1[v[i]:v2[i+1],5])

###   segundo periodo:
i=2
factores1[v[i]:v2[i+1],5]
#View(factores1)
factores2[v[i]:v2[i+1],5]
##    Comprobación manual (coinciden los primeros coeficientes)
(0.104451-mean(factores1[v[i]:v2[i+1],5]))/sd(factores1[v[i]:v2[i+1],5])

###   Ejemplo para el último periodo:
i=3479
factores1[v[i]:v2[i+1],5]
#View(factores1)
factores2[v[i]:v2[i+1],5]
##    Comprobación manual (coinciden los primeros coeficientes)
(0.041651-mean(factores1[v[i]:v2[i+1],5]))/sd(factores1[v[i]:v2[i+1],5])

###   Necesitamos denotar que los valores para las variables beta, volatilidad, netdebt
###   son al contrario que el resto, menores valores, mayor rendimiento.

###   Creamos un data set con los valores normalizados, y damos la vuelta a esas variables
zScore<-factores2
zScore$NET.DEBT..EBITDA.NTM<-zScore[,8]*(-1)
zScore$Beta.6m.Daily<-zScore[,10]*(-1)
zScore$Volatility.1yr<-zScore[,11]*(-1)

###   El siguiente paso es realizar las ponderaciones de cada factor en función de su comportamiento.
###   Explicado en el trabajo.
zScore[,5]<-0.2*zScore[,5]#EY=3.21
zScore[,6]<-+0.15*zScore[,6]#DY=2.04
zScore[,7]<-+0.3*zScore[,7]#MM=4.05
zScore[,8]<-+0.12*zScore[,8]#ND=1.86
zScore[,9]<-+0.05*zScore[,9]#ROE=1.3
zScore[,10]<-+0.08*zScore[,10]#Beta=1.8
zScore[,11]<-+0.1*zScore[,11]#Vol=1.85
```

```
###   Construccción del Elemental Smart beta factor

###   Primera prueba manual solo de los dos primeros periodos
###   Creamos un dataset conde almacenamos el identificador, la fecha, el periodo y la posición
###   donde almacenaremos los valores para el nuevo elemental smart beta factor
fctor<-data.frame(ID=zScore$ID,Fecha=zScore$Period..YYYYMMDD.,NF=zScore$NF,Fractiles=c(rep(0,nrow(zScore))))

###   Calculamos los quintiles, es decir, las probabilidades que hacemos servir para la clasificación
###   de los valores.

###   Como te comentaba utilizo los dos vectores donde tengo almacenado los cambios de fecha: v[1]:v2[2]
a<-quantile(fctor$NF[v[1]:v2[2]],probs = c(0.2,0.4,0.6,0.8),na.rm = T);a

###   Y ahora para las 10 primeras observacines, si es menor que fractil 20, le asigno 5,
###   si está entre el 20 y 40 le asigno un 4, entre en el 4 y 3, le asigno un 3,
###   entre el 2 y el uno le asigno un 2, y si es mayor que el 20 le asigno un 1.
###   Pero de momento solo a las 10 primeras observaciones para ir comprobando.

fctor[which(fctor$NF[1:10] <= (a[1])),4]<-5
fctor[which(fctor$NF[1:10] > a[1] & fctor$NF[1:10] <= a[2]),4]<-4
fctor[which(fctor$NF[1:10] > a[2] & fctor$NF[1:10] <= a[3]),4]<-3
fctor[which(fctor$NF[1:10] > a[3] & fctor$NF[1:10] <= a[4]),4]<-2
fctor[which(fctor$NF[1:10] > a[4]),4]<-1

###   Periodo de comprobación del cálculo

##    Fractiles:
a
##    Asignación:
fctor$NF[1:10]

##    Realmente lo ha hecho bien

###   Realizamos el cálculo para todo el dataset: vamos cogiendo periodo a periodo, y calculamos
###   las probabilidades parra ellos. A continuación cogemos los valores de cada periodo y los
###   clasificamos en función de dichas probabilidades.

