Università degli Studi di Pavia

Dipartimento di Matematica

PhD in Computational Mathematics and Decision Sciences - XXXV Ciclo

Machine Learning: from image recognition to financial markets

Relatori:
Chiar.ma Prof. Silvia Figini

PhD Thesis
Mishel Qyrana

Anno accademico 2021-2022

La verità ti renderà libero,
ma solo quando avrà finito con te.

David Foster Wallace

Contents

Abstract

This work collects the four papers that best represent my PhD path.
Feature selection variable based on multicollinearity was my very first paper. Starting from a measure born in the financial field and in particular referring to systemic risk, I reinterpreted it in the statistical field, providing an endogenous criterion based on Random Matrix Theory for choosing the fundamental parameter of the indicator. From this reinvention then I readapted the measure to the financial context from which it came, in a working paper that is not included here, where the modifications made to the original indicator proved to be fruitful in predictive terms.
Deep Learning approaches to detect damaged smartphone screens was the first project I worked on for Generali, and it introduced me deeply into Deep Learning topics. A partner of the company had requested a model for the recognition of damaged screens dmartphone from images. According to our knowledge, although the task seems to be quite common, nothing particularly significant in this sense has been proposed in literature. The final solution started from a well-known convolutional architecture, enriched and readapted for the specifici problem. The results obtained were extremely good in terms of performance, and the model went into production.
Invoices default forecasting for credit factoring was born from a project that I followed for the company KEDA S.r.l. from the very first developments - personnel selection and team training - up to the final production. The project aimed at enhancing the rich dataset of an Italian company, made by the records of a lot of transactions in the Italian business to business context. The idea was to propose a predictive model about the future default of the invoices, so to favour the implementation and risk management of credit factoring practices. This purpose required me - in addition to the aforementioned ones - to model the problem, define its criticalities and a priori elaborate the desirable properties of the solution, and finally to implement and test it. The results obtained were a marked improvement over the benchmark model employed by the company and provided by a well-known name in the fintech world, and the model passed the challenges of recognized academics and practitioners.
Generalized non linear low Beta portfolio strategies through autoencoder neural networks is the epitome of my PhD journey. In the short time hiatus following the project related to the detection of smartphones with damaged screens that had introduced me to Deep

Learning, it was very natural for me trying to experiment with these new techniques in the financial field, given my background. In a short time I discovered that Financial Artificial Intelligence was very different from simply applying AI algorithms on financial datasets. The exploration of these new paradigms on one hand projected me into what would have been my future research, on the other hand it also shed a precious retrospective light on what had been my previous studies. The paper and its idea arose from the perhaps chaotic but always constant jumble of research and ideas. In the work I examine an age-old problem of asset allocation and a widely documented winning anomaly on financial markets. Then I propose a market model based on an autoencoder neural network that offers a solution to the asset allocation problem by exploiting a data-driven generalization of the aforementioned anomaly. The prototype version of this model convinced Generali that it would be more fruitful to exploit the hybrid strangeness of my academic training, making me work on Financial AI.

Part I

Financial Machine Learning.

Chapter 1

Generalized non linear low Beta portfolio strategies through autoencoder neural networks.

In this paper we address the problem of asset allocation, trying to avoid the curse of dimensionality by selecting a small basket from a universe of securities, such that they can guarantee true diversification and preserve good portfolio performances. The selection criterion is based on a generalization of β anomaly. Starting from a universe of securities and momentum indicators, a market model is built on the basis of the endogenous factors obtained through the non linear structure of an autoencoder neural network. The securities are then sorted according to their mean square error with respect to their observed value. Low Generalized Beta stocks are therefore defined as those for which the mean square error is greater. The concept of low β proposed includes in this way securities that are at the same time poorly correlated with each other and with the starting universe itself, on the basis of the non-linear autoencoder model. The paper is then organized as follows: section 1 proposes a literature review on asset allocation, β anomaly and Financial Machine Learning techniques; in the second section the basis structure of autoencoder neural networks , the particular version used in the paper and the momentum indicators used as stock characteristics are presented, then the portfolio selection models and covariance estimation techniques across which the stock picking technique is tested are introduced; in the third section a main empirical application and out-of-sample backtest are discussed and analyzed from various perspectives, as well as some other applications; the fourth section collects the conclusions and some suggestions for future research.

1.1 Asset allocation, Beta anomalies and Financial Machine Learning.

A truly rigorous study on the asset allocation problem comes with a certain delay compared to the development of other branches of the economy. From the time of Adam Smith to the neoclassical economists, there has been focus only on how consumers allocate their resources by choosing between goods and services, in order to maximize their satisfaction given a certain spendable budget. The other typical consumer decision, how to choose investing between different activities, has been shrouded in nebulosity for years, based only on intuition and common sense - and then arbitrariness, Then deterministic tools through which we tried to explain the mechanism of decisions of the first type were unfit to deal with it.

Except for some attempts like Bachelier's [8] to find a probability measure of risk, or Fisher's 1906 [49] of constructing one distribution of expected returns, for a long time it was not possible to converge to a a theory that intertwined the concept of risk with that of yield. After Marschak's 1938 studies [83], the true Big Bang of the asset allocation theory was due to one of his students. A graduate in economics from the University of Chicago was looking for a thesis topic for his PhD. The student met a broker who suggested him to focus on the stock market. Harry Markowitz followed that advice, thus developing a theory that became the basis of financial economics and a revolutionary investment practice [82], effectively marking the beginning of the discipline of Mathematical Finance, on which about twenty years after Black, Scholes and Merton [22] would have placed another fundamental piece, this time in the context of option pricing.

The work produced then earned Markowitz the Nobel Prize in Economics in 1990. At a high level, what Markowitz did was to rationally identify and define the compromise that the investor must face: risk and expected return.

The problem of how to allocate the investor's wealth among a the securities on a given universe is called *Portfolio Selection*; hence the title of the fundamental and pioneering article by Markowitz, published in the March 1952 issue of the *Journal of Finance*.

In that article and in the subsequent works, Markowitz extends the techniques of linear programming for the development of the critical line algorithm. The critical line algorithm identifies all eligible portfolios that minimize risk, measured by variance or standard deviation, for a given level of expected return and, vice versa, they maximize the expected return for a given level of risk. The graphical representation of the portfolios thus constructed constitutes the so-called efficient frontier.

Markowitz developed mean-variance analysis in the context of selecting a portfolio of ordinary shares but, in the last two decades, this approach has been used in practice in an increasingly general sense, applying it to asset classes or to portfolios and ETFs. The reason for this use is given by the fact that a similar approach allows diversification by building the portfolio with a limited number of assets, thus mitigating one of the intrinsic

problems of the model, which we will discuss later.
Markowitz's main contribution was therefore to give a structure to a concept that until then had remained on an intuitive and naive level: the diversification of one's own investment portfolio.
Despite its simplicity, there is no single definition of diversification available in the literature, resulting in the production of many contributions on the subject. [115]
Beyond the definition, however, the idea behind diversification and the benefits it brings are clear.
Take for example a toy market consisting of only A and B stocks, with volatility measured by standard deviation σ_A and σ_B respectively, and with the correlation coefficient between the two stocks equal to ρ. Furthermore, suppose that you invest in each of the securities a percentage w_A and w_B respectively, and that these have expected returns μ_A and μ_B. Then the return of this portfolio will be given by the weighted average of the two expected returns:

$$mu_P = w_A\mu_A + w_B\mu_B$$

what is more interesting, however, is what happens to the standard deviation. Let's take the case in which the two stocks are perfectly correlated, that is $\rho = 1$, which means that the stocks have the same type of behavior, going to deviate from their average return moving on average in the same direction. By calculating in terms of variance, we will therefore have:

$$\sigma_P^2 = w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_Aw_B\sigma_A\sigma_B = (w_A\sigma_A + w_B\sigma_B)^2$$

hence the volatility will therefore be:

$$\sigma_P = w_A\sigma_A + w_B\sigma_B$$

that is, exactly as was the case for yields, the weighted average of the standard deviations of the individual securities. If, however, we were in the much more general case of $\rho < 1$, then the standard deviation of the new portfolio obtained would be lower by construction than the simple combination of the two single volatilities:

$$\sigma_P^2 = w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_Aw_B\sigma_A\sigma_B > w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_Aw_B\sigma_A\sigma_B\rho$$

Although extremely important for its pioneering nature in the problem of capital allocation, the Markowitz model presents several problems related to the assumptions underlying the model, which have made it unattractive for practioners. Among others, three significant problems of the model can be isolated.
Contrary to what is assumed in the hypotheses, the empirical analysis reveals that the Gaussian distribution is not a good probability density function for financial returns. [84]
The latter in fact reveal a substantially different shape in distribution, with higher densities concentrated around the average value, more drooping distribution shoulders and

above all thicker distribution tails, compared to the Gaussian. Then we have basically leptocuric distributions, which testify that the returns from the most extreme magnitude are actually much more probable than those postulated by the Gaussian associated with the given mean and variance. Therefore, since there is no Normality, the allocation problem cannot be based only on the first two moments of the distribution of returns. From this it is inferred that the underlying utility function for the investor - which is implicit in the Markowitz model - can no longer be quadratic. However, the limit of the Gaussian distribution is not insurmountable, on the one hand it is in fact possible to reduce the granularity of the returns by taking them for example on a monthly basis and on the other hand generalizations of the model have been proposed in the literature that also take into account the third-order moment [53] up to the moment of fourth order [122] linked to leptocurtosis. The problem with both these approaches, however, is revealed in the fact that they clash even more strongly than the Markowitz model with the other limitations of the model, as for example the curse of dimensionality.

About this, the historical covariance matrix is not a good estimate of the true covariance matrix, and this is problem number two. Once a certain number n of securities to be put in the portfolio is fixed, then the number of parameters to be estimated for the covariance matrix will be given by $\frac{n(n+1)}{2}$ [84]. The number of parameters to be estimated - and which physiologically carries within an error and noise component - grows quadratically with respect to the number of securities to be held in the portfolio. There are techniques and methodologies to filter the covariance matrices from erratic components without signal [41], [74], [72], [40] as well as various dimensionality reduction techniques. From this problem, it is easy to understand how the solutions offered for the non-Gaussianity of returns generate an even greater problem in terms of covariance matrix estimation. In fact, recording the returns on a monthly basis requires very long time series, which by construction must be based on information that is probably obsolete with respect to the time of the estimate; on the other hand, the generalizations proposed would require estimating the coasymmetry and cocurtosis matrixes, doubling the already hypertrophic number of parameters to be estimated and sharpening the effects of the so-called *curse of dimensionality* [122].

As happens for the covariance matrix, it can be shown that the average return recorded by a security on a historical basis is not a good estimate of the expected return of that security [87]. This is also somehow implicit in the very need to build a portfolio: if it were easy to correctly estimate the expected return of the securities and the variance were on the appropriate order of magnitude, then it would be known from time to time a priori on which securities to invest. Furthermore, the methodology by which to choose the required return for one's portfolio is not clear.

That said, both literature and practice have tried to deal with these problems. We already mentioned some covariance filtering techniques. For the non Gaussianity of returns, risk management techniques have been proposed, in order to mitigate the heavy-tails related problems.

For the choice of the optimal portfolio among those on the efficient frontier - which basically is equivalent to ask for an ideal required return - the maximum Sharpe Ratio portfolio have been proposed [107], [124]. This portfolio, even if is theoretically the best, in fact leads to very unbalanced portfolio, especially if unconstrained in terms of short positions. This is due to the fact that it relies on the expected returns even heavily than Markowitz. Indeed in the classical Markowitz model the expected returns are just used to build a constraint, while in the maximum Sharpe ratio portfolio they are part of the objective function.

The Minimum Variance portfolio is a first and natural restriction to the Markowitz model which allows in one fell swoop not having to choose a return required by the portfolio and not having to estimate the expected returns of the assets involved. From the point of view of optimization, the idea is to reformulate the problem by emancipating it from the constraint linked to returns.

Several other methods which exclude expected returns calculations from the portfolio construction and just focus on the risk part have then been proposed.

The Maximum Diversification portfolio [33] reinvents the allocation paradigm proposed by Markowitz, explicitly disregarding the role - and therefore the estimate - of expected returns. The immediate effect of such an approach is in the balancing of the weights obtained. In fact, the Markowitz portfolio tends not only to result in extreme weights, but often to allocate the largest part of the capital in a few stocks compared to those available, selecting for the others negligible allocations, and effectively reducing the real diversification obtained. On the other hand, the MD portfolio ensures that all the securities available rationally receive a non-negligible portion of the investor's resources.

Risk parity portfolios [101] can be read as a generalization of the equally weighted portfolio, but the purpose is no longer to allocate the same capital among the securities but to equally distribute the risk among the securities available.

Another different approach is provided by the Black-Litterman portfolio model [65]. It can be said that this other approach takes its roots precisely from the fact that, systematically, it is very difficult for a portfolio to beat, for example, an index such as the SP500. In light of this fact then, what the Black-Litterman model allows to start from a benchmark portfolio - for example a portfolio weighted by market capitalization (as in fact it is the SP500) or a naive portfolio, i.e. allocations that in history for one reason or another have always proved more effective than more sophisticated approaches - and then to modify these positions only on the base of specific and circumscribed beliefs and their confidence level.

Strictly connected to portfolio theory - especially from an historical point of view - Sharpe derived the CAPM model. Even if the model has deeper theoretical roots and consequences, a widely accepted and simple interpretation among the practitioners is that the risk of a stock can be entirely represented by a coefficient β which linearly relates the asset

performance with the related market performance.
A first methodology to estimate this β coefficient was in the historical [46], while a first well-known criticism was in [99], which highlights how the need of a proxy of the market portfolio - such as, for example SP500 - stultifies the proposed methodology.
A natural while conceptually different extension to the CAPM was given by the multi-factor risk analysis and so called Arbitrage Portfolio Theory, which was introduced in [44], [43] [45], [99] and [100]. Furthermore, a lot of literature about the identification of these factors have been proposed. [55] describes and analyzes the return predictive signals (RPS) publicly identified in the period 1970–2010 and shows how more than 330 signals have been reported and a lot of them show properties which are stable over time, as well as highlights how RPS with higher mean returns have both larger standard deviations and higher Sharpe ratios; [60] analyzes different factors, trying to discriminate them in a *lucky vs skill* discriminant analysis; [120] explores the out-of-sample performance of a lot of factors, finding that not a single one would have helped in outprediction.
One of the empirical facts that the CAPM struggles to explain is the so called *Beta anomaly*, that is the low β stocks have better adjusted performance than the high β ones. It's clear that this empirical evidence conflicts with the well-known risk-reward relationship. A lot of researches tried to explain this fact: in [31] the authors claims to have solved the Beta anomaly by showing that the conditional beta for the high-minus-low beta portfolio shows negative covariance with respect to the equity premium and a positive one with respect to the market volatility; in [15] is suggested that investors' demand for lottery-like stocks is a non negligible driver for the beta anomaly, which is no longer detected when portfolios ranked on the base of the Betas are neutralized to lottery demand or factor models including a lottery demand factor. This anomaly is then concentrated in stocks with low levels of institutional ownership and it persists only when the price impact of lottery demand is concentrated in high-beta stocks; [23] incorporates in a multifactor model the agency effect for which the equilibrium relation between CAPM beta and expected stock returns becomes flat, instead of linearly positive; [28] starts from the fact that the equity of a levered firm can be modeled as a call option on firm assets and that option returns are non-linearly related to underlying stock returns, and for this reason highlights how for these cases the linear CAPM-type regressions are generally misspecified; [4] highlights the well-known Beta anomaly by showing that stocks with recent past high idiosyncratic volatility have low future average returns around the world, and then shows that there is a strong covariation in the low returns to high-idiosyncratic-volatility stocks across countries, suggesting that some not easily diversifiable factors lie behind the phenomenon; a more general analysis about the implications in investing practices produced by the anomaly are analyzed in [36]; [5] observes that safer assets must offer higher risk-adjusted returns than riskier assets because the consumption of high risk-adjusted returns which characterize safer assets requires leverage, then producing an opportunity for investors which apply leverage, and furthermore highlight how Risk parity portfolios naturally exploit this fact

by construction; in [6] The authors refute the idea that low-risk investing delivers high returns because of industry bets that favor stable industries by showing that a strategy of betting against beta has delivered positive returns both as an industry-neutral bet within each industry and as a pure bet across industries; a further clue in this sense is provided in [11], where the authors decomposes the anomaly into micro and macro components, where the micro component comes from the selection of low-beta stocks, while the macro ones from the selection of low-beta countries or industries, and both parts contribute to the anomaly; in [12] the authors explain why leverage is inversely related to systematic risk; [13] collects a comprehensive list of evidences which support the Beta anomaly; in [3] the Beta is splitted in its two components - correlation and standard deviation ratio, and the Beta anomaly is exploited to show how to construct interesting portfolio strategies by relying on the aforementioned decomposition.

In the recent years, Machine Learning methodology strongly entered the asset allocation problem and, more in general, financial markets, investing and trading related problems. The usefulness and power of ML lie basically on two pillars strictly related: the possibility to deal with high-dimensional data and the capacity to decouple the search for features from the search for specification.

A lot of contributions in this sense have been provided by ML. Starting from penalized regressions such as Lasso [52], [20], [113], [118], Ridge, Elastic Net [125], Group Lasso [79], [7] to dimensionality reduction based techniques as PCR [9], [54] and PSR [68] to non parametric techniques such Random Forests [25], [26] as well as the more recent XGboost [32].

Last, also Deep Learning techniques based on Neural Networks have been tested and proposed in financial context. Neural Networks, with their property of universal aproximators [64] have shown huge potential in a lot of predictive tasks indeed.

A lot of these models, have been used in asset allocation, and in particular for asset pricing. A very comprehensive review in this sense is provided in [56].

In this paper we address the problem of asset allocation. The idea is to reduce the problems related to the curse of dimensionality by selecting a small basket from a universe of securities, such that they can guarantee real diversification and maintain good portfolio performances while remaining in a small number. The selection criterion is based on the mimesis of the strategies that exploit the β anomaly, based however on a more general concept of β which we propose here.

Starting from a universe of securities and features obtained endogenously, a market model is built on the basis of a handful of endogenous factors obtained through the non linear structure of an autoencoder neural network. Based on these factors, the securities are reproduced using linear models, and sorted according to their tracking error with respect to their observed value. Low β stocks are therefore defined as those for which the tracking error is greater. The concept of low β proposed includes securities that by construction are at the same time poorly correlated with each other and with the market as a whole, on the basis of the proposed non-linear autoencoder model.

1.2 Autoencoder neural network for low Beta stock picking and true diversification achievement

In this section the classic structure of the Autoencoders [80] is introduced, defining the modified structure of the Autoencoder proposed and its interpretation as a market model, as well as its use for stock picking procedure.

Neural network models characterized by a number L of hidden layers can be written in a recursive way.

Let K^l be the number of neurons which are in each layer l, with $l = 1, ..., L$, $r_t^{l,k}$ the output of neuron k in layer l as for a fixed observation t and the vector $r_t^l = (r_t^{l,1}, ..., r_t^{l,K(l)})'$ which collects all the outputs of the layer . At each hidden layer, on outputs from the previous layer are applied nonlinear transformations - activation functions - $g(\bullet)$ and the obtained result becomes the input for the next layer. To initialize the network, the input layer is feeded with the cross section of returns $r_0^t = (r_t^{0,1}, ..., r_t^{0,n})'$, where n is the number of securities in the investment universe. Then, for a layer l such that $l > 0$ the recursive output formula is:

$$r_t^l = g\left(b^{l-1} + W^{l-1} r_t^{l-1}\right) \tag{1.1}$$

where W^{l-1} is a $K^l \times K^{l-1}$ matrix of weight parameters, and b^{l-1} is a $K^l \times 1$ vector of the so-called bias parameters. These are tipically called *dense layers*.