for(i in 2:length(v)){

  b<-quantile(fctor$NF[v[i]:v2[i+1]],probs = c(0.2,0.4,0.6,0.8),na.rm = T);b

  fctor[v2[i]+which(fctor$NF[v[i]:v2[i+1]] <= (b[1])),4]<-5
  fctor[v2[i]+which(fctor$NF[v[i]:v2[i+1]] > b[1] & fctor$NF[v[i]:v2[i+1]] <= b[2]),4]<-4
  fctor[v2[i]+which(fctor$NF[v[i]:v2[i+1]] > b[2] & fctor$NF[v[i]:v2[i+1]] <= b[3]),4]<-3
  fctor[v2[i]+which(fctor$NF[v[i]:v2[i+1]] > b[3] & fctor$NF[v[i]:v2[i+1]] <= b[4]),4]<-2
  fctor[v2[i]+which(fctor$NF[v[i]:v2[i+1]] > b[4]),4]<-1

  print(i) #Contador (algo personal que siempre utilizo en loops)
}


###   Periodo de comprobación del cálculo:
#Periodo 1:
quantile(fctor$NF[v[1]:v2[1]],probs = c(0.2,0.4,0.6,0.8),na.rm = T)
fctor$NF[v[1]:v2[2]]
fctor$Fractiles[v[1]:v2[2]]
#Periodo 2:
quantile(fctor$NF[v[2]:v2[3]],probs = c(0.2,0.4,0.6,0.8),na.rm = T)
fctor$NF[v[2]:v2[3]]
fctor$Fractiles[v[2]:v2[3]]
#Periodo 15:
quantile(fctor$NF[v[15]:v2[16]],probs = c(0.2,0.4,0.6,0.8),na.rm = T)
fctor$NF[v[15]:v2[16]]
fctor$Fractiles[v[15]:v2[16]]

###   Podemos comprobar como el looop ha funcionado a la perfección, por lo que el siguiente paso es,
###   almacenar los valores dentro de la base de datos original, y llamar a nuestra función base.

data$NF<-fctor$NF
data$NF.Fractiles<-fctor$Fractiles
attach(data)

###   Llamada a la función aportando como parámetros de entrada, el nombre del factor que es objeto
###   de estudio, y los portfolio que deseeamos comprobar.
simulacion(obj = NF.Fractiles ,top=T,bottom = T)


###   Machine learning and random forest

###   Librerias y/o paquetes utlizados
```

```
library(rpart)
library(lme4)
library(party)
library(partykit)
library(ggplot2)
library(glm2)
library(car)
library(caret)
library(knitr)
library(rpart.plot)
library(randomForest)
library(inTrees)
library(MASS)
library(lmtest)
library(pscl)
library(stargazer)
library(data.table)

###RANDOM FOREST

###   En primer lugar, apuntar que necesitamos los casos completos, es decir, la función de R
###   no acepta NAS

###   Creamos un data set con todos los casos completos del data set original
enteros<-data[complete.cases(data),]

###   Por facilitar la computación extraemos aquellas variables que no son factores quantitativos
columnas<-c(1,3,7,8)
enteros<-enteros[,-columnas]

###   Primer periodo de computación
periodo1<-data[which(data$Period..YYYYMMDD.=="2000-12-29"),]
completeData <- periodo1[complete.cases(periodo1),]
attach(completeData)

###   Calculo del los bosques aleatorios para este periodo
bosque1<-randomForest(Universe.Returns~EY.Med.Ntma+DY.Fwd.NTMA+X12.1m.Price.Mom+NET.DEBT..EBITDA.NTM
                      +ROE.NTMA+Beta.6m.Daily+Volatility.1yr,data=completeData,proximity=TRUE,
                      importance=TRUE,type="classification")

###   Variables importantes
bosque1$importance #Tabla
require(extrafont)
varImpPlot(bosque1,main = "Variable Importance") #Gráfico

##Grafico de MSE VS ntrees para este periodo
MSE<-data.frame(N.trees=seq(1,500,by=1),MSE=bosque1$mse)
grafico<-ggplot(MSE) +
  geom_point(aes(x=N.trees,y=MSE,colour=MSE))

grafico+theme(text = element_text(size=6)) + # Tamaño de fuente del grafico por defecto
  ggtitle("% MSE decrease") + # Título del gráfico
  theme(plot.title = element_text(family="Comic Sans MS",
                                  size=rel(2), #Tamaño relativo de la letra del título
                                  vjust=2, #Justificación vertical, para separarlo del gráfico
                                  hjust=0.5,
                                   face="bold", #Letra negrilla. Otras posibilidades "plain", "italic", "bold" y "bold.italic"
                                  color="black", #Color del texto
                                  lineheight=1.5)) + #Separación entre líneas
  labs(x = "Number of trees",y = "MSE (%)") + # Etiquetas o títulos de los ejes
  #theme(axis.title = element_text(face="italic", colour="brown", size=rel(1.5))) # Tamaño de los títulos de los ejes
  theme(axis.title.x = element_text(face="bold", vjust=-0.5, colour="black", size=rel(1.5))) +
  theme(axis.title.y = element_text(face="bold", vjust=1.5, colour="black", size=rel(1.5)))