Usually, the neural network defined needs three conditions to be an Autoencoder.

First, for a certain intermediate layer l^*, such that $0 < l^* < L$, $K^{l^*} << n$. This condition means that the number of neurons of an intermediate layer needs to be a lot smaller then the number of initial input data, which in our case are the returns of the n assets.

Second, $K^L = n$, the number of neurons in the final layer has to be equal to the number of columns in initial input data.

Last, the neurons in the final layer try to aproximate the initial input data, that is that the Autoencoder is looking for that set of parameters $\mathbf{W}, \mathbf{b}$ such to find:

$$\min \sum_{t=1}^{T} d(r_t^L, r_t^0) \tag{1.2}$$

where d is a generic distance measure.

An Autoencoder looks for a K^{l*} dimensional representation of the n dimensions given by the number of assets in the investment universe. The connection with the Principal Component Analysis is clear: both are dimensionality reduction techniques, the PCA is linear, and the Autoencoder is potentially non-linear due to the $g(\bullet)$ activation functions [14].

In fact, it can be easily proven that an Autoencoder model with a single hidden layer and a linear activation function corresponds to the PCA method [57]. So that, by extension, there is a clear connection among Autoencoders and latent factors for asset pricing models.
In our model, we have proposed a slightly modified version of a classic Autoencoder, including additional starting features to the logarithmic returns. The dataset is given then by a three-dimensional array (or tensor) $\mathbf{R}$, of size $T \times n \times M$, where T represents the number of observations, i.e. the time period under consideration, n the number of securities of the investment horizon and M the features considered for each security.
For a fixed time t and a given security i, let R_t^i to be the $M \times 1$ vector of features observations fo the i asset at times t. Then a linear combinations is applied to the features:

$$r_t^{0,i} = w_c' R_t^i \tag{1.3}$$

where w_c has dimensions $M \times 1$. Note that the weights of this linear combination are the same for all the n assets in the investment universe. In practice, this is achieved with a 1×1 convolutional layer [73], which extract a unique feature from the M original features, by applying the same combination for all the n stocks.
At this point, from $r_t^0 = (r_t^{0,1}, ..., r_t^{0,n})'$, the traditional Autoencoder part as previously described starts. Thus the modified version of Autoencoder consists of a first convolutional layer of size 1×1 and the rest of the standard structure, and then the parameters to be estimated are $\mathbf{W}$ and $\mathbf{b}$ from the proper Autoencoder part plus the M parameters w_c.
In particular, after a convolutional layer on top of which a ReLu activation function [1] is applied - that is $g(x) = max(0, x)$ - we have a single hidden layer with K neurons and linear activation function. The final layer has a linear activation function, and n neurons as the number of the securities in the investment universe.
Now, let the m-th among the M features to be that one corresponding to logarithmic returns, so that R_t^m is the $1 \times n$ vector which collects the n logarithmic returns at time t, then the loss function to be minimized is given by the Mean Square Error between the logarithmic returns and the final layer r^2:

$$\min \sum_{t=1}^{T} \left(R_t^m - r_t^2\right)^2 \tag{1.4}$$

which is equivalent to minimize the Frobenius norm of the difference among the $T \times n$ matrix of returns and the matrix provided by the final layer. The parameters are updated through Backpropagation [102], and the chosen optimization routine is Adam [69].

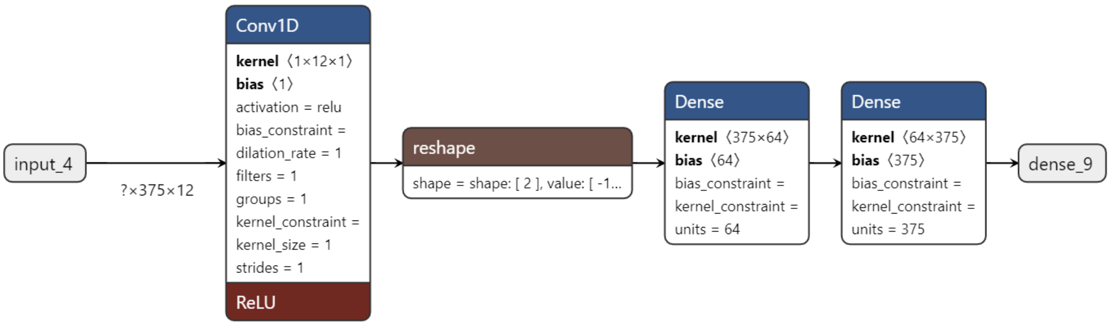

Figure 1.1: Modified Autoencoder Structure

Then, to summarize, the input data is R with dimension $T \times n \times M$. First, a convolutional 1×1 layer with ReLu activation function is applied to obtain r^0, with dimension $T \times n$. Next, a dense layer with linear activation function compress the previous layer output into r^1 which has $T \times k$ dimension. Finally, another dense layer with linear activation is applied, and r^3 with dimensions $T \times n$ is obtained. The objective function is the mean square error among the final layer r^2 and the R_t^m which collects the logarithmic returns for the T observations, across the n assets.

The neural network therefore linearly combines all the features for each security in the universe of assets, looking for a unique representation for all stocks. At this point, the ReLu function takes into account any non-linear relationships between the representations of the individual securities, and from these composes K latent factors. Starting from these, the model tries to reconstruct the series of returns of all n stocks.

The factors are then produced endogenously, starting from the non-linear representations of the combination of features for each security. They are therefore implicitly induced by the universe of assets considered.

The K latent factors must therefore reconstruct n series of logarithmic returns. For this reason the neural network has to *choose* the K factors such to serve for as many stocks as possible. In other words, each of the factors must be the basis from which reproducing on average more than $[n/K]$ series of returns.

The intuition is therefore that the logarithmic returns of those securities reproduced less accurately (with greater Mean Square Error) by the neural network share less exposure to implicit factors extracted from the investment universe, and are therefore less correlated. The lower correlation is both with respect to the other securities that make up the initial basket and within the group of securities with the greatest Mean Square Error. In fact, if these, taken in a sufficiently large number, had been correlated with each other, the Autoencoder would have been interested in producing factors capable of reconstructing them more accurately.

This duality allows to get at the same time a basket of asset which permits to achieve true diversification - through the fact that the securities are poorly correlated with each other - and at the same time poorly correlated with the market as a whole, encapsulating in this way a low beta strategy.

1.2.1 Generalized Correlations

Let preserve our previous definitions, so that $\mathbf{R}$ is the tensor which collects T observations for each of the M features across the n securities, and r_t^2 represents the cross-sectional logarithmic series, that is one of the M matrices feature-stacked to produce the tensor, so that $r_t^{2,j}$ is the same but for the particular security j. Define $\mathbf{f}(\mathbf{R})$ as a market model which tries to approximate r^2, i.e. the logarithmic returns of n stocks during the same homogeneous T period. Last, let R_i to be the i matrix of observations and features for the i-th asset and $\mathbf{R} \setminus R_i$ is the $T \times (n-1) \times M$ matrix which collects features and observations for all the securities except the i-th.
Then the following concepts are defined.

Definition 1.2.1 ($\epsilon(\theta)$-generalized pairwise correlation)**.** Let $\epsilon(\theta)$ to be a piece-wise continuous positive defined function of a generic set of parameters θ, then the asset i is $\epsilon(\theta)$-generalized pairwise correlated to asset j if:

$$\left|r^{2,j} - \mathbf{f}_j(\mathbf{R} \setminus R_i)\right| - \left|r^{2,j} - \mathbf{f}_j(\mathbf{R})\right| > \epsilon(\theta)$$

where || represents here a generic distance measure or loss function.

Definition 1.2.2 ($\delta(\xi)$-generalized correlation)**.** Let $\delta(\xi)$ to be a piece-wise continuous positive defined function of a generic set of parameters ξ, then the asset j is $\delta(\xi)$-generalized market correlated if:

$$\left|r^{2,j} - \mathbf{f}_j(\mathbf{R})\right| < \delta(\xi)$$

where again || represents a generic distance measure or loss function.

Notice that the properties of the definitions here provided are depending on the market model which is chosen by the researcher. Pairwise correlation for example, does not need to be symmetric.
At the same time, θ and ξ may collect parameters or hyperparameters of the market model itself. In our case for example, fix $\theta = \xi = \frac{nM}{sK^{l*}}$, where s is the number of securities to be included in our investment. Suppose for simplicity ϵ and δ are both $\mathbb{C}^1$ functions, then the following properties are desirable:

$$\begin{cases} \frac{\mathrm{d}\epsilon(\theta)}{\mathrm{d}\theta} \leq 0 \\ \frac{\mathrm{d}\delta(\xi)}{\mathrm{d}\xi} \geq 0 \end{cases}$$

In our case, all the choices in this sense are implicitely dictated by the choice of s, such that $s >>> \frac{n}{K^{l*}}$, taking n, M, and K^{l*} as given; at the same time, keep s sufficiently small to avoid market correlations arising. After a certain treshold for s, then both market and pairwise generalized correlations arise, until the maximum for both the correlations, which is recorded for the naive case $s = n$.

1.2.2 Connors RSI

In order to generate the features for each asset we chose to use the Connors Relative Strength Index - from now CRSI - is used. CRSI is a momentum oscillator which generalizes the well known Relative Strength Index - from now RSI.
Consider a series r_t, with $t \in [0, T]$, then the CRSI is the average of three distinct components which capture different momentum concepts with different memory lengths:

$$C(r_t; p, q, k) = \frac{1}{3}R(r_t, p) + \frac{1}{3}S(r_t, q) + \frac{1}{3}P(r_t, k) \tag{1.5}$$

with $R, S, P : \mathbb{R} \to [0, 1] \subset \mathbb{R}$, where $p, q, k \in \mathbb{N}_+$ are the time related parameters which define the memory length for each component.
The R component represent the classic RSI with exponentially weighted moving average. In particular, let r_t^+ and r_t^- to be defined as:

$$r_t^+ = \max\{r_t, 0\},\ r_t^- = |\min\{r_t, 0\}| \tag{1.6}$$

and ρ_t^+ and ρ_t^- as:

$$\rho_t^+ = \alpha r_t^+ + (1 - \alpha)\rho_t^+ \tag{1.7}$$

$$\rho_t^- = \alpha r_t^- + (1 - \alpha)\rho_t^- \tag{1.8}$$

with $\alpha = \frac{2}{p+1}$, then:

$$R(r_t, p) = \frac{\rho_t^+}{\rho_t^+ + \rho_t^-} \tag{1.9}$$

Let define the following indicators:

$$I_t^+ = \begin{cases} 1 \ \mathit{with} \ r_t \geq 0 \\ 0 \ \mathit{otherwise} \end{cases} \quad I_t^- = \begin{cases} -1 \ \mathit{with} \ r_t < 0 \\ 0 \ \mathit{otherwise} \end{cases} \tag{1.10}$$

The indicator I_t^+ is a vector of length T which at any time t highlights if the value r_t is non negative, while I_t^- is a signal for a negative value of r_t. Then:

$$\tau_{t+1}^+ = I_{t+1}^+(I_{t+1}^+ + \tau_t^+),\ \tau_{t+1}^- = |I_{t+1}^-|(I_{t+1}^- + \tau_t^-) \tag{1.11}$$

and finally:

$$S(r_t, q) = R(\tau_t^+ + \tau_t^-, q) \tag{1.12}$$

so that S measures how much positive values consecutive days have been recorded from the last negative value, updating the count each day.
$P(r_t, k)$ is the percentile score value of r_t with respect to the values $\{r_{t-i}\}_{i=0}^k$.
CRSI measures the series momentum from different perspectives. The first one is the classic RSI, which informs if on average the positive values are dominating or not the negative ones. The second component measures the dominance of the positive days on a streak against the negative ones. While the first component is influenced by the magnitude of the values, this second one captures how long the positive periods have last with respect to the negative ones. The last term completely changes the perspective, by measuring the how muche extreme the value is compared to the last k periods.

1.2.3 Asset allocation models.

Markowitz Model.

Let $\mathbf{S} \subset \mathbf{R}^{T \times n}$ to be the matrix which collects for T periods the price of n assets, with $T > n$. Let $\mathbf{X} \subset \mathbf{R}^{(t-1) \times n}$ to be the log-returns matrix, i.e. $X_{i,j} = \log S_{i+1,j} - \log S_{i,j}$, per $i = 1, ..., t$ e $j = 1, ..., n$, with $T - 1 \equiv T > n$, then we define the vector of expected returns $\mathbf{r} \subset \mathbf{R}^{n \times 1}$ as $r_{j,1} = \frac{1}{t} \sum_{i=1}^{t} X_{i,j}$, with $j = 1, ..., n$. Then we define the Covariance Matrix as $\Omega = \mathbf{X}^{*\prime}\mathbf{X}^{*}$ with $\mathbf{X}^{*}$ equal to $\mathbf{X}$ demeaned, where $'$ represents the transpose, and suppose Ω is invertible.

Finally, the return required by the investor from its portfolio w is R_p and the risk associated with it is σ_p^2, and they are defined as follows:

$$R_p = \mathbf{w}'\mathbf{r}$$

$$\sigma_p^2 = \mathbf{w}'\mathbf{\Omega}|\mathbf{w}$$

In particular, Markowitz's idea was to build a model that minimizes portfolio risk for a given return or, conversely, maximizes the portfolio's expected return for a given level of risk. It is therefore a constrained optimization problem that can be formulated in two specular ways:

Min σ_p^2	Max R_p
s.t	s.t
$w'\mathbf{1} = 1$	$w'\mathbf{1} = 1$
$w'\mathbf{r} = R_p$	$w'\mathbf{\Omega}w = \sigma_p^2$

We choose the first approach, which is the most used in practice. [34].

To find the variance that minimizes the risk of the portfolio, the above constrained optimization problem can be solved by exploiting the Lagrange multiplier method, for which:

$$L = \mathbf{w}'\mathbf{\Omega}\mathbf{w} - \lambda_1(\mathbf{w}'\mathbf{r} - R_p) - \lambda_2(\mathbf{w}'\mathbf{1} - 1)$$

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{\Omega}\mathbf{w} - \lambda_1\mathbf{r} - \lambda_2\mathbf{1} = \mathbf{0} \tag{1.13}$$

$$\frac{\partial L}{\partial \lambda_1} = R_p - \mathbf{r}'\mathbf{w} = 0 \tag{1.14}$$

$$\frac{\partial L}{\partial \lambda_2} = 1 - \mathbf{w}'\mathbf{1} = 0 \tag{1.15}$$

From the equation (1.13):

$$\mathbf{w} = \frac{1}{2}\mathbf{\Omega}^{-1}(\lambda_1\mathbf{r} + \lambda_2\mathbf{1}) = \frac{1}{2}\mathbf{\Omega}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix} \tag{1.16}$$

In the equation (1.16) we reported the term $(\lambda_1\mathbf{r}+\lambda_2\mathbf{1})$ in matricial terms in order to get $\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}$ from (1.14) and (1.15). In this way (1.14) and (1.15) can be rewritten as:

$$\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}'\mathbf{w} = \begin{bmatrix}R_p\\ 1\end{bmatrix} \tag{1.17}$$

For sake of simplicity we define:

$$\mathbf{A} \equiv \begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}'\mathbf{\Omega}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix} \tag{1.18}$$

the symmetric matrix 2×2 constitued by:

$$\begin{bmatrix}a & b\\ b & c\end{bmatrix} = \begin{bmatrix}\mathbf{r}'\mathbf{\Omega}^{-1}\mathbf{r} & \mathbf{r}'\mathbf{\Omega}^{-1}\mathbf{1}\\ \mathbf{r}'\mathbf{\Omega}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}\end{bmatrix} \tag{1.19}$$

Given that $\mathbf{A}$ is positive definite for each y_1, y_2 because of:

$$\begin{aligned}\begin{bmatrix}y_1 & y_2\end{bmatrix}\mathbf{A} &= \begin{bmatrix}y_1 & y_2\end{bmatrix}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}'\mathbf{\Omega}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}\\ &= \begin{bmatrix}y_1\mathbf{r}+y_2\mathbf{1}\end{bmatrix}'\mathbf{\Omega}^{-1}\begin{bmatrix}y_1\mathbf{r}+y_2\mathbf{1}\end{bmatrix} > 0\end{aligned}$$

we can write

$$\frac{1}{2}\mathbf{A}\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}R_p\\ 1\end{bmatrix}$$

Given that $\mathbf{A}$ is invertible, we can obtain the Lagrange multipliers:

$$\frac{1}{2}\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \mathbf{A}^{-1}\begin{bmatrix}R_p\\ 1\end{bmatrix} \tag{1.20}$$

Thus the portfolio $\mathbf{w}$ which minimizes the portfolio risk for the given return is obtained as:

$$\begin{aligned}\mathbf{w} &= \frac{1}{2}\mathbf{\Omega}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}\\ &= \mathbf{\Omega}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}\mathbf{A}^{-1}\begin{bmatrix}R_p\\ 1\end{bmatrix}\end{aligned}$$

Then, by substituting the required return R_p, we can compute the minimal risk σ_p^2 of the portfolio:

$$\begin{aligned}\sigma_p^2 &= \mathbf{w}'\mathbf{\Omega}\mathbf{w}\\ &= \begin{bmatrix}R_p & 1\end{bmatrix}\mathbf{A}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}'\mathbf{\Omega}^{-1}\mathbf{\Omega}\mathbf{\Omega}^{-1}\begin{bmatrix}\mathbf{r} & \mathbf{1}\end{bmatrix}\mathbf{A}^{-1}\begin{bmatrix}R_p\\ 1\end{bmatrix}\\ &= \begin{bmatrix}R_p & 1\end{bmatrix}\mathbf{A}^{-1}\begin{bmatrix}R_p\\ 1\end{bmatrix}\end{aligned}$$

Minimum Variance

The Minimum Variance portfolio is a first and natural restriction to the Markowitz model which does not ask you to choose a required return from the portfolio and does not employ the expected returns of the assets involved. From the point of view of optimization, the idea is to reformulate the problem by emancipating it from the constraint linked to returns, namely:

$$\text{Min } \sigma_p^2 \textit{ such that } \langle w, \mathbf{1} \rangle = 1 \tag{1.21}$$

where $\sigma_p^2 = \mathbf{w}'|\mathbf{\Omega}\mathbf{w}$. The idea therefore is to simply find the Minimum Variance portfolio without binding it to the behavior of the stock returns. The rationale behind such a proposal is linked to various empirical analyzes that have shown that portfolios with low variance tend to have more satisfactory long-term returns. From a formulistic point of view, this paradigm shift can be trivially implemented in the formula by making the following reasoning: to force the optimization to focus solely on the variance minimization problem, it is sufficient to assume that the expected returns of the individual assets are identical, so that the yield does not become a distinction, which is equivalent to writing:

$$\mathbf{r} = \mathbf{1}$$

and:

$$R_p = 1,$$

from which we can immediately infer the resulting formula:

$$\mathbf{w} = \mathbf{\Omega}^{-1} \begin{bmatrix} \mathbf{1} & \mathbf{1} \end{bmatrix} \mathbf{A}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

with the $\mathbf{A}$ suitably modified by making the aforementioned replacements in terms of the vector of the expected returns and the required return.