###   Ahora almaceno los resultados de las variables importantes
a<-importance(bosque1)
a<-data.frame(a[,1]) #  Como se ha explicado en el trabajo solo la primera

###   Extraigo el nombre del factor mas importante
b<-rownames(a)[apply(a, 2, which.max)];b

###   Segundo periodo de computación (Mismo procedimiento)
periodo2<-data[which(data$Period..YYYYMMDD.=="2001-01-31"),]

completeData <- periodo2[complete.cases(periodo2),]
attach(completeData)
bosque1<-randomForest(Universe.Returns~EY.Med.Ntma+DY.Fwd.NTMA+X12.1m.Price.Mom+NET.DEBT..EBITDA.NTM
                      +ROE.NTMA+Beta.6m.Daily+Volatility.1yr,data=completeData,proximity=TRUE,
                      importance=TRUE,
                      type="classification")
```

```
bosque1$importance
importance(bosque1)
varImpPlot(bosque1,main = "Variable Importance")

#Guardo en A las variables importancia
a<-importance(bosque1)
a<-data.frame(a[,1])
#Extraigo el nombre del factor mas importante
b<-rownames(a)[apply(a, 2, which.max)];b

###   Automatización del proceso para todos y cada uno de los periodos

###   Necesitamos conocer todas las fechas sin repetir que tenemos en el data set, para ello cargamos
###   el vector donde se contienen
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/fechas.RData")
fechas[1:4]

### Creación del loop donde se automatiza los procedimientos realizados para cada periodo anteriormente:
i=0

varimp<-c() # Vector donde almacenarempos la variable importante en cada uno de los peroidos
for(i in 1:length(fechas)){

  ### Cogemos fecha a fecha los complete cases
  periodo<-data[which(data$Period..YYYYMMDD.== fechas[i]),]
  completos<-periodo[complete.cases(periodo),]

  ### Calculamos los bosques aleatorios para cada fecha anterior
  bosque1<-randomForest(Universe.Returns~EY.Med.Ntma+DY.Fwd.NTMA+X12.1m.Price.Mom+NET.DEBT..EBITDA.NTM
                        +ROE.NTMA+Beta.6m.Daily+Volatility.1yr,data=completos,proximity=TRUE,  importance=TRUE,
                        type="classification")

  ### Guardamos los resultados (ambas medidas de importanci)
  importancia<-importance(bosque1)
  ### Especificamos que deseamos la INCMSE (explicado trabajo)
  importancia<-data.frame(importancia[,1])
  ### Almacenamos solo el nombre
  varimp[i]<-rownames(importancia)[apply(importancia, 2, which.max)]

  ### Comprobación in-situ del funcionamiento del loop (personal)
  print(varimp[i])
  print(i)
}

### Con el fin de no repetir el proceso las almacenamos
save(varimp1, file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/varimp1.RData")

###   A partir de ahora solo tendremos que cargarlas.
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/varimp1.RData")

###   Por saber qûe factores y el número han tenido cada una hacemos lo siguiente (curiosidad)
class(varimp1)
prueba<-factor(varimp1)
summary(prueba)

###   Comprobación primeras 5 variables importancia
varimp1[1:5]

###   Ahora hay que coger periodo a periodo los fractiles de las variables mas importantes
###   para cada uno de dichos periodos.

###   Creamos un data set inicial
prueba<-data.frame(data)

###   Creamos la posición donde almacenaremos los resultados para el mismo
prueba$RF<-c(rep(0,nrow(prueba)))

###   Como siempre hacemos, calculamos los dos primeros periodos de manera manual antes
###   de automatizar el procesp.

###   En el nuevo data set creado donde el periodo  coincida con la fecha almacenada en el vector
###   de fechas, almacenas los resultados del periodo anterior de la variable más importante
###   almacenada por el loop anterior.

###   Es decir, en el periodo siguiente, invertimos en función del factor más importante en el periodo
###   inmediatamente anterior.
prueba[which(prueba$Period..YYYYMMDD.==fechas[2]),28]<-prueba[which(prueba$Period..YYYYMMDD.==fechas[1]),varimp1[1]]


###   Automatización del proceso para todos los periodos
```

```
i=2
for(i in 2:length(varimp1)){

    prueba[which(prueba$Period..YYYYMMDD. == (fechas[i])),28]<-prueba[which(prueba$Period..YYYYMMDD.==fechas[i]),varimp1[i-1]]


}

###   Una vez calculados, de nuevo guardamos los resultados en la base de datos original, y hacemos
###   la llamada a la función base.
data$RF.Fractiles<-prueba$RF
attach(data)

####    Llamada a la función base con el parámetro de entrada esta vez del nuevo factor, y los
###   portfolio que deseeamos observar
simulacion(obj = RF.Fractiles ,top=T,bottom = T)

###   Finalmente ampliamos el rolling windows a 3 meses tal y como se explica en el trabajo, es decir,
###   la inversión en este periodo, se hará en función de la variable más importante durante los 3 periodos
###   inmeediantamente anteriores a este

###   Como el metodo de computación es exactamente el mismo, pasamos directamente a la automatizacion
###   del proceso para todos los periodos

###   Cargo las fechas sin repetir
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/fechas.RData")

### LooP

i=1
varimp3<-c() #Vector donde almacenaremos la variable importante en los 3 periodos
for(i in 1:(length(fechas)-2)){

  ###   Ahora el periodo de estudio reune 3 periodos en sí
  periodo<-subset(data, data$Period..YYYYMMDD. == fechas[i] | data$Period..YYYYMMDD. == fechas[i+1] |
                    data$Period..YYYYMMDD. == fechas[i+2])

  ###   Calculamos los casos completos
  completos<-periodo[complete.cases(periodo),]

  ###   Calculamos los bosques aleatorios
  bosque1<-randomForest(Universe.Returns~EY.Med.Ntma+DY.Fwd.NTMA+X12.1m.Price.Mom+NET.DEBT..EBITDA.NTM
                        +ROE.NTMA+Beta.6m.Daily+Volatility.1yr,data=completos,proximity=TRUE,
                        type="classification")

  ###   Obtenemos ambas medidas de importancia
  importancia<-importance(bosque1,type = 2)

  ###   Pero solo almacenamos la que nos interesa
  varimp3[i]<-rownames(importancia)[apply(importancia, 2, which.max)]

  ###   Comprobación personal del proceso
  print(i)
}

###   De nuevo una vez calculados procedemos a almacenar dicho vector
save(varimp3, file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/varimp3.RData")

###   Después solo debemos cargar dicho vector cuando se necesite con la siguiente función
load(file="E:/Universitat Autónoma Barcelona/Trabajo Fin de Grado/trabajo/varimp3.RData")

###   De nuevo comprobamos la frecuencia de los factores más importantes durante el horizonte
class(varimp3)
proba<-factor(varimp)
summary(proba)

###   Ahora necesitamos decirle que lo que queremos son los quintiles asociados a cada factor importante
###   no su valor. Se diferencian por el nombre

##Decimos que son los fractles no los coeficientes
varimp3<-c(paste(varimp3,"..Fractiles.",sep=""))
varimp3[1:5]

###   Creamos un data set nuevo de prueba
prueba3<-data.frame(data)

###   Creamos la posición donde almacenar los resultados
prueba3$RF3<-c(rep(0,nrow(prueba3)))

###   Computo del primer periodo manualmente
```

```
## Ha metido en los 3 primeros periodos, los correspondientes a la variable mas importante 1 (por tener algo)

prueba3[which(prueba3$Period..YYYYMMDD.==fechas[1]|prueba3$Period..YYYYMMDD.
              ==fechas[1+1]|prueba3$Period..YYYYMMDD. == fechas[1+2]),29]<-
prueba3[which(prueba3$Period..YYYYMMDD.==fechas[1]| prueba3$Period..YYYYMMDD.
              == fechas[1+1]|prueba3$Period..YYYYMMDD. == fechas[1+2]),varimp3[1]]

###   Automatizacion del proceso para cada periodo: Loop
i=3
for(i in 3:length(varimp3)){

  ### Es el mismo proceso anterior, solo que ahora dejamodos periodos por el medio en lugar de uno.
  prueba3[which(prueba3$Period..YYYYMMDD. == (fechas[i+1])),29]<-
    prueba3[which(prueba3$Period..YYYYMMDD. ==fechas[i+1]),varimp3[i-2]]
}

###   Una vez obtenidos los resultados seguimos el procedimiento habitual de guardarlos en el dataset
###   original, y haremos la llamada a la función.
data$RF3.Fractiles<-prueba3$RF3
attach(data)
simulacion(obj = RF3.Fractiles ,top=T,bottom = T)


#######     FIN   ##########
```