Maximum Sharpe

The so-called Efficient Frontier of portfolios is made up of all those portfolios for which, given a certain return, the minimum possible variance is associated with it and, conversely, given a certain variance, the highest achievable return is coupled to it. The Efficient Frontier is therefore defined by all those portfolios that solve the two optimization problems posed by the Markowitz model, and assumes the formula of a parabola with axis of symmetry parallel to the abscissa axis, where this hosts the variances while the corresponding ordinate axis records returns. However, it is worth asking whether, among this plethora of efficient portfolios, there is one better than others. A natural answer may be to identify the portfolio with the best associated risk-reward ratio. The problem set up up to now therefore changes form: the lowest possible variance is no longer sought, keeping the required yield fixed, for example, but a fruitful equilibrium is sought between the two measures. This ratio is called the Sharpe Ratio, and therefore has the form:

$$SR = \frac{w'\mathbf{r}}{\sqrt{w'\mathbf{\Sigma}w}} = \frac{R_p}{\sigma_p}. \tag{1.22}$$

in the case in which the risk-free rate is set to be 0, as it will be considered in the present work. In order to find the portfolio $\mathbf{w}$ wihch maximizes the Sharpe Ratio $\frac{R_p}{\sigma_p}$ we have to solve the following optimization problem:

$$argmax_{\mathbf{w}} \frac{\mathbf{w'r}}{\sqrt{\mathbf{w'\Omega w}}} \tag{1.23}$$

with $\mathbf{w^T 1} = 1$. This is equivalent to minimizing the Lagrangian:

$$L = -\frac{\mathbf{w'r}}{\sqrt{\mathbf{w'\Omega w}}} - \lambda(\mathbf{w'1} - 1) \tag{1.24}$$

And by exploiting the power operations properties we can rewrite the Lagrangian in a more useful way:

$$L = -\frac{\mathbf{w'r}}{(\mathbf{w'\Omega w}^{1/2})} - \lambda(\mathbf{w'1} - 1) \tag{1.25}$$

Then, the first order conditions are given by:

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = \frac{-\mathbf{r}(\mathbf{w'\Omega w})^{1/2} + \mathbf{w'r}\frac{1}{2}(\mathbf{w'\Omega w})^{-1/2}}{(\mathbf{w'\Omega w}^{1/2})^2} - \lambda\mathbf{1} = 0 \\ \frac{\partial L}{\partial \lambda} = \mathbf{w'1} - 1 = 0 \end{cases} \tag{1.26}$$

The first of the two conditions can be rewritten by using the power properties:

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{-\mathbf{r}(\mathbf{w'\Omega w})^{1/2} + \mathbf{w'r}\frac{1}{2}(\mathbf{w'\Omega w})^{-1/2}\mathbf{\Omega w}}{\mathbf{w'\Omega w}} - \lambda\mathbf{1} =$$

$$= -\mathbf{r}(\mathbf{w'\Omega w})^{1/2-1} + \mathbf{w'r}\frac{1}{2}(\mathbf{w'\Omega w})^{-1/2-1}\mathbf{\Omega w} - \lambda\mathbf{1} =$$

$$= -\mathbf{r}(\mathbf{w'\Omega w})^{-1/2} + \mathbf{w'r}\frac{1}{2}(\mathbf{w'\Omega w})^{-3/2}\mathbf{\Omega w} - \lambda\mathbf{1} =$$

$$= -\frac{\mathbf{r}}{(\mathbf{w'\Omega w})^{1/2}} + \frac{\mathbf{w'r}}{\frac{1}{2}(\mathbf{w'\Omega w})^{3/2}}\mathbf{\Omega w} - \lambda\mathbf{1}$$

so that we have:

$$\frac{\partial L}{\partial \mathbf{w}} = -\frac{\mathbf{r}}{(\mathbf{w'\Omega w})^{1/2}} + \frac{\mathbf{w'r}}{(\mathbf{w'\Omega w})^{3/2}}\mathbf{\Omega w} - \lambda\mathbf{1} = 0 \tag{1.27}$$

where $\frac{1}{2}$ vanished because the equivalence is invariant to it. Now, if we multiply from the left by $\mathbf{w'}$ we are able to obtain the value of λ:

$$-\frac{\mathbf{w'r}}{(\mathbf{w'\Omega w})^{1/2}} + \frac{\mathbf{w'r}}{(\mathbf{w'\Omega w})^{3/2}}\mathbf{w'\Omega w} - \lambda\mathbf{1} = 0$$

$$\lambda\mathbf{1} = 0$$

so that $\lambda = 0$ necessarily, and then by pluggin in the valute of λ the first condition becomes:

$$\frac{\partial L}{\partial \mathbf{w}} = -\frac{\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{1/2}} + \frac{\mathbf{w}'\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{3/2}}\boldsymbol{\Omega}\mathbf{w} = 0 \tag{1.28}$$

If we now premultiply it by Ω^{-1} we obtain the following equation:

$$-\frac{\boldsymbol{\Omega}^{-1}\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{1/2}} + \frac{\mathbf{w}'\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{3/2}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\mathbf{w} = 0$$

$$-\frac{\boldsymbol{\Omega}^{-1}\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{1/2}} + \frac{\mathbf{w}'\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{3/2}}\mathbf{I}\mathbf{w} = 0$$

$$-\frac{\boldsymbol{\Omega}^{-1}\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{1/2}} + \frac{\mathbf{w}'\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{3/2}}\mathbf{w} = 0$$

Now we express the solution w in its most natural implicit form, by rearranging the terms and multiplying both sides by $\frac{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}^{3/2}}{\mathbf{w}'\mathbf{r}}$:

$$\frac{\mathbf{w}'\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{3/2}}\mathbf{w} = \frac{\boldsymbol{\Omega}^{-1}\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{1/2}}$$

$$\frac{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}^{3/2}}{\mathbf{w}'\mathbf{r}}\frac{\mathbf{w}'\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{3/2}}\mathbf{w} = \frac{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}^{3/2}}{\mathbf{w}'\mathbf{r}}\frac{\boldsymbol{\Omega}^{-1}\mathbf{r}}{(\mathbf{w}'\boldsymbol{\Omega}\mathbf{w})^{1/2}}$$

from which we have:

$$\mathbf{w} = \frac{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}}{\mathbf{w}'\mathbf{r}}\boldsymbol{\Omega}^{-1}\mathbf{r} = \left(\frac{\mathbf{w}'\mathbf{r}}{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}}\right)^{-1}\boldsymbol{\Omega}^{-1}\mathbf{r} \tag{1.29}$$

Now we multiply from left by $\mathbf{1}'$, so that we obtain:

$$\mathbf{1}'\mathbf{w} = \left(\frac{\mathbf{w}'\mathbf{r}}{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}}\right)^{-1}\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{r}$$

But now we can exploit the condition: $\mathbf{1w} = 1$, so that we get:

$$1 = \left(\frac{\mathbf{w}'\mathbf{r}}{\mathbf{w}'\boldsymbol{\Omega}\mathbf{w}}\right)^{-1}\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{r} \tag{1.30}$$

Now we are almost done. The only thing left to do is to express the solution which was in the implicit form by making use of this last equation:

$$w = 1\frac{\boldsymbol{\Omega}^{-1}\mathbf{r}}{\mathbf{1}'\boldsymbol{\Omega}\mathbf{r}} \tag{1.31}$$

Which can be immediately verified by substituting the correspondent expression for 1 and check that we can go back to the solution in the implicit form. The portfolio thus obtained has the advantage of freeing the investitor from the embarrassment of having to set a certain level of return required as a side effect of a more complete and rational choice criterion, on the other hand, however, differently from other models that will be addressed here, it depends on somehow even more strongly by a sensible estimate of expected returns. For this reason, Sharpe's portfolio usually suffers as much and more than Markowitz from a further shadow problem: the portfolio weights obtained are decidedly extreme and a rational investor could hardly rely on them. For this reason, in the empirical application we will restrict the model to positive weights only.

Equally Weighted

The Equally Weighted portfolio, also called *naive* portfolio, is obtained by choosing w such that $w_i = \frac{1}{n}$ for each $i = 1, ..., n$. Surprising as it may be, such a portfolio proved to be a difficult benchmark to beat for far more sophisticated and complex models, as well as some of those presented here [39]. The reasons for this success, however, on closer inspection are less obscure and more intuitive than it may seem. In fact, such a model totally renounces any sophistication in the face of the factual problems of correctly estimating expected returns and covariances. In a certain sense, in its brutality and rigidity, it has the advantage of not offering the side in any way to errors of estimation and at the same time offering - at least potentially - a certain diversification.
A further reason to include this model in this analysis, is that it is particularly useful to assess a Stock Picking strategy, being free from estimation troubles.

Risk Parity

Portfolios at Equal Risk [101] can be read as a generalization of the Equally Weighted portfolio, the purpose of which is no longer to allocate the same capital among the n securities on which it is chosen to invest but instead of equally distributing the risk among the securities available. And it is precisely this interpretation that provides the correct formalization of the problem. In fact, let p_i be defined as follows:

$$p_i = \frac{w_i^2 \sigma_i^2 + \sum_{j,i \neq j}^{n} w_i w_j \sigma_{ij}}{\sigma_p}$$

where σ_i^2 is the variance of the i-th asset, σ_{ij} is the i-th row and j-th column of Ω - that is, the covariance among the i-th and j-th asset - and σ_p^2 the portfolio variance. It is therefore evident that p_i is the relative risk contribution brought by the i-th security to the total portfolio risk. The Risk Parity portfolio is then obtained as that vector w such for which it holds:

$$p_i = \frac{1}{n}$$

for each $i = 1, ..., n$. The Risk Parity portfolio therefore focuses exclusively on the risk component and its balanced distribution, being moreover intrinsically constructed to restrict the possibilities of short selling.

Maximum Diversification Portfolio.

The Maximum Diversification portfolio [33] reinvents the allocation paradigm proposed by Markowitz, explicitly disregarding the role - and therefore the estimate - of expected returns. The immediate effect of such an approach is in the balancing of the weights obtained. In fact, the Markowitz portfolio tends not only to result in extreme weights, but often to allocate the largest part of the capital in a few stocks compared to those available, deliberating for other decidedly negligible allocations, and effectively reducing the real diversification obtained. On the other hand, the MD portfolio ensures that all the securities available rationally receive a non-negligible portion of the investor's resources, and in the reference paper it also consists of a constraint on short selling that will be investigated more generally in the empirical section. In general, however, the w portfolio of maximum diversification is the one obtained by maximizing the so-called diversification ratio:

$$D = \frac{\mathbf{w}'\sigma}{\sqrt{\mathbf{w}'\mathbf{\Omega}\mathbf{w}}} \tag{1.32}$$

where σ collects the standard deviations of the n assets in portfolio.

Probabilistic Sharpe Ratio

Sharpe ratio Efficient Frontier permits the selection of optimal portfolios under non-Gaussian and leveraged returns. The portfolio optimization differs from other higher-moment methods because skewness and kurtosis are incorporated through the standard deviation of the Sharpe ratio estimator. This avoids making arbitrary assumptions about the relative weightings that higher moments have in the utility function.
SEF can be explained as that set of portfolios that maximizes the expected Sharpe Ratio for different confidence levels. The maximum Sharpe ratio portfolio lies in the SEF, but it may differ from the portfolio that maximizes the PSR. PSR is valid under stationary and ergodic returns [85] and [92].
So, in PSR asset allocation model, the Sharpe Ratio is treated as an estimator and then it is a random variable. The distribution of the Sharpe Ratio estimator $\widehat{SR}(\mathbf{w}, \mathbf{X})$ which depends on the weights $\mathbf{w}$ and on the matrix of returns $\mathbf{X}$ is Gaussian even if the returns X are far from Gaussianity, and the $\widehat{SR}$ variance depends on the returns skewness and kurtosis. So, what we are looking for is in fact:

$$\mathbf{w}* = argmax_{\mathbf{w}} \frac{\widehat{SR}(\mathbf{w}, \mathbf{X})\sqrt{n-1}}{\sqrt{1 - \gamma_3(\mathbf{w}, \mathbf{X})\widehat{SR}(\mathbf{w}, \mathbf{X}) + \frac{\gamma_4(\mathbf{w}, \mathbf{X})-1}{4}\widehat{SR}(\mathbf{w}, \mathbf{X})}} \tag{1.33}$$

where $\gamma_3(\mathbf{w}, \mathbf{X})$ and $\gamma_4(\mathbf{w}, \mathbf{X})$ are respectively the skewness and the kurtosis of the portfolio $\mathbf{w}$ on stocks with returns $\mathbf{X}$.
For a detailed treatment of this model, see [10], and for a derivation of the Sharpe Ratio estimator distribution see [78].
The optimization has to be numerical, and gradient descent based-algorithms are suitable for the task.

Hierarchical Risk Parity

The condition number of a covariance (or correlation matrix) is lowest for a diagonal correlation matrix, which is its own inverse. As add correlated investments are added, the condition number grows. At some point, the condition number is so high that numerical errors make the inverse matrix too unstable, that is a small change on any entry will lead to

a very different inverse. This is another curse for portfolio models: the more correlated the investments, the greater the need for diversification, and yet the more likely the solutions will be unstable. The estimation erros totally dominate the benefits of diversification.
As already said by mentioning the *curse of dimensionality*, an increase the number of assets will only jeopardize the problem, as each covariance coefficient is estimated with fewer degrees of freedom.
Furthermore, correlation structures do not remain invariant over such long periods. A clear clue in this direction is given by the overperformance of the Equally Weighted portfolios with respect to th mean-variance models and risk-based optimization out-of-sample. These instability concerns have received substantial attention in recent years as the several proposed filtering techniques for covariance matrix testify.
One reason for the instability of quadratic optimizers is that the vector space is modelled as a complete (fully connected) graph, where every node is a potential candidate to substitute another. In algorithmic terms, inverting the matrix means evaluating the partial correlations across the complete graph. Small estimation errors are magnified, leading to incorrect solutions. Intuitively it would be desirable to drop unnecessary edges, given that some investments seem closer substitutes of one another, and other investments seem complementary to one another.
And yet, to a correlation matrix, all investments are potential substitutes to each other. In other words, correlation matrices lack the notion of hierarchy.
Let $\mathbf{X}$ to be the $T \times n$ matrix which collects the returns for T periods of n assets, and let ρ to be its correlation matrix whith entries ρ_{ij} with $i, j \in \{1, ..., n\}$. Then we define a proper distance measure - for a proof that this is a proper distance measure see [38] -

$$d(X_i, X_j) = \sqrt{\frac{1}{2}(1 - \rho_{ij})} \tag{1.34}$$

with d that have values in $[0, 1]$ by construction. In this way we can construct D, a $N \times N$ symmetric matrix which collects the correlation-based distances among the assets, which entries are of course d_{ij}.
Now we need an Euclidean distance d_d, which is applied this time to the columns of D, and that has to be interpreted as a distance of distances:

$$d_d(D_i, D_j) = \sqrt{\sum_{k=1}^{n} (d_{ki} - d_{kj})^2} \tag{1.35}$$

and that by construction take values in $[0, \sqrt{n}]$. Now, we know that each row and column of the dymmetric matrix D_d which entries are $d_{d(i,j)}$, represents the Euclidean distance between the column vectors define by the correlation-based distances for the securities i and j, so that $d_{d(ij)}$ measures the difference between the correlation-based distance that the asset i has with respect to all the other securities and the correlation-based distance that the asset j has with all the other securities.

From matrix D, we will form the clusters that will arise in a hierarchical structure, through recursive bisection, so that each cluster can be constitued only by two elements: two securities, one security and one cluster, two clusters.
We start by identifying the first cluster, which will be given by:

$$(i*, j*) = argmin_{ij, i \neq j}(D_{d(i,j)}) \tag{1.36}$$

so that $(i*, j*) \in C(1)$ where $C(1)$ is the first cluster we identified.
In the next step we need to find out the second cluster. This one can be given by two other securities, of by one of the residual securities (so, nor $i*$ or $j*$) and the first cluster $C(1)$.
In order to do this, we have to update the matrix D_d, so that it includes not only the distances between the assets but also the distances between the assets and the new cluster, which have now to be thought as a new asset. This means that we have to choose a criteria which provides a reference for the cluster from which a distance can be inferred. There are several ways to do this, and we choose the so called *nearest point algorithm*, so that as reference for the distances among a given security and the first cluster, we take the distance among the security and each of the two assets in the cluster and then we choose the minimum value:

$$m_{i,C(1)} = \min(d_{d(i,j)_{j \in C(1)}}) \tag{1.37}$$

so that i is the given security, and j iterate over the two assets which are in the first cluster.
Now that we have a vector with a number of rows equal to the securities for which we computed the distance from the first cluster, we add this column vector as last column of D_d and its transpose as last row, and then we add a 0 along the diagonal which represents the distance among the cluster and itself.
At this point we eliminate from D_d the $i*$ and $j*$ rows and columns, because we don't need anymore those securities, given that they are already in $C(1)$.
At this point our D_d has been updated: it presents a new pseudo asset given by the first cluster which has encapsulated $i*$ and $j*$ which are no more available now, so that we can we can start again checking for the closest assets in D_d and repeat the procedure until each security belongs to a cluster.
At this point everything is set in order to obtain the portfolio weights. We start from the first cluster which collects two securities and compute some initial weights $\bar{\mathbf{w}}$ by selecting the weights as the inverse variance of each assets (normalized to sum to 1). Please note that this choice is equivalent to a Risk Parity subportfolio. This equivalence is clearly given by the fact that we only have to assets. Then we compute the total variance of the cluster, which is of course: $\sigma^2_{C1} = \bar{\mathbf{w}}' \mathbf{\Omega_{C1}} \bar{\mathbf{w}}$.
Now we proceed in the same way for the remaining securities. If the next cluster is among two other stocks, we repeat this procedure. If otherwise its among a stock and the first

cluster, the variance from which to computer the weight is given by the total variance of the cluster which we just defined. In this case this is equivalent to choose the Minimum Variance portfolio.

1.2.4 Covariance estimation

We have already discussed the problems of historical estimation of the covariance matrix and its inversion in asset allocation problems, detailing them in particular in 1.1 and 1.2.3. In short, given a number of assets, the estimate of the covariance matrix needs a sufficiently large number of observations so that the estimate is not distorted by the noise in the series. However, as the length of the observations increases, the data used sink too far into the past and the asset allocation models end up based on estimates that do not reflect the current situation.
For each of these problems, several notable solutions have been proposed in the literature. Here we will briefly explore two of them. The Exponentially Weighted Covariance Matrix which tries to limit the impact of older observations on the estimation and a filtering technique based on the Random Matrix Theory, which tries to reduce the impact of noise on short series. However, it is possible to effectively merge these two techniques, to obtain an estimation method that tries to tackle both problems jointly [93]. Since the model proposed in this paper offers a technique for stock selection so as to reduce the number of assets in the portfolio starting from a certain universe and thus contain the problem of estimating covariances while preserving the effectiveness of the portfolio, we have considered it appropriate to test it against alternative covariance estimation models, to verify if the model was still able to improve performance. We introduce very shortly these methods in the next two sections.

Exponentially weighted covariance matrix

While the sample covariance matrix attributes uniform weights to the co-deviations from the mean during the computations, the Exponentially weighted covariance matrix gives more weight to the last observations, and then the weight exponentially decreases with observations getting older.
In this way, more importance is attributed to more recent observations, but the true length of data is shortened, then making the estimation more prone to noise [93].

Covariance matrix with Random Matrix Theory based filtering

Let T to be the number of observations for n i.i.d. random variables, e.g. the T daily log returns for the n contituents of a portfolio. Then let $Q = T/n$, taking a particular limit for $T, n \to \infty$ such that $1 < Q < \infty$, the eigenvalues of the correlation matrix are distributed according to the Marcenko-Pastur distribution [81], [121] as well as bounded between a minimum and maximum eigenvalue:

$$\lambda_{\pm} = \left(1 \pm \sqrt{\frac{1}{Q}}\right)^2$$

where λ_- and λ_+ are respectively the minimum and maximum theoretical eigenvalues. In financial applications, the eigenvalues distribution has been widely used to deal with some of the well-known covariance matrix related issues in asset allocation models and the maximum theoretical eigenvalue has been used as treshold to clean covariance matrices through various recipes [72, 71, 29, 96]. The main idea is to filter those eigenvalues which are in the range among λ_- and λ_+, because they carry only noise, given that they are basically the same obtained with randomly i.i.d. generated process.
In practice, also the eigenvalues which are lower than the minimum eigenvalue are filtered out. This happens because usually there is not a clear separation among the last eigenvalues in the range and those smaller than λ_-.
The filtering techniques are various. We will use in particular the approach proposed in [96], for which all the eigenvalues lower than λ_+ are set to be 0.

1.3 Empirical analyzes

In this section several empirical analyzes and backtests are proposed to verify the model effectivness and if its underlying intuition is confirmed by data. Most of these validation processes are performed across different datasets, asset allocation models - those presented in **??** -, covariance estimation methods - introduced in 1.2.4 - and two different experiments setups.
Crossing different estimation techniques of the covariance matrix with asset allocation models helps to prove that the benefits of the stock-picking are novel and distinguished from those provided by portfolio selection and covariance estimation. For the maximum Sharpe Ratio model and the Maximum Diversification, only the case where short sales were restricted were included, because the free of constraints cases produced totally irrealistic weights and equity lines. Furthermore, for any case on which was required the risk-free rate, it has been considered as 0.
On the first setup the stock-picking is based on the MSE obtained by the securities on the train set, while on the second one the stock-picking is done with regard to MSE recorded on the validation set.
For each of these choices, two different kinds of backtests are presented.
In the first one, the portfolio models built on the basket of 40 securities obtained via autoencoder are tested against the portfolios obtained with the remaining securities. The first ones will be called *portfolio 40*, while those constructed on the residual assets will be called *portfolio n*, where n is the number of remaining assets, varying across the datasets employed.

The second analysis compares the portfolios obtained on the basket of 40 securities against other ones with the same number of assets. The ideal setting for this experiment would have been to compare portfolios 40 against all the possible combinations with the same number of stocks.

For the resources at our disposal, such an analysis would have been in fact not feasible because of the huge number of combinations. It was therefore decided to sort the residual assets in descending order based on Mean Square Error produced by the autoencoder model. Then the portfolios have been generated through 40 securities windows scrolled in steps of 1, so that the portfolios obtained gradually changed from those built on securities with highest MSE - excluding of course the 40 highest ones, selected through the autoencoder approach - up to those with smaller MSE values.

For this second analysis the employed asset allocation models have been restricted only to those which have a closed form solution, with the only exception of the maximum Sharpe ratio portfolio. This choice was again dictated by computational resources scarcity.

In order to measure the portfolios performances the following indicators have been used. The first four statistical moments of returns distribution, maximum and average drawdown, Value at Risk, conditional Value at Risk - both computed with historical method and for a 95% confidence interval -, Sharpe Ratio, Probabilistic Sharpe Ratio, Information Ratio, Sortino Ratio and Calmar Ratio.

In tables 1.1 and 1.2 the acronims used henceforth for asset allocation models and performance indicators are reported.

Abbreviation	Indicator
maxD	Maximum Drawdown
meanD	Average Drawdown
VaR	Value at Risk
cVaR	Conditional Value at Risk
mean	Mean Return
std	Standard Deviation
skew	Skewness
kurt	Kurtosis
SR	Sharpe Ratio
PSR	Probabilistic Sharpe Ratio (Treshold: 1)
SoR	Sortino Ratio
CR	Calmar Ratio

Table 1.1: Abbreviations for performance indicators.

Abbreviation	Portfolio
MAR	Markowitz
SR	Maximum Sharpe Ratio
MV	Minimum Variance
RP	Risk Parity
MD	Maximum Diversification
EW	Equally Weighted
PSR	Probabilistic Sharpe Ratio
HRP	Hierarchical Risk Parity

Table 1.2: Abbrevations for asset allocation models.

For the portfolio models, the suffix "NS" stands for "short sales are not allowed during the optimization" while a capital a into parenthesis - (A) - indicates that the portfolio have been constructed on the 40 securities selected through the Autoencoder approach. .
A negative value convention is used for the Drawdowns, Value at Risk and Conditional Value at Risk (both computed with the historical method and at 95% level); the mean return is multiplied by 100; the Probabilistic Sharpe Ratio has a treshold value equal to 1 if not differently specified.
Three other performance indicators were developed. Through the trend filtering algorithm based on Lasso penalty proposed in [90], the test sample has been labeled into bull market periods and bear market ones. Then the transition probabilities for the four possible combinations have been computed. On this basis the following indicators have been

constructed:

- **BullToBull/BearToBear**: this indicator is the ratio between the transition probability from a bull market day to another and the transition from a bear market day to another. Basically it compares the positive persistence of the equity lines with the negative one.

- **BearToBull/BullToBear**: this indicator is the ratio between the transition probability from a bear market day to a bull one and the transition from a bull market day to a bear one. Basically it compares the positive recovery of the equity lines with the instability of positive performances.

- **Bull period**: this indicator is the percentage of bull market days along the entire equity line.

Some last worlds on the methodology applied on the empirical analyzes and backtests. Everything was built and performed *first shoot* and there was no fine-tuning. The Autoencoder was immediately constructed with its final form. The Validation and Test set have not been explored in any way or form prior to their actual use. The same Train dataset was not analyzed but immediately used in order to obtain the parameters of the autoencoding neural network. The analysis proposed in 1.4.4 was carried out only at the end of all the backtests that precede it in the discussion.
These measures were adopted in order to reduce the probability of a backtest overfitting. Any potential bias present in the used datasets of which the author is aware has been reported in the appropriate section.
Every difference among the setups presented here and those employed in the backtests and experiments for each datasets will be disclosed in the proper sections.
Transaction costs, fees and taxation have not been considered. Given that the strategies presented require to buy assets only ones, transactions and fees can be negligible. Taxation problems are no beyond the scope of the present work. Furthermore, given that any issue related to taxation or other costs affect both the compared portfolios, they cancel out.

1.3.1 Features engineering

Let O_t, A_t, H_t, L_t to be respectively the open, adjusted close, high, low price and V_t the volume at day t for each asset, all of them strictly positive. We define furtherly E_t, which is the exponentially weighted moving average of the adjust closing price - henceforth we will use *close* and *adjusted close* interchangeably - with a parameter α:

$$E_t = \alpha A_t + (1 - \alpha)E_{t-1}$$

with $\alpha = \frac{2}{m+1}$ where $m \equiv 10$ in our case.
Then we define the following returns:

$$r_t = \ln(A_t) - ln(A_{t-1})$$

$$r_t^O = \ln(O_t) - ln(O_{t-1})$$

$$r_t^H = \ln(H_t) - ln(H_{t-1})$$

$$r_t^L = \ln(L_t) - ln(L_{t-1})$$

$$r_t^V = \ln(V_t) - ln(V_{t-1})$$

$$r_t^{HL} = \ln(H_t) - ln(L_{t-1})$$

$$r_t^{AE} = \ln(A_t) - ln(E_{t-1})$$

Then we computed the following Connors RSI:

$$C(r_t; p, q, k),\ C(r_t^O; p, q, k),\ C(r_t^L; p, q, k),$$

$$C(r_t^H; p, q, k),\ C(r_t^V; p, q, k),$$

with $P \equiv 5$, $Q \equiv 7$, $k \equiv 90$. Then the final features are all the returns previously computed except for r_t^{HL} and all the Connors RSI oscillators plus $P(r_t^{HL}, k)$, for a total of 12 features obtained endogenously from the initial dataset.
The idea behind the choice of these features is to inform the algorithm for each time t in which phase or cycle the security is, and how much it is supported by the volume trend. All the provided features are stationary.

1.4 SP500: 25-04-2007 to 24-09-2021

The main dataset collects the daily series of open, close, high, low and adjusted close prices, as well as the volumes of the 375 stocks of the SP500 that have been listed in the index continuously over the period ranging from 25-04-2007 to 24-09-2021. Note that the need to select these particular stocks includes a survival bias in the analysis, but in light of the available data providers the choice was unavoidable.
The time period has been chosen so to include both extremely strong crisis periods and very intense rally periods in the train, validation and test datasets, in order to make the analysis robust across different market conditions.
The train dataset ranges from 25-04-2007 to 26-05-2017, then collecting 2541 observations (70% of total), and it has been used to train the autoencoder neural network.
The validation set ranges from 30-05-2017 to 29-07-2019, then collecting 544 observations (15% of total), and it has been used to assess the autoencoder hyperparameters and to correctly stop its optimization, as well as to estimate the parameters for the portfolio models and covariance estimation techniques employed in this analysis.

The observations from 30-07-2019 to the last day are collected in the test dataset, for a total of 546 observations, used to perform the backtests.

1.4.1 Stock-picking on train set

This analysis compares the portfolios obtained on the 40 stocks with highest MSE recorded on the train set to those build on the remaining 335 stocks of the original universe.
The performance description is organized into sections based on the three methods for covariance matrix estimation.
In each section, first of all, an overview of the performances of portfolio 40 against portfolio 335 is presented, supported by the equity lines plots and a wide range of performance indicators.
Then, a focus on pairs of equivalent asset allocation models is presented: Markowitz on the 40 stocks with highest MSE is compared against Markowitz on the remaining 335, and so on for all the other models. The analysis of the pairs of portfolios is supported by the graphs of the equity lines and the QQ plot of returns. In the pairs comparisons, for sake of brevity, it was decided to keep as performance indicators only the main statistical moments of returns, maximum and average drawdown, Value at Risk, conditional Value at Risk and Sharpe Ratio. Anyway, the complete range of indicators is still present in the general overview.
Finally, a series of econometric tests seeks to assess whether or not there is a significant difference between the distributions of returns of the pairs of portfolios using the Kolmogorov-Smirnov test [18] and the k-sample Anderson-Darling test [104]. For the latter, p-value is capped at 25%, so that when this value is recorded it could be even higher (an asterisk is used as reminder each time).
The second type of analysis compare portfolio 40 with other portfolios of 40 assets. This procedure allows to rule out the hypothesys that portfolio 40 asset allocation models perform better just because the reduced number of assets, given that they are less prone to curse of dimensionality.
As said, the ideal analysis would have been to compare portfolio 40 against all possible combinations of 40 assets portfolios on the remaining 335 stocks. This operation is computationally too expensive so that a different procedure has been implemented.
Given the 335 stocks of the homonymous portfolio sorted in descending order based on the Mean Square Error, the stocks were selected in groups of 40, scrolling one security at a time. The first of these portfolios was therefore composed of the 40 stocks with the highest Mean Square Error among the 335 under examination. The second portfolio loses the stock with maximum Mean Square Error and acquires the 41st stock, and so on.
With respect to each of these assets baskets, Markowitz, Minimum Variance, Sharpe, Risk Parity and Equally Weighted weights only were calculated for the three methods of covari-

ance matrix estimation. In this way the computational power required has been largerly reduced..

For each basket of stocks and each allocation, the mean and variance of the returns were calculated, as well as the Sharpe Ratio. The results of the indicators obtained by different allocation models were averaged.

Three indicators are therefore associated with each basket of securities: the average mean return, to average variance returns and the average Sharpe Ratios.

With respect to these indicators some comparisons between the original 40 portfolio and its competitors were produced, as usual organized with respect to the covariance matrix estimation methods.

Historical Covariance Matrix

In this section the portfolios obtained through historical covariance matrix estimation are examined.

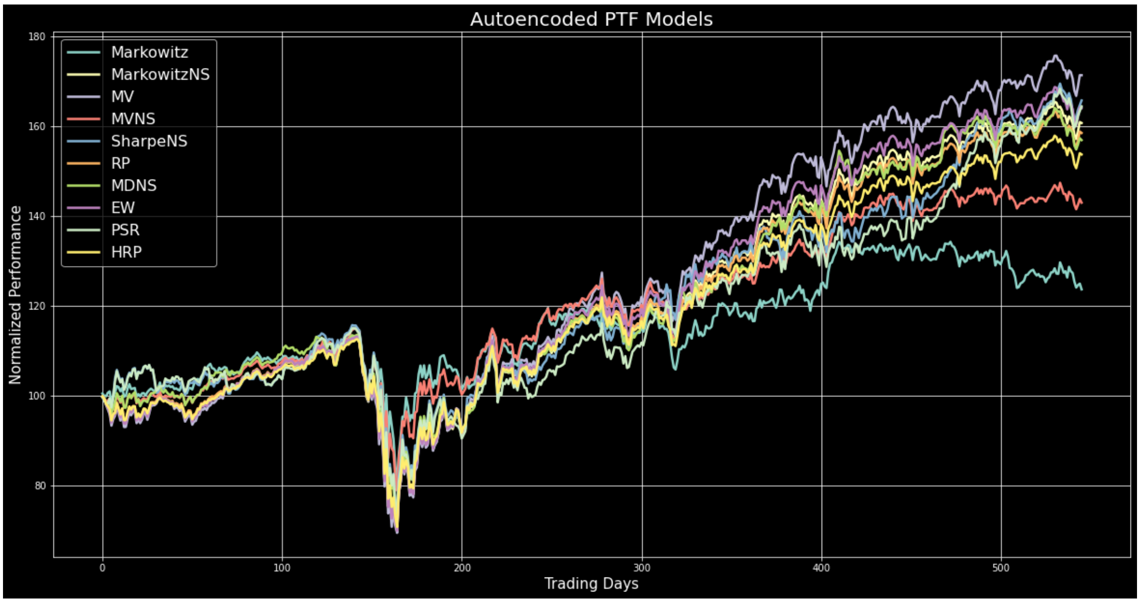

Figure 1.2: Equity lines produced by the portfolios on 40 assets.

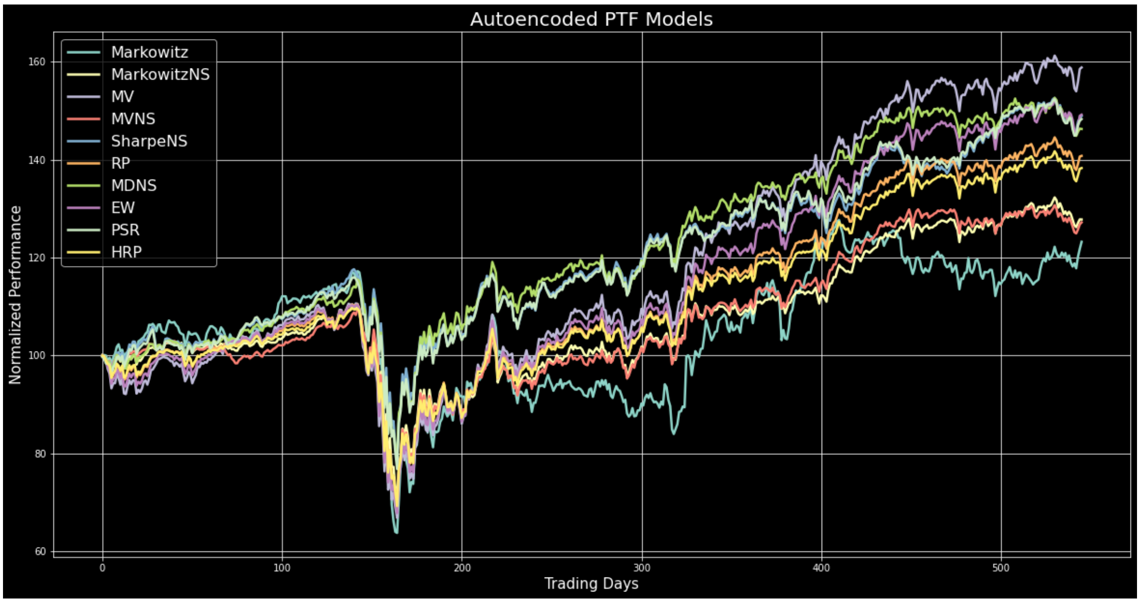

Figure 1.3: Equity lines produced by the portfolios on 335 assets.

The plots show a certain dominance of the portfolios based on 40 stocks. This is highlighted by the final equity level reached and from the mitigated reaction to the pandemic crisis. The dominance of the first group is even more evident looking at the performance indicators.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt
Markowitz	-0.30	-0.034	-0.020	-0.036	0.0381	0.015	-0.65	12.99
MarkowitzNS	-0.38	-0.039	-0.024	-0.043	0.0876	0.017	-1.19	16.05
MV	-0.38	-0.040	-0.026	-0.045	0.0996	0.018	-1.16	15.71
MVNS	-0.31	-0.030	-0.018	-0.037	0.0659	0.015	-0.89	15.72
SharpeNS	-0.35	-0.043	-0.023	-0.043	0.0931	0.017	-0.79	11.19
RP	-0.37	-0.038	-0.023	-0.043	0.0849	0.016	-1.22	16.29
MDNS	-0.37	-0.040	-0.022	-0.041	0.0832	0.016	-0.71	13.42
EW	-0.38	-0.039	-0.025	-0.044	0.0914	0.017	-1.20	16.07
PSR	-0.36	-0.046	-0.024	-0.044	0.0917	0.018	-0.75	10.55
HRP	-0.37	-0.037	-0.022	-0.042	0.0792	0.016	-1.22	16.61

Table 1.3: Historical Covariance: risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	0.401061	0.00	-0.57	0.575760	0.893176
MarkowitzNS	0.810342	1.79	0.48	1.048760	1.620289
MV	0.878966	10.03	0.75	1.148087	1.814349
MVNS	0.694543	0.01	-0.14	0.948660	1.480147
SharpeNS	0.825205	1.34	0.41	1.133741	1.834810
RP	0.803024	1.47	0.42	1.032308	1.594957
MDNS	0.795873	0.57	0.26	1.084583	1.564518
EW	0.836370	3.82	0.58	1.082632	1.687287
PSR	0.807292	0.58	0.36	1.114666	1.790694
HRP	0.757377	0.28	0.25	0.977133	1.504395

Table 1.4: Historical Covariance: performance indicators adjusted for risk for portfolios 40

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.006597	1.869464	0.735780
MarkowitzNS	1.081423	19.973684	0.930275
MV	1.083942	20.554054	0.932110
MVNS	1.027957	7.500000	0.882569
SharpeNS	1.039660	20.760000	0.954128
RP	1.106788	25.700000	0.944954
MDNS	1.082768	20.760000	0.954128
EW	1.081423	19.973684	0.930275
PSR	1.090323	15.000000	0.937615
HRP	1.106788	25.700000	0.944954

Table 1.5: Historical Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.45	-0.10	-0.025	-0.047	0.0381	0.019	-0.59	14.00
MarkowitzNS	-0.36	-0.04	-0.020	-0.040	0.0448	0.016	-1.07	17.16
MV	-0.39	-0.04	-0.025	-0.046	0.0854	0.018	-0.95	13.08
MVNS	-0.37	-0.04	-0.019	-0.041	0.0440	0.016	-0.96	17.90
SharpeNS	-0.33	-0.03	-0.021	-0.039	0.0733	0.016	-0.79	18.28
RP	-0.37	-0.04	-0.022	-0.042	0.0629	0.016	-1.09	15.81
MDNS	-0.31	-0.02	-0.016	-0.035	0.0699	0.014	-0.56	18.77
EW	-0.38	-0.04	-0.023	-0.044	0.0738	0.017	-0.98	14.87
PSR	-0.33	-0.03	-0.021	-0.039	0.0729	0.016	-0.82	18.40
HRP	-0.36	-0.04	-0.021	-0.041	0.0596	0.016	-1.04	16.30

Table 1.6: Historical Covariance: risk indicators and mean returns for portfolios 335

	SR	PSR	ISR	SoR	CR
Markowitz	0.306303	0.00	-0.37	0.432071	0.589488
MarkowitzNS	0.445047	0.00	-0.65	0.561285	0.875811
MV	0.736904	0.05	0.39	0.964208	1.523491
MVNS	0.424774	0.00	-0.56	0.544307	0.829478
SharpeNS	0.727181	0.06	0.04	0.940965	1.540999
RP	0.595316	0.00	-0.26	0.757018	1.183745
MDNS	0.761140	0.27	-0.03	0.998191	1.552187
EW	0.668725	0.00	0.07	0.852066	1.351297
PSR	0.721532	0.05	0.19	0.927101	1.522145
HRP	0.578651	0.00	-0.36	0.734309	1.142051

Table 1.7: Historical Covariance: performance indicators adjusted for risk for portfolios 335

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.007620	1.484018	0.598165
MarkowitzNS	1.051383	13.315789	0.930275
MV	1.059523	14.820000	0.908257
MVNS	1.054655	14.111111	0.933945
SharpeNS	1.051383	13.315789	0.930275
RP	1.058190	14.500000	0.906422
MDNS	1.035341	6.586667	0.906422
EW	1.062366	15.500000	0.911927
PSR	1.051383	13.315789	0.930275
HRP	1.049880	12.948718	0.928440

Table 1.8: Historical Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335

With the exception of the Markowitz portfolio, there are no major differences between portfolios 40 and portfolios 335 in terms of risk indicators, although there is a slight predominance on average of the 40-securities group, especially as regards kurtosis.
The difference is remarkable as regards the Sharpe and Probabilistic Sharpe ratio, becoming even clearer on the Calmar ratio and Sortino ratio, proving that the risks due to downward fluctuations affect the second group more. Furthermore the Information Sharpe Ratio provides that only 4 models out of 10 overperform the benchmark for portfolio 335, while this happens in 8 cases for portfolio 40, and in In 4 of these 8 models, ISR is larger or equal to 0.4, while for portfolio 335 this never happens.
Finally, with respect to the stability and recovery indicators, still a slight dominance of portfolio 40 has to be noted.
More details follow on the pairs comparison focus.

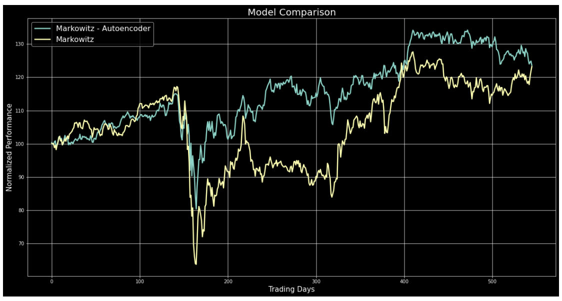

Figure 1.4: Markowitz Equity Lines.

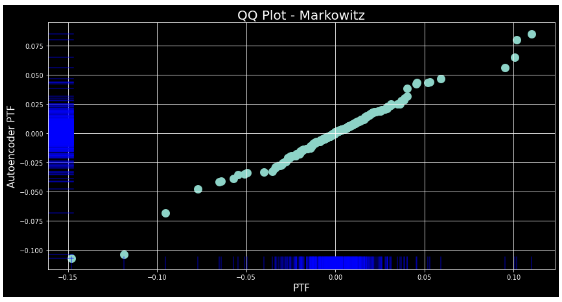

Figure 1.5: Markowitz QQ plot.

Markowitz model. The equity line shows a clear improvement provided by the stock-picking procedure. Despite both portfolios reaching the same final profit, portfolio 40 has a much less risky trend and heavily contains losses in the pandemic period than its counterpart. Then the dominance starts in the post-pandemic rally.
The QQ plot shows how the returns distributions deviate immediately after the shoulders of the distributions.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mar(A)	-0.30	-0.034	-0.020	-0.036	0.0381	0.015	-0.65	12.99	0.40
mar	-0.45	-0.102	-0.025	-0.047	0.0381	0.019	-0.59	14.00	0.30

Table 1.9: Markowitz model: performance indicators.

The performances indicators provides a very strong dominance of the 40 portfolio under every perspective except for the mean returns, especially for the extreme risk measures.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.16	0.045

Table 1.10: Markowitz model: econometric tests.

The two tests confirm that if the two portfolios are generated from the same distribution the observed returns for the pair would be unlikely.

Markowitz model, no short sales The improvement of portfolio 40 is strong even for the Markowitz model with no short sales allowed.

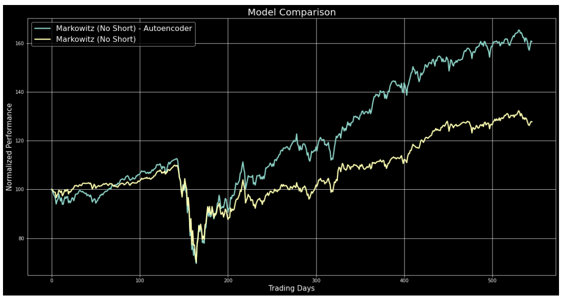

Figure 1.6: Markowitz (NS) equity lines.

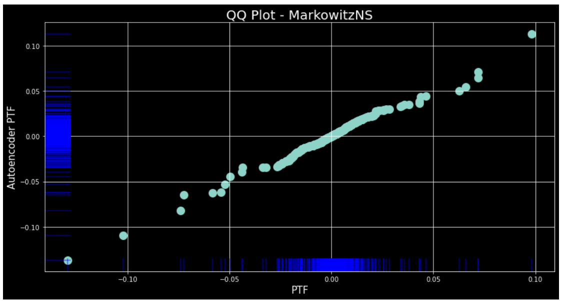

Figure 1.7: Markowitz (NS) QQ plot.

In that case, the reaction to the pandemic has been basically the same for the two portfolios, but the after-pandemic rally has been clearly stronger for portfolio 40, which performed better than the 335 for a 28%.
The QQ plot shows how the deviation among the two distributions is concentrated on the left shoulder and the right tail, while it lightens in the central part and in the left tail.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
marNS(A)	-0.38	-0.039	-0.024	-0.043	0.0876	0.017	-1.19	16.05	0.81
marNS	-0.36	-0.047	-0.020	-0.040	0.0448	0.016	-1.07	17.16	0.44

Table 1.11: Markowitz model with no short sales: performance indicators.

The risk measures are pretty balanced, but the difference becomes great in terms of Sharpe ratio, strongly driven by the very large difference in returns.

	Kolmogorov-Smirnov	Anderson-Darling
ρ-value	0.14	0.096

Table 1.12: Markowitz model with no short sales: econometric tests.

Based on both the KS and the AD tests it is unlikely that the two series of returns source from the same distribution, and then that probably the indicators performance differences are significative.

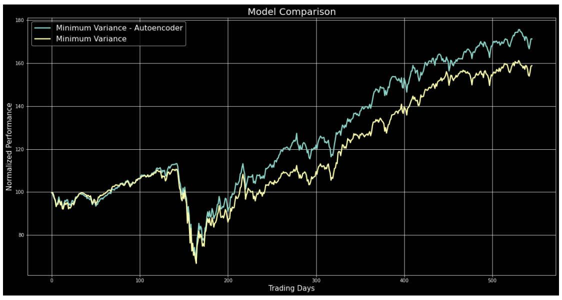

Figure 1.8: MV equity lines.

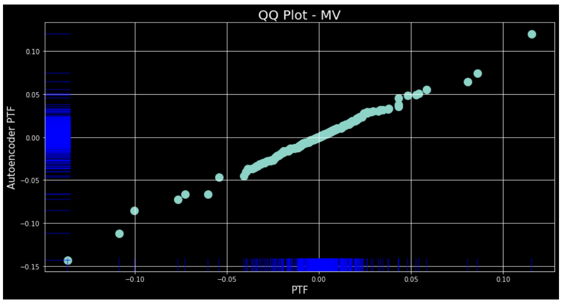

Figure 1.9: MV QQ plot.

Minimum Variance portfolio. In this case the differences are quite small. The 40 MV performs of course better in terms of final return and takes the lead when the rally starts, but the difference coudl be considered negligible. The QQ plot shows how the only meaningful deviation among the two distributions is at the end of the right shoulder.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mv(A)	-0.38	-0.040	-0.026	-0.045	0.0996	0.018	-1.16	15.71	0.88
mv	-0.39	-0.045	-0.025	-0.046	0.0854	0.018	-0.95	13.94	0.73

Table 1.13: Minimum Variance model: performance indicators.

In terms of risk measures there is on average a light improvement of the 40 portfolio against its competitor, but basically the more clear difference is given by the Sharpe ratio, driven by the mean return, showing how for the same amount of expected risk the 40 portfolio is more remunerative. The two tests confirm what already deduced through the QQ plot and the analysis of the statistical moments and the other indicators. Even if the portfolio 40 has not produced a significant improvement compared to the portfolio 335,

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.99	0.25*

Table 1.14: Minimum Variance model: econometric tests.

with a change of perspective we can also conclude that constructing the portfolio with 335 stocks does not bring any substantial benefit, and that indeed the overall risk of the out-of-sample portfolio is slightly increased while losing a certain portion of return.

Minimum Variance model, no short sales Restricting the short sales in the MV portfolio produces a stronger impact of the stock-picking procedure.

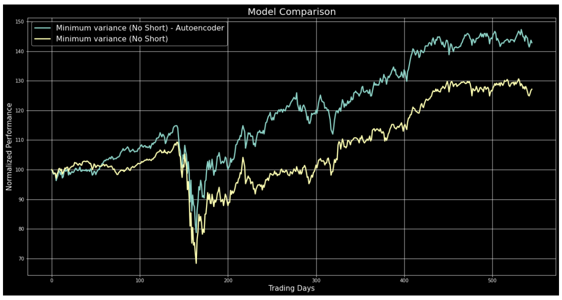

Figure 1.10: MV(NS) equity lines.

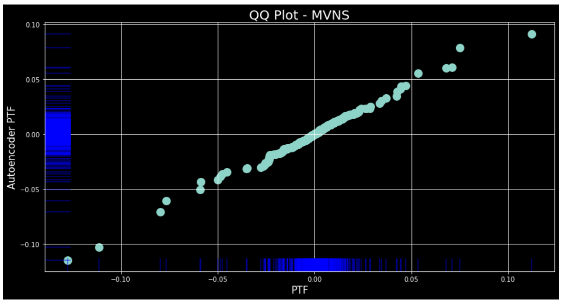

Figure 1.11: MV(NS) QQ plot.

In this case the dominance of portfolio 40 starts before the pandemic, and during the crisis the drawdown is strongly mitigated with respect to its competitor. Furthermore, the market rally after the crisis is clearly more intense, even if a little bit unstable.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mvNS(A)	-0.31	-0.030	-0.018	-0.037	0.0659	0.015	-0.89	15.72	0.69
mvNS	-0.37	-0.049	-0.019	-0.041	0.0440	0.016	-0.96	17.90	0.42

Table 1.15: Minimum Variance model with no short sales: performance indicators.

The performance measures are strongly biased toward portfolio 40 which outperform its competitor across all the indicators. In the extreme risk measures the dominance is clear, especially on the drawdowns side.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.71	0.25*

Table 1.16: Minimum Variance model with no short sales: econometric tests.

Even if the differences in terms of performance measures are quite clear, the tests struggle to reject the null hypothesis that the two series of returns come from different distributions. This apparent contraddiction is probably due to the fact that the differences among the two equity lines are produced in a few peculiar periods. However, it should be considered that the difference in improvement produced by portfolio 40 between the cases of minimum variance and minimum variance with short sales not allowed is in any case not negligible. This may be given by the fact that the constraint on shorts has subtracted degrees of freedom useful to the optimization algorithm to build the ideal portfolio, and that in this context the choice of stocks has had a more relevant impact.

Maximum Sharpe Ratio model. In the portfolio of maximum Sharpe ratio, the improvement produced by the stock-picking procedure is not very significant.

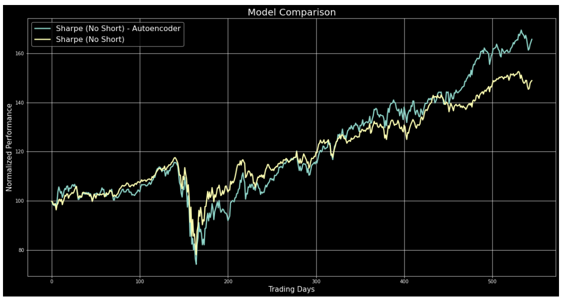

Figure 1.12: SR equity lines.

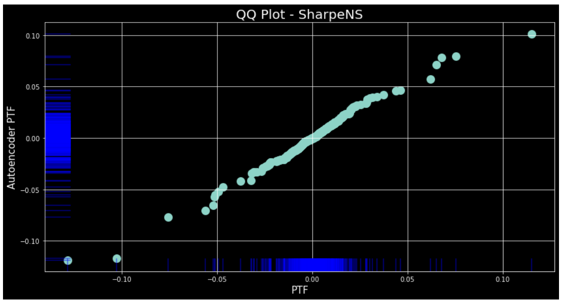

Figure 1.13: SR QQ plot.

A clear dominance of portfolio 40 starts only in the second half of the last year. Furthermore, during the pandemic it performed slightly worse. The QQ plot shows that the two portfolios returns distributions deviate from each other on the tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
srNS(A)	-0.35	-0.043	-0.023	-0.043	0.0931	0.017	-0.79	11.19	0.82
srNS	-0.33	-0.031	-0.021	-0.039	0.0733	0.016	-0.79	18.28	0.72

Table 1.17: Maximum Sharpe Ratio model: performance indicators.

In terms of performance measures, portfolio 40 performed slightly worse in terms of extreme risk measures, except for the conditional Value at risk indicator and the kurtosis. It improved the Sharpe Ratio with respect to its competitor due to an higher return that, as already said, is concentrated on the last part of the test set.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.23	0.078

Table 1.18: Maximum Sharpe Ratio model: econometric tests.

The econometric tests assess that is reasonable to think that the two portfolios returns source from different probability distributions.

Risk Parity model. The two portfolios are extremely similar during the first year of the test set, then portfolio 40 starts a period of overperformance with respect to the its competitor, and the dominance continues until the end of the test period.

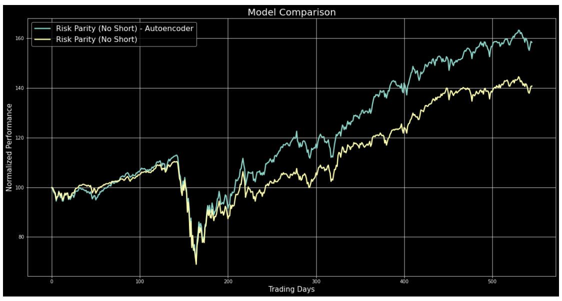

Figure 1.14: Risk parity equity lines.

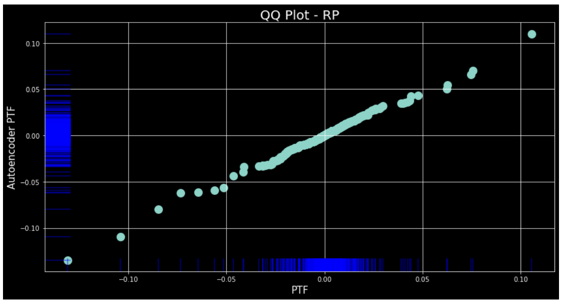

Figure 1.15: Risk parity QQ plot.

The QQ shows some difference among the two distributions on the left shoulder and the left tail.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
rp(A)	-0.37	-0.038	-0.023	-0.043	0.0849	0.016	-1.22	16.29	0.80
rp	-0.37	-0.043	-0.022	-0.042	0.0629	0.016	-1.00	15.81	0.59

Table 1.19: Risk Parity model: performance indicators.

The performance measures are quite close except for the mean return and the Sharpe Ratio, which shows an important improvement of 0.21.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.66	0.25*

Table 1.20: Risk Parity model: econometric tests.

The tests agree on the fact that it is not unlikely to observe the two return series given the null hypothesis that they are generated from the same process.

Maximum Diversification model. Even in the Maximum Diversification portfolio the dominance of portfolio 40 starts only in the last part of the considered sample.

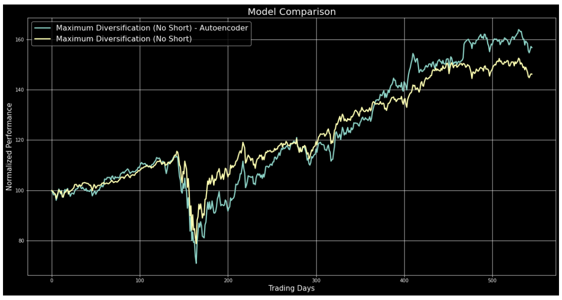

Figure 1.16: MD equity lines.

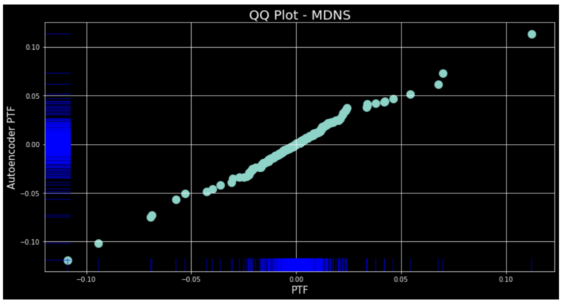

Figure 1.17: MD QQ plot.

Something interesting to notice is that before the last two years portfolio 40 is dominated by its competitor, especially during the pandemic and in the rally immediately after. In the QQ plot it is noticeable a divergence on the right should and on the right tail.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
md(A)	-0.37	-0.040	-0.022	-0.041	0.0832	0.016	-0.71	13.42	0.79
md	-0.31	-0.025	-0.016	-0.035	0.0699	0.014	-0.56	18.77	0.76

Table 1.21: Maximum Diversification model: performance indicators.

In terms of risk metrics, the larger portfolio entirely dominates its competitor. The situation is reversed with respect to returns and consequently to Sharpe Ratio, where in any case the difference is negligible. A possible interpretation is that the MD portfolio, being the model which more than others relies on diversification and which has a diversification measure as objective function, intrinsecally benefits from a larger basket.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.49	0.10

Table 1.22: Maximum Diversification model: econometric tests.

The KS and AD tests strongly disagree with respect to the null hypothesis that the two returns series come from different generating data processes, but as already said the KS test is by construction too conservative, and so this kind of behavior is expected.

Equally Weighted model. The Equally Weighted portfolio is useful in order to assess if the stock picking have been successfull, given that it is agnostic to any asset allocation model and covariance matrix estimation.

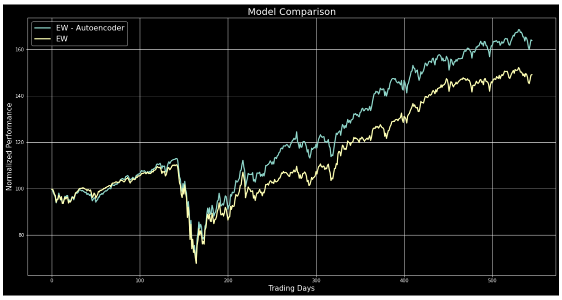

Figure 1.18: EW equity lines.

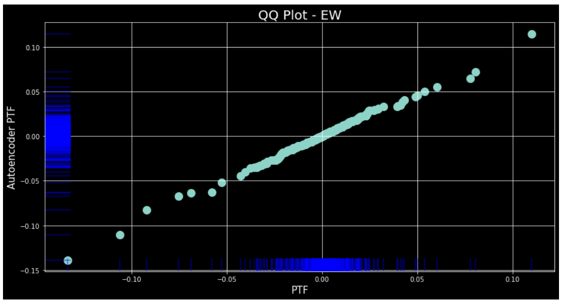

Figure 1.19: EW QQ plot.

The equity lines show how the portfolio 40 have been superior especially during the rally after the pandemic crisis. At the same time the QQ plot does not reveal significant deviations among the two returns series distribution, except for the left tail.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
ew(A)	-0.38	-0.039	-0.025	-0.044	0.0914	0.017	-1.20	16.07	0.84
ew	-0.38	-0.043	-0.023	-0.044	0.0738	0.017	-0.98	14.87	0.66

Table 1.23: Equally Weighted model: performance indicators.

In terms of risk measures, there are no evident differences. As in the other cases, the dominance comes out on the mean return and then on the Sharpe Ratio, where a not trivial 0.18 improvement has been recorded for portfolio 40. Probably the market trends have been so massive in the test set period that influenced almost all the stocks according to it. At the same time, the stock selection can reveal advantages when combined with portfolio models, and in any case the overperformance is clear even for the Equally Weighted portfolio.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.92	0.25*

Table 1.24: Equally Weighted model: econometric tests.

As already deduced, standing to the econometrics tests seem to be no evidences to refuse the idea the two portfolios' returns are produced by the same probability distribution.

Probabilistic Sharpe Ratio model. The PSR shows a behavior which is quite similar to what has been observed for maximum Sharpe Ratio portfolio, and the conclusions are mostly the same.

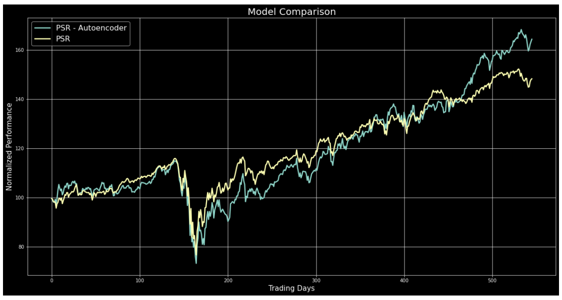

Figure 1.20: PSR equity lines.

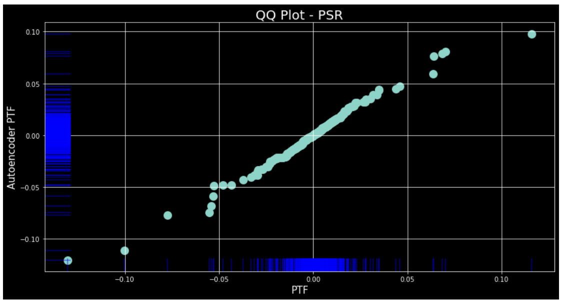

Figure 1.21: PSR QQ plot.

The QQ plot shows a L figure at the start of the left tail, while the dominance of portfolio 40 starts only in the last period. During the initial rally the larger portfolio overperforms its competitor.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
psr(A)	-0.36	-0.046	-0.024	-0.044	0.0917	0.018	-0.75	10.55	0.81
psr	-0.33	-0.030	-0.021	-0.039	0.0729	0.016	-0.82	18.40	0.72

Table 1.25: Probabilistic Sharpe Ratio model: performance indicators.

On the risk perspective portfolio 40 is outperformed, except for the kurtosis. The dominance is - as usual - with respect to the expected return and Sharpe Ratio.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.16	0.05

Table 1.26: Probabilistic Sharpe Ratio model: econometric tests.

The tests show that it is reasonable to reject the null hypothesis, with the KS as usual much more conservative in rejection.

Hierarchical Risk Parity model. The HRP portfolio performance analysis is quite interesting, because the HRP already implements a Machine Learning approach in order to achieve true diversification through clustering.

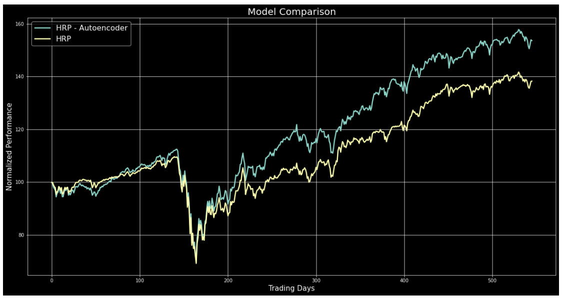

Figure 1.22: HRP equity lines.

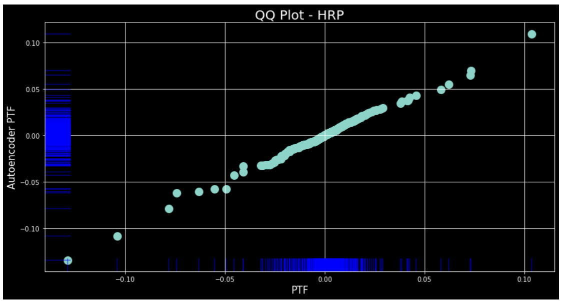

Figure 1.23: HRP QQ plot.

Even in this case the dominance of the portfolio based on stock-picking starts after the pandemic, and becomes deeper in time. The QQ plot shows a divergence starting in the left tail.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
hrp(A)	-0.37	-0.037	-0.022	-0.042	0.0792	0.016	-1.22	16.61	0.77
hrp	-0.36	-0.042	-0.021	-0.041	0.0596	0.016	-1.04	16.30	0.57

Table 1.27: Hierarchical Risk Parity model: performance indicators.

The performance measures are very close for the risk part, while the portfolio 40 overperforms its competitor as regards expected return and Sharpe Ratio, which value is increased by 0.20.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.71	0.25*

Table 1.28: Hierarchical Risk Parity model: econometric tests.

Even if the difference in terms of Sharpe ratio is not negligible, the tests struggle to distinguish two different distributions for the returns of the two investments, even if the KS test which is usually very punishing stops at 0.71.

Historical covariance: rolling portfolios analysis. As for the rolling portfolios analysis, the following results have been obtained.

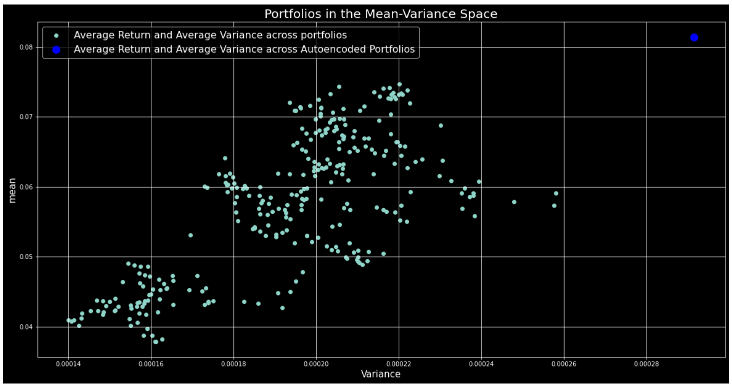

Figure 1.24: Historical Covariance, Mean-Variance Space: portfolio 40 against competitors.

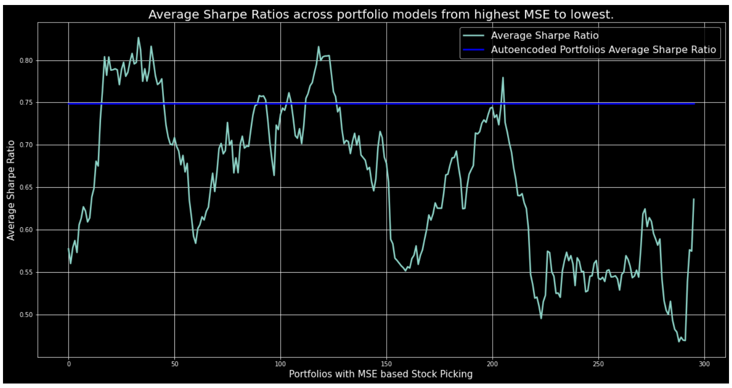

Figure 1.25: Historical Covariance, Average Sharpe ratio: portfolio 40 against competitors.

It can be seen that portfolio 40 is by far the riskiest in the space of average mean return and and variance - with an abuse of notation, this space will be called *Mean-Variance Space*. At the same time, however, it is the most profitable portfolio. In the dichotomy between the two measures, the profitability of the portfolio prevails, as can be seen from the second plot, where the average Sharpe Ratio of the portfolio 40 is compared with the their competitors corresponding. It can be seen that portfolio based on stock-picking is able to outperform 82% of its competitors in terms of risk adjusted return. It is also interesting to note how the trend of the competitors' Sharpe Ratio seems to confirm the theory underlying the model: a slightly downward trend can in fact be noticed, even if the series is rather noisy.

Exponentially Weighted Covariance Matrix

In this section the performances obtained by portfolio 40 and portfolio 335 are explored across the allocation models with covariance matrix calculated according to the Eponentially Weighted method.

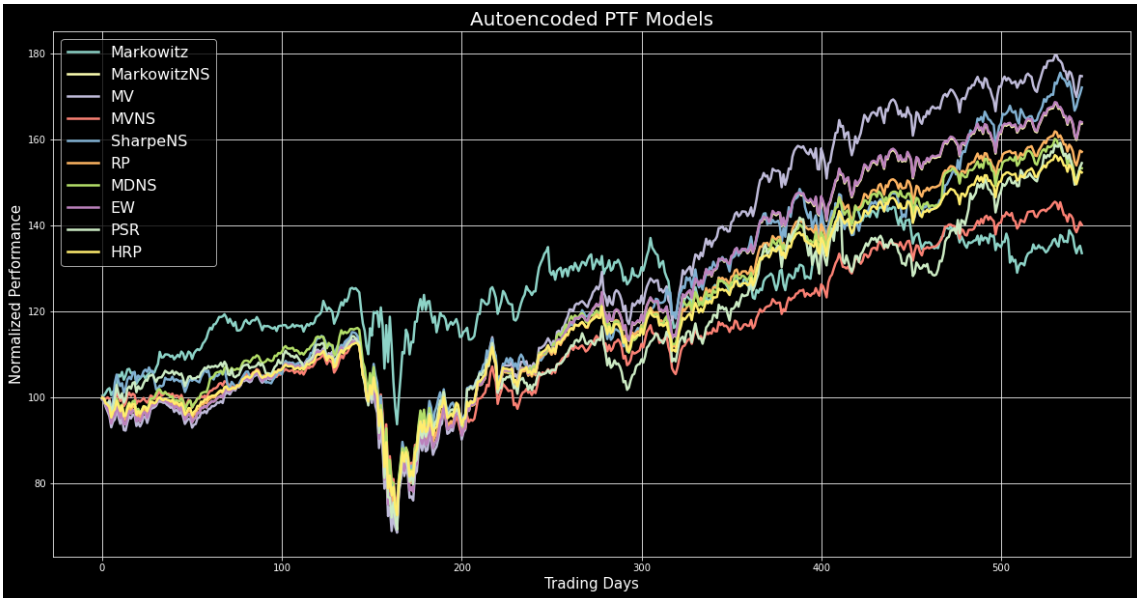

Figure 1.26: Equity lines produced by the portfolios on 40 assets.

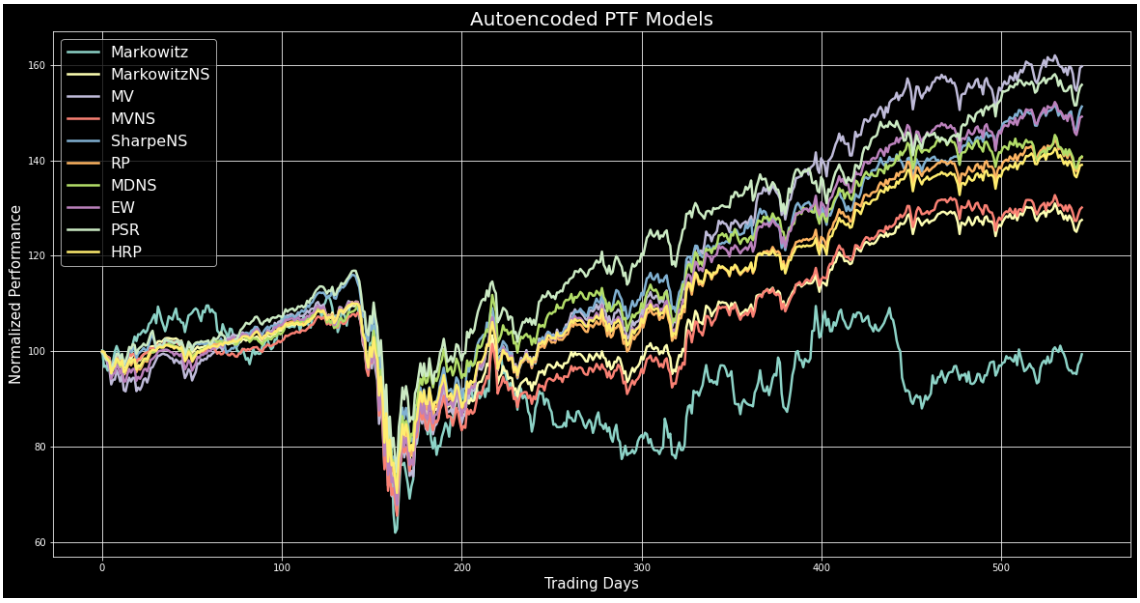

Figure 1.27: Equity lines produced by the portfolios on 335 assets.

The plots show how the models obtained on the 40 stocks dominated on average the performance in terms of yield as well as of risk mitigation during the pandemic. The resilience of the portfolio 40 Markowitz model is surprising, even if the rest of the path is quite disappointing compared to the other models. Despite this, the dominance over its counterpart is stark.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt
Markowitz	-0.25	-0.040	-0.023	-0.038	0.0524	0.015	-0.09	8.75
MarkowitzNS	-0.38	-0.039	-0.025	-0.044	0.0910	0.017	-1.20	16.08
MV	-0.39	-0.042	-0.027	-0.047	0.1034	0.018	-1.15	15.80
MVNS	-0.34	-0.039	-0.021	-0.040	0.0619	0.015	-1.16	15.40
SharpeNS	-0.37	-0.041	-0.023	-0.043	0.0999	0.018	-0.73	11.50
RP	-0.37	-0.037	-0.023	-0.042	0.0833	0.016	-1.23	16.08
MDNS	-0.37	-0.040	-0.024	-0.042	0.0790	0.017	-0.86	15.38
EW	-0.38	-0.039	-0.025	-0.044	0.0914	0.017	-1.20	16.07
PSR	-0.39	-0.051	-0.026	-0.046	0.0801	0.018	-0.71	12.25
HRP	-0.35	-0.035	-0.021	-0.040	0.0774	0.015	-1.30	16.21

Table 1.29: Exponentially Weighted Covariance: risk indicators and mean returns for portfolios 40.

	SR	PSR	ISR	SoR	CR
Markowitz	0.526354	0.00	-0.26	0.798897	1.462213
MarkowitzNS	0.833923	3.59	0.46	1.081416	1.681572
MV	0.885665	11.48	0.76	1.165472	1.834455
MVNS	0.618936	0.00	-0.27	0.804639	1.259394
SharpeNS	0.875299	6.41	0.51	1.221511	1.892302
RP	0.797905	1.22	0.39	1.027611	1.587008
MDNS	0.734157	0.05	0.19	0.991731	1.485826
EW	0.836370	3.82	0.58	1.084786	1.687287
PSR	0.673827	0.00	0.06	0.932760	1.434857
HRP	0.773416	0.54	0.20	0.995496	1.536042

Table 1.30: Exponentially Weighted Covariance: performance indicators adjusted for risk for portfolios 40.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.017677	2.400000	0.748624
MarkowitzNS	1.081423	19.973684	0.930275
MV	1.093400	15.151515	0.919266
MVNS	1.057369	9.880000	0.908257
SharpeNS	1.039660	20.760000	0.954128
RP	1.102827	24.822581	0.943119
MDNS	1.052973	13.702703	0.932110
EW	1.081423	19.973684	0.930275
PSR	1.049880	12.948718	0.928440
HRP	1.074836	36.857143	0.948624

Table 1.31: Exponentially Weighted Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.44	-0.137	-0.032	-0.054259	-0.0016	0.023	-0.675936	9.41
MarkowitzNS	-0.38	-0.055	-0.020	-0.041655	0.0446	0.016	-1.06	15.81
MV	-0.40	-0.047	-0.025	-0.047105	0.0865	0.018	-0.91	13.58
MVNS	-0.39	-0.059	-0.021	-0.043084	0.0483	0.017	-1.25	16.23
SharpeNS	-0.38	-0.048	-0.023	-0.043953	0.0765	0.017	-1.27	18.70
RP	-0.37	-0.043	-0.021	-0.042021	0.0627	0.016	-1.02	15.92
MDNS	-0.34	-0.034	-0.019	-0.039119	0.0631	0.015	-0.85	13.29
EW	-0.38	-0.043	-0.023	-0.044088	0.0738	0.017	-0.98	14.87
PSR	-0.36	-0.036	-0.023	-0.041768	0.0821	0.016	-0.99	17.72
HRP	-0.36	-0.040	-0.020	-0.040571	0.0607	0.016	-1.01	15.74

Table 1.32: Exponentially Weighted Covariance: risk indicators and mean returns for portfolios 335

	SR	PSR	ISR	SoR	CR
Markowitz	-0.010631	0.00	-0.57	-0.015980	-0.024851
MarkowitzNS	0.424185	0.00	-0.63	0.533035	0.825828
MV	0.731259	0.03	0.39	0.956902	1.515160
MVNS	0.445363	0.00	-0.49	0.564927	0.866024
SharpeNS	0.706894	0.04	0.12	0.883197	1.400174
RP	0.595960	0.00	-0.27	0.747940	1.185207
MDNS	0.633871	0.00	-0.22	0.834802	1.279643
EW	0.668725	0.00	0.07	0.845635	1.351297
PSR	0.788575	0.93	0.25	0.997977	1.605234
HRP	0.600757	0.00	-0.35	0.752216	1.185406

Table 1.33: Exponentially Weighted Covariance: performance indicators adjusted for risk for portfolios 335.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.011970	2.073446	0.675229
MarkowitzNS	1.054655	14.111111	0.933945
MV	1.054502	13.611111	0.900917
MVNS	1.054655	14.111111	0.933945
SharpeNS	1.049880	12.948718	0.928440
RP	1.062366	15.500000	0.911927
MDNS	1.054655	14.111111	0.933945
EW	1.062366	15.500000	0.911927
PSR	1.051383	13.315789	0.930275
HRP	1.049880	12.948718	0.928440

Table 1.34: Exponentially Weighted Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335.

The performance measures confirm what could be inferred from the equity lines. Beyond the Markowitz model, the overperformance of portfolio 40 across all the models and various indicators can be noted.

Where short sales are not allowed on the Markowitz model, the stock-picking based portfolio shows a more than double expected return with respect to its counterpart, and the same holds true with respect to the Sharpe Ratio. The difference is even amplified when the Sortino Ratio is taken in account. With respect to the extreme risk measures, the two portfolios are balanced, except for the not negligible difference in terms of average drawdown.

The same happens for all the other asset allocation models, with the only distinction being the Probabilistic Sharpe Ratio, which also obtains the maximum result in terms of the homonymous performance indicator. In general, the PSR reacts positively with a vast universe of assets, where differences in terms of asymmetry and kurtosis can be better exploited.

However, with regard to the PSR performance measure, it is quite significant the outcome of such an accurate indicator in the comparison between the two portfolios blocks: while in the 335 portfolio, with the only notable exception of the aforementioned model, the probability values are all firmly anchored to zero, in portfolio 40 there are several models that have non-zero probability of a true Sharpe Ratio greater than 1, and all of them have a larger probability with respect to the best model in terms of PSR of the larger group. As for ISR, results are basically unchanged with respect to the historical covariance matrix estimation method.

As regards the bull persistence and bear resilience indicators, the dominance of portfolio 40 stocks concerns almost all the asset allocation models.

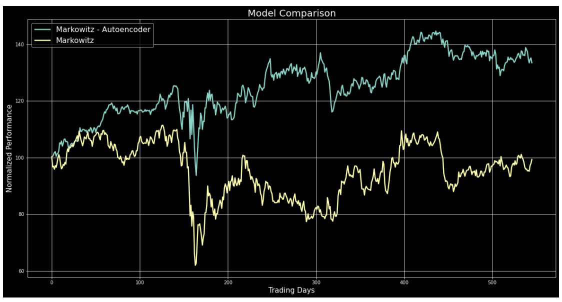

Figure 1.28: Markowitz equity lines.

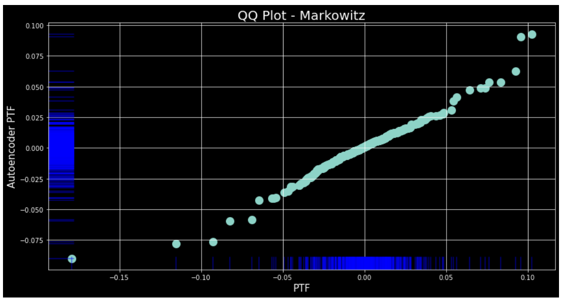

Figure 1.29: Markowitz QQ plot.

Markowitz model. For the Markowitz model portfolio 335, after a slow start, is hardly hit during the pandemic period - the worst portfolio of all those analyzed in this paper from this perspective. On the contrary, portfolio 40 has a bombastic start, thanks to which it is able to strongly limit the damage due to the pandemic crisis, even if the continuation of the performance is rather fluctuating and betrays the expectations of the first period. The QQ plot shows how the divergence between the distributions of the returns of the two portfolios begins already in the final part of the shoulders, and then becomes clear and unambiguous in the tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mar(A)	-0.25	-0.043	-0.023	-0.038	0.0524	0.015	-0.09	8.75	0.52
mar	-0.44	-0.137	-0.032	-0.054	-0.0016	0.023	-0.68	9.41	-0.01

Table 1.35: Markowitz model: performance indicators.

The difference in performance is clear across almost all indicators. On the risk side - extreme and not - the dominance of the portfolio 40 is clear, as can be seen especially on the drawdown side.
The difference remains equally evident on the expected return, and is consequently reflected together with the lower standard deviation on the Sharpe Ratio.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.05	0.01*

Table 1.36: Markowitz model: econometric tests.

Econometric tests confirm what emerged from the visual impression: it is quite reasonable to infer that the two series of returns come from very distinct distributions.

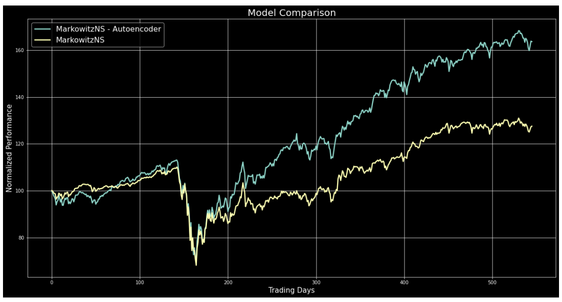

Figure 1.30: Markowitz(NS) equity lines.

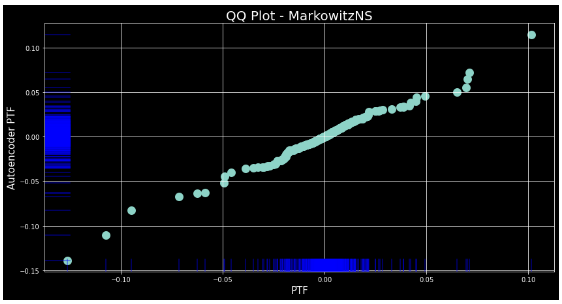

Figure 1.31: Markowitz(NS) QQ plot.

Markowitz model, no short sales. When short sales are not allowed in the Markowitz model, it can be seen how the two portfolios behave in a very similar way in the first period, during the pandemic and for a very short period after the crisis. During the rest of the

rally, however, the two equity lines diverge more and more markedly, with a net dominance of portfolio 40.
The QQ plot shows how apart from the very central percentiles the two distributions deviate strongly from each other in both directions, with a strong domain of portfolio 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
marNS(A)	-0.38	-0.039	-0.025	-0.044	0.0910	0.017	-1.20	16.08	0.83
marNS	-0.38	-0.055	-0.020	-0.041	0.0446	0.016	-1.06	15.81	0.42

Table 1.37: Markowitz model with no short sales: performance indicators.

Based on the performance indicators, the two portfolios show quite similar results as regards risk indicators, where portfolio 40 performs slightly worse albeit with the notable exception of the mean drawdown. The difference is instead large as regards the average return and consequently the Sharpe Ratio, with respect to which portfolio 40 achieves a double performance compared to its competitor.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.10	0.19

Table 1.38: Markowitz model with no short sales: econometric tests.

Econometric tests confirm that it is less likely to have observed similar returns than the opposite, assuming that the series of returns of the two portfolios come from the same data generating process. Curiously, this is one of the very few cases where the rho -value obtained from the KS test is less than the AD.

Minimum Variance model. Turning to the Minimum Variance portfolio, it can be observed that also in this case the two investments behave similarly after all.

Figure 1.32: MV equity lines.

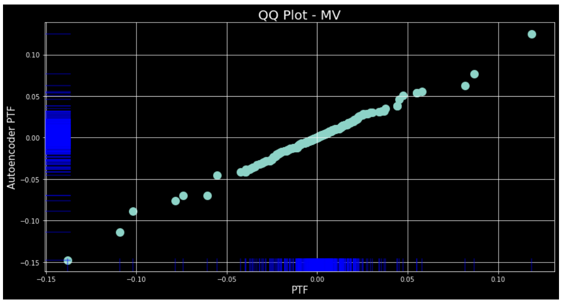

Figure 1.33: MV QQ plot.

After a very similar initial trend, portfolio 40 begins to detach from 335, with steeper climbs and more contained descents. Towards the final part - after the 400-th trading day - the autoencoder based portfolio exploits a rather intense rise in the same period during which its competitor presents a drawdown.
However, the QQ plot shows how the distributions of the two portfolios diverge slightly in terms of returns.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mv(A)	-0.39	-0.042	-0.027	-0.047	0.1034	0.018	-1.15	15.80	0.88
mv	-0.40	-0.047	-0.025	-0.047	0.0865	0.018	-0.91	13.58	0.73

Table 1.39: Minimum Variance model: performance indicators.

The risk indicators stand on similar values more or less for both portfolios. The difference as already is in the expected returns, which lead to a non-trivial increase of 0.15 on the Sharpe Ratio.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.97	0.25*

Table 1.40: Minimum Variance model: econometric tests.

Econometric tests confirm the visual impression: in probabilistic terms it is difficult to distinguish the returns of one portfolio from the other.

Minimum variance model, no short sales. The differences between the two portfolios become a little more substantial when short selling is inhibited.

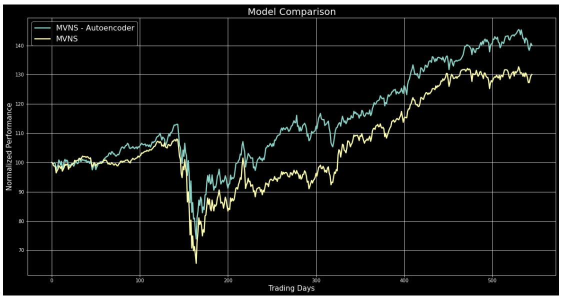

Figure 1.34: MV(NS) equity lines.

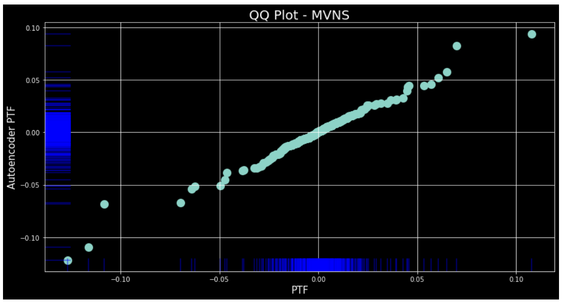

Figure 1.35: MV(NS) QQ plot.

In this case, the dominance of portfolio 40 begins very soon, and persists during the pandemic thanks to the advantage accumulated. The difference between the two equity lines shows periods of expansion as well as contraction, but the preferability of the portfolio 40 is not in any case called into question. In general, this superiority is built above all those periods for which the 335 portfolio is in a lateral trend.
The QQ plot shows how a really noticeable divergence starts at the beginning of the distribution tails, with an advantageous structure for the portfolio 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mvNS(A)	-0.34	-0.039	-0.021	-0.040	0.0619	0.015	-1.16	15.40	0.61
mvNS	-0.39	-0.059	-0.021	-0.043084	0.0483	0.017	-1.25	16.23	0.44

Table 1.41: Minimum Variance model with no short sales: performance indicators.

Risk indicators, with the sole exception of Value at Risk, clearly reward portfolio 40, especially as regards drawdowns. As almost always happened, however, it is once again the expected return that clearly discriminates the two investments, reverberating once again on the Sharpe Ratio, where there is a gain of 0.16 return units for each unit of risk assumed.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.24	0.25*

Table 1.42: Minimum Variance model with no short sales: econometric tests.

The econometric tests reveal in this case an ambiguous result, but which nevertheless leads to consider at least likely that the two series of returns come from different probability distributions.

Maximum Sharpe Ratio model. In the maximum Sharpe Ratio portfolio again there is a dominance of portfolio 40, which becomes very substantial in the last year of test set data.

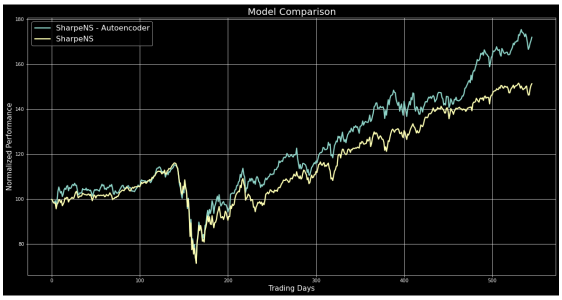

Figure 1.36: SR equity lines.

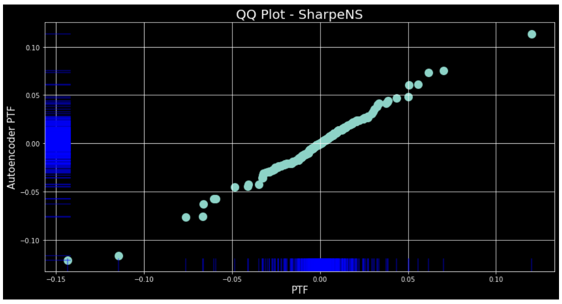

Figure 1.37: SR QQ plot.

After a better start for portfolio 40, the two equity lines end up collimating until the pandemic crisis. In the immediately following bullish period, the difference in performance immediately begins to be glimpsed, becoming more and more clear. In the final part, portfolio 40 makes a leap as for performance that clearly distances its competitor.
The QQ plot then shows how dissimilarity between the two series of returns begins already in the final part of the distributions shoulders and becomes glimpsed in some areas of the tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
srNS(A)	-0.37	-0.041	-0.023	-0.043	0.0999	0.018	-0.73	11.50	0.87
srNS	-0.38	-0.048	-0.023	-0.044	0.0765	0.017	-1.27	18.70	0.70

Table 1.43: Maximum Sharpe Ratio model: performance indicators.

Compared on the basis of the risk indicators, portfolio 40 performs better than 335. Once again portfolio 40 is able to do better especially on the expected returns and consequently the Sharpe Ratio, with an increase of 0.18 units of expected return adjusted for unit of risk assumed.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.16	0.16

Table 1.44: Maximum Sharpe Ratio model: econometric tests.

The KS and AD tests agree in the ρ -value, which would allow to reject with a confidence level of 84 % the hypothesis that the two series of returns were generated by the same process.

Risk Parity model. For the Risk Parity model, results are quite similar to those obtained with historical covariance estimation.

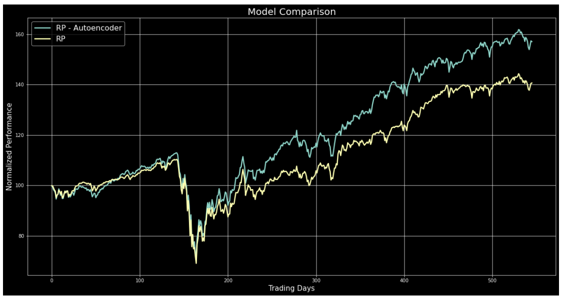

Figure 1.38: RP equity lines.

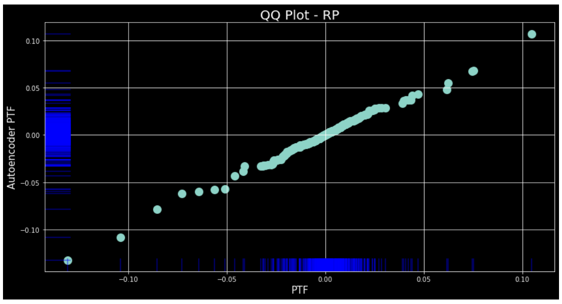

Figure 1.39: RP QQ plot.

Also in this case, the dominance of portfolio 40 takes shape towards the start of the post-pandemic rally, while in the previous period the two portfolios were essentially the same. The QQ plot shows a certain similarity between the distributions of the returns of the two investment strategies, only partially broken at the beginning of the left tail.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
rp(A)	-0.37	-0.037	-0.023	-0.042	0.0833	0.016	-1.23	16.08	0.79
rp	-0.37	-0.043	-0.021	-0.042	0.0627	0.016	-1.02	15.92	0.59

Table 1.45: Risk Parity model: performance indicators.

Even in this case, the risk indicators of the two portfolios are essentially the same, with the exception of the average drawdown where portfolio 40 performs better. Also in this case, the average yield discriminates the smaller basket portfolio from its competitor. This contribution is reflected in the Sharpe Ratio, where portfolio 40 outperforms 335 by 0.20 units.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.77	0.25*

Table 1.46: Risk Parity model: econometric tests.

Econometric tests do not substantially allow us to reject the null hypothesis, according to which the two series of returns come from the same probability distribution.

Maximum Diversification model. The Maximum Diversification model is also in this case the one where differences among two portfolios are blurred.

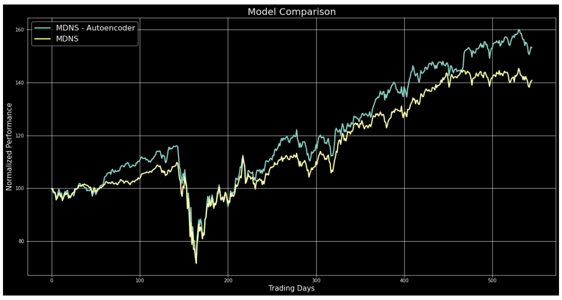

Figure 1.40: MD equity lines.

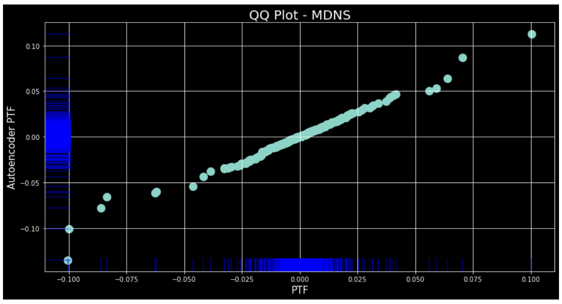

Figure 1.41: MD QQ plot.

After a better start, the dominance of the portfolio 40 actually materializes only after the post pandemic rally already started. During the critical period it shows a greater resilience, and in the last part is able to still grow when its competitor starts a lateral trend. The QQ plot shows a certain equivalence in the distributions around the central area which then ends up breaking up closer to tails, especially the left one.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
md(A)	-0.37	-0.040	-0.024	-0.042	0.0790	0.017	-0.86	15.38	0.73
md	-0.34	-0.034	-0.019	-0.039	0.0631	0.015	-0.85	13.29	0.63

Table 1.47: Maximum Diversification model: performance indicators.

In terms of risk indicators, the overperformance of the competitor portfolio is unquestionable. Again, portfolio 40 does better as for the expected return, and this is sufficient to obtain a slightly higher Sharpe Ratio.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.14	0.25*

Table 1.48: Maximum Diversification model: econometric test.

Finally, the econometric tests would seem to suggest that the returns of the two investments are generated by different probability distributions, even if the verdict is not entirely clear, given the discrepancy between the two tests, moreover in the opposite direction to the usual one, i.e. the KS ρ-value is lower than the AD one.

Probabilistic Sharpe Ratio model. The Probabilistic Sharpe Ratio model essentially replicates what happened for the Maximum Diversification portfolio.

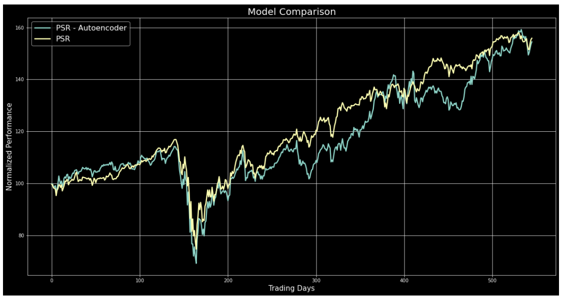

Figure 1.42: PSR equity lines.

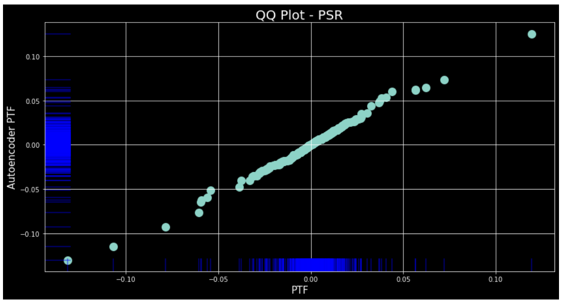

Figure 1.43: PSR QQ plot.

The auto-encoder strategy is totally dominated by the larger portfolio, as the QQ plot also demonstrates. This is true especially in terms of risk, both in the pandemic period and in some phases of the subsequent uptrend.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
psr(A)	-0.39	-0.051	-0.026	-0.046	0.0801	0.018	-0.71	12.25	0.71
psr	-0.36	-0.036	-0.023	-0.0417	0.0821	0.016	-0.99	17.72	0.78

Table 1.49: Probabilistic Sharpe Ratio model: performance indicators.

The performance measures confirm the clear dominance of portfolio 335 with regard to risk and expected return, with a resulting Sharpe Ratio improved by 0.07 with respect to portfolio 40.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.09	0.06

Table 1.50: Probabilistic Sharpe Ratio model: econometric tests.

Econometric tests suggest that it would not be far-fetched to reject the null hypothesis, according to which the returns of the two investments would come from the same probability distribution.

Hierarchical Risk Parity model. Finally, the Hierarchical Risk Parity model maintains a behavior quite similar to its historical covariance-based counterpart.

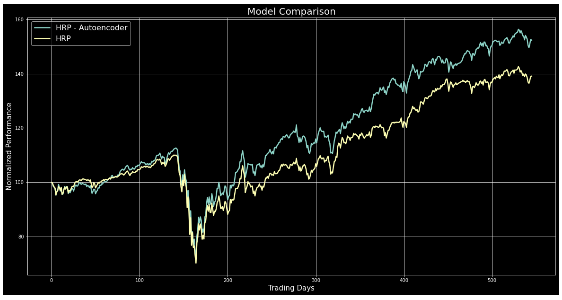

Figure 1.44: HRP equity lines.

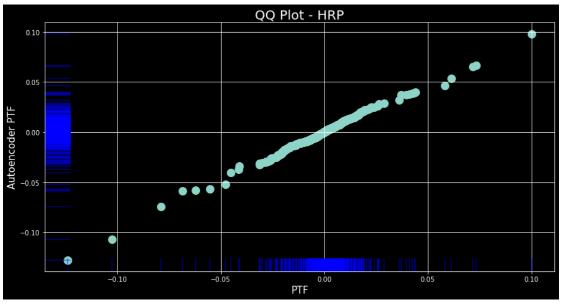

Figure 1.45: HRP QQ plot.

The dominance of portfolio 40 is not negligible. It is triggered immediately after the pandemic and then gradually underlines in an increasingly evident manner. This overperformance is sometimes cumulated on those days on which portfolio 335 shows a lateral trend while portfolio 40 takes advantage. The QQ plot shows how the best performance of portfolio 40 was built also by partially limiting the collapses of the left tail compared to the competing model.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
hrp(A)	-0.35	-0.035	-0.021	-0.040	0.0774	0.015	-1.30	16.21	0.77
hrp	-0.36	-0.040	-0.020	-0.040	0.0607	0.016	-1.01	15.74	0.60

Table 1.51: Hierarchical Risk Parity model: performance indicators.

With respect to risk indicators, the ability of the autoencoder-based portfolio to contain average drawdowns is highlighted. The difference in terms of average return, combined with a slight overperformance in terms of standard deviation, allowed a substantial increase in the Sharpe Ratio of 0.16.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.77	0.25*

Table 1.52: Hierarchical Risk Parity model: econometric tests.

The KS and AD tests agree in the impossibility of rejecting the null hypothesis, even if in a not totally decisive way.

Exponentially weighted covariance: rolling portfolios analysis. Here the results of the test carried out by Exponentially Weighting covariance estimation are summarized.

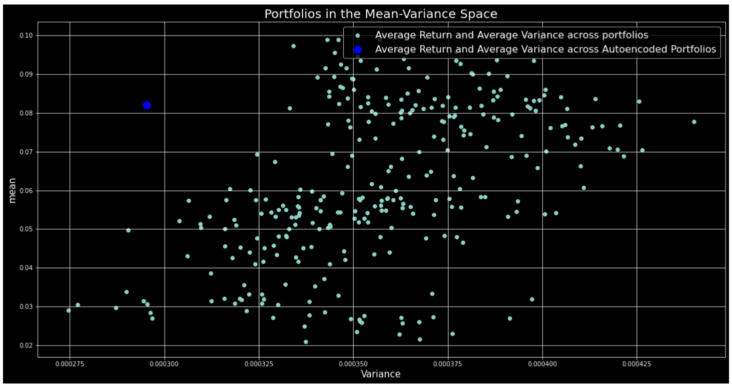

Figure 1.46: Exponentially Weighted Covariance, Mean-Variance Space: portfolio 40 against competitors.

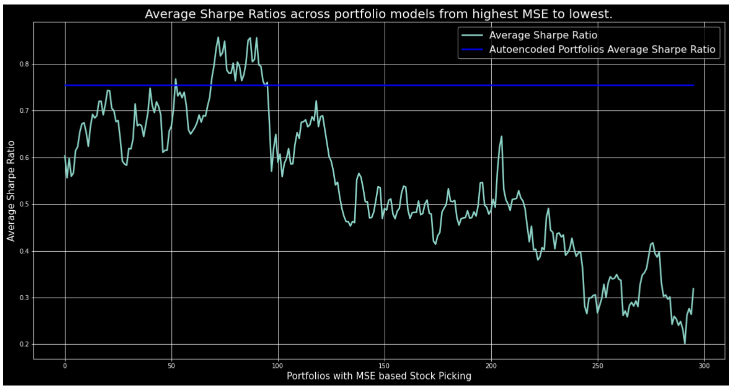

Figure 1.47: Exponentially Weighted Covariance, Average Sharpe Ratio: portfolio 40 against competitors.

In this case, the profitability and risk of the portfolio is much more balanced than in the historical estimation case. It can be seen from the plot in the Mean-Variance Space how portfolio 40 performs better than the vast majority of competitors in terms of profitability, and the same happens to an even greater extent if the focus is shifted to variance.
The average Sharpe Ratio confirms the visual intuition: portfolio 40 outperforms 94 % of its competitors in terms of return adjusted to standard risk.
Also in this case, the downward trend of the average Sharpe Ratio seems to corroborate the theoretical intuition behind the proposed model.

RMT Covariance Matrix

This section collects the performances obtained by the two portfolios, with covariance matrix filtered according to Random Matrix Theory.

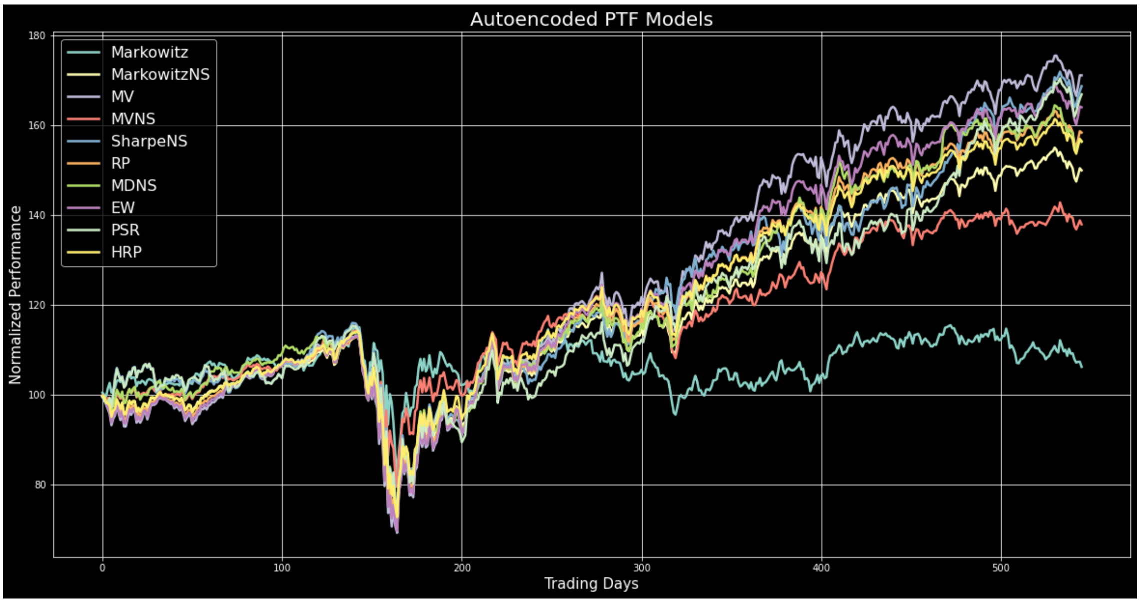

Figure 1.48: Equity lines produced by the portfolios on 40 assets.

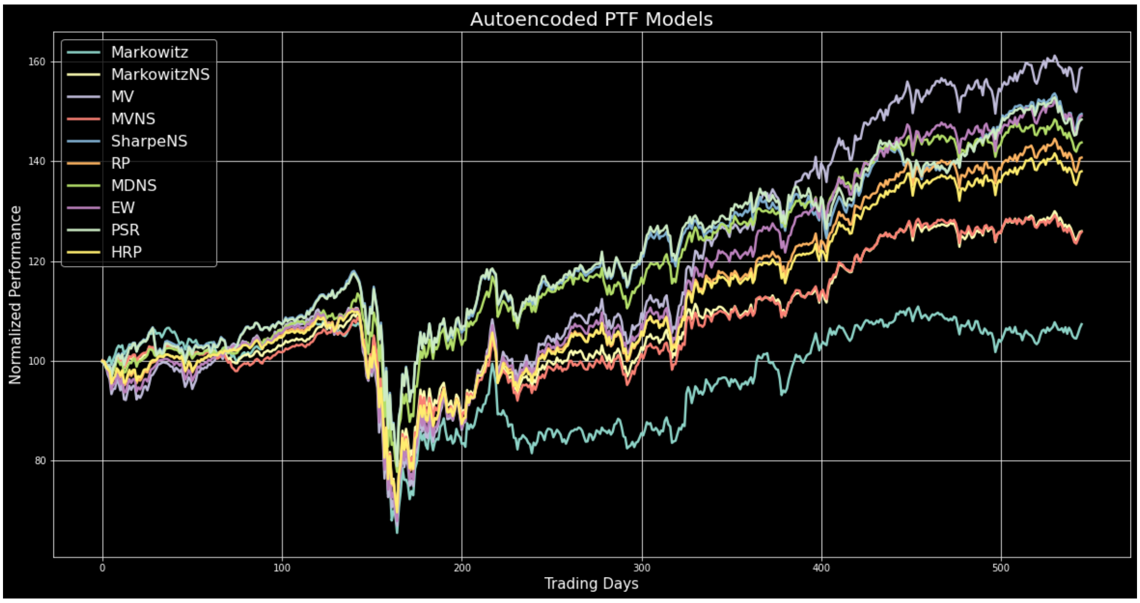

Figure 1.49: Equity lines produced by the portfolios on 335 assets.

The plots show on average a certain overperformance of the models based on the autoencoder, which becomes clear both in the ability to slightly better resist the pandemic crisis and in the total return achieved during the test period. The performance indicators confirm the visual information.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt
Markowitz	-0.28	-0.053	-0.021	-0.035	0.0101	0.014	-0.69	10.95
MarkowitzNS	-0.36	-0.036	-0.022	-0.041	0.0745	0.016	-1.30	16.32
MV	-0.38	-0.040	-0.026	-0.045	0.0993	0.017	-1.16	15.72
MVNS	-0.30	-0.032	-0.017	-0.036	0.0592	0.014	-0.91	15.40
SharpeNS	-0.36	-0.043	-0.023	-0.044	0.0966	0.018	-0.83	11.04
RP	-0.37	-0.037	-0.023	-0.043	0.0848	0.016	-1.22	16.32
MDNS	-0.37	-0.039	-0.021	-0.041	0.0825	0.016	-0.75	13.74
EW	-0.38	-0.039	-0.025	-0.044	0.0914	0.017	-1.20	16.07
PSR	-0.36	-0.047	-0.023	-0.045	0.0947	0.018	-0.76	10.32
HRP	-0.36	-0.034	-0.022	-0.041	0.0825	0.016	-1.25	16.70

Table 1.53: RMT based Covariance: risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	0.109097	0.00	-0.57	0.156249	0.249848
MarkowitzNS	0.738809	0.12	0.48	0.947172	1.436648
MV	0.877123	9.66	0.77	1.138962	1.807828
MVNS	0.629733	0.00	-0.14	0.860329	1.372569
SharpeNS	0.849583	3.08	0.42	1.149506	1.890398
RP	0.803340	1.49	0.42	1.027373	1.596565
MDNS	0.796942	0.64	0.26	1.076381	1.565908
EW	0.836370	3.82	0.58	1.076245	1.687287
PSR	0.827163	1.25	0.37	1.129469	1.843023
HRP	0.809108	1.88	0.26	1.036695	1.611743

Table 1.54: RMT based Covariance: performance indicators adjusted for risk for portfolios 40.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.008345	1.784161	0.702752
MarkowitzNS	1.106788	25.700000	0.944954
MV	1.083942	20.554054	0.932110
MVNS	1.028493	7.634921	0.884404
SharpeNS	1.064798	16.548387	0.943119
RP	1.106788	25.700000	0.944954
MDNS	1.050180	6.805556	0.899083
EW	1.081423	19.973684	0.930275
PSR	1.093542	15.484848	0.939450
HRP	1.077911	38.296296	0.950459

Table 1.55: RMT based Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt
Markowitz	-0.39	-0.095	-0.021	-0.041825	0.0131	0.016	-1.28	13.47
MarkowitzNS	-0.36	-0.045	-0.020	-0.040498	0.0423	0.016	-1.01	19.03
MV	-0.40	-0.045	-0.025	-0.046322	0.0855	0.018	-0.95	13.94
MVNS	-0.37	-0.047	-0.020	-0.040881	0.0421	0.016	-0.95	18.67
SharpeNS	-0.33	-0.030	-0.021	-0.039799	0.0742	0.016	-0.77	18.31
RP	-0.38	-0.044	-0.022	-0.042154	0.0629	0.017	-1.00	15.82
MDNS	-0.32	-0.025	-0.017	-0.035159	0.0670	0.014	-0.60	18.98
EW	-0.39	-0.044	-0.023	-0.044088	0.0738	0.018	-0.99	14.87
PSR	-0.33	-0.030	-0.022	-0.039695	0.0729	0.016	-0.67	17.92
HRP	-0.37	-0.043	-0.022	-0.041324	0.0593	0.016	-1.03	16.07

Table 1.56: RMT based Covariance: risk indicators and mean returns for portfolios 335.

	SR	PSR	ISR	SoR	CR
Markowitz	0.127347	-0.90	-0.37	0.168047	0.234936
MarkowitzNS	0.416420	-0.65	-0.65	0.524960	0.818854
MV	0.737047	0.39	0.39	0.955335	1.523809
MVNS	0.412075	-0.63	-0.59	0.523595	0.809748
SharpeNS	0.736257	0.07	0.04	0.942732	1.576640
RP	0.595422	-0.25	-0.26	0.750927	1.183914
MDNS	0.734151	-0.10	-0.04	0.951963	1.496701
EW	0.668725	0.07	0.08	0.844050	1.351297
PSR	0.726195	0.04	0.19	0.936779	1.562913
HRP	0.575051	-0.37	-0.36	0.723566	1.137848

Table 1.57: RMT based Covariance: performance indicators adjusted for risk for portfolios 335.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.028493	7.634921	0.884404
MarkowitzNS	1.054655	14.111111	0.933945
MV	1.059523	14.820000	0.908257
MVNS	1.054655	14.111111	0.933945
SharpeNS	1.051383	13.315789	0.930275
RP	1.058190	14.500000	0.906422
MDNS	1.054655	14.111111	0.933945
EW	1.062366	15.500000	0.911927
PSR	1.051383	13.315789	0.930275
HRP	1.049880	12.948718	0.928440

Table 1.58: RMT based Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335.

As for the maximum drawdown, there is an improvement of portfolio 40 compared to 335 as regards the Markowitz model and the two Minimum Variance portfolios, while for the other allocation methods of the portfolio the results are quite mixed. The dominance of the autoencoder-based portfolio is more evident for the average drawdown, where 335 records better results only for the Maximum Diversification and the Probabilistic Sharpe Ratio models. Again, the differences between two portfolios sharpen when considering the risk adjusted performance indicators. Furthermore, also in this case, the result of PSR measure is rather indicative. While in fact for portfolio 40 different values are greater than zero, for portfolio 335 this happens only three times and at a negligible level. Indeed with regard to the PSR (to be intended as the allocation model) which is the method portfolio 335 comes closest to dominating its competitor, the probability that the Sharpe Ratio is greater than 1 is 0.05%, while in portfolio 40 this same probability is 1.25%. As for the ISR, the results are also slightly better than the previous ones for portfolio 40, which shows this time 5 models with an ISR value greater than 0.4.
Finally, the indicators linked to the bullish and bearish periods also highlight a certain dominance of portfolio 40, especially as regards the recover capacity from a bearish period.

Markowitz model. Going into the detail of the pairs of portfolios distinguished by asset allocation techniques, we see how the Markowitz model has generated extremely violent equity lines in both cases.

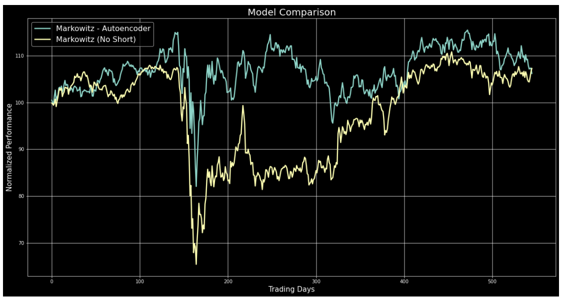

Figure 1.50: Markowitz equity lines.

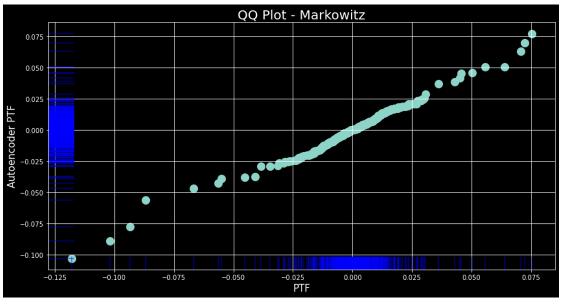

Figure 1.51: Markowitz QQ plot.

More specifically, portfolio 40 clearly shows greater resistance to pandemic stress, however helped by the advantage gained up to a few days before the crisis. Furthermore, its equity line has dominated for longer time. On the other aspects the performance is disappointing, full of side trends with a lot of volatility inside the trend. The QQ plot shows how the link between the distributions of the two series of returns tends to become vaguely sinusoidal, and how these then diverge on both tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mar(A)	-0.28	-0.053	-0.021	-0.035	0.0101	0.014	-0.69	10.95	0.10
mar	-0.39	-0.094	-0.021	-0.041	0.0131	0.016	-1.27	13.47	0.12

Table 1.59: Markowitz model: performance indicators.

The slightly better performance of the larger portfolio in terms of expected return impacts the Sharpe Ratio enough for the portfolio to outperform its competitor by 0.02 units. Given that neither of the two portfolios is even remotely close to any investor's ideal, it is true that portfolio 40's absolute dominance in terms of risk profile cannot be overlooked.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.92	0.25*

Table 1.60: Markowitz model: econometric tests.

Last, the econometric tests are in agreement in not distinguishing the distributions generating the series of returns of the two portfolios.

Markowitz model, no short sales. Restricting short-selling strongly improves the the results on both portfolios.

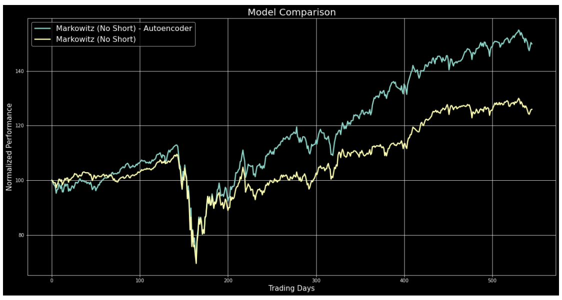

Figure 1.52: Markowitz(NS) equity lines.

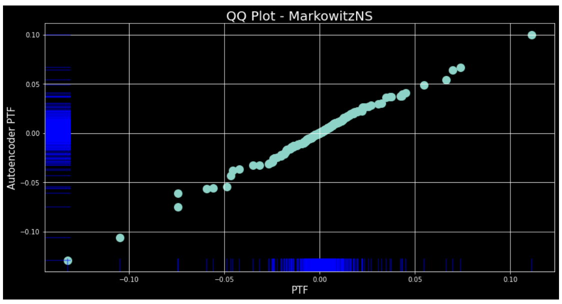

Figure 1.53: Markowitz(NS) QQ plot.

Portfolio 40 obtains a significantly better performance than its competitor. This overperformance is also in this case based on those periods where the larger one has a lateral or tentatively upward trend. In fact, in these trading days, portfolio 40 shows significant growth.

The QQ plot shows distortions in the percentiles of the returns for the two investment strategies especially at the beginning of the tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
marNS(A)	-0.36	-0.036	-0.022	-0.041	0.0745	0.016	-1.30	16.32	0.73
marNS	-0.36	-0.044	-0.019	-0.040	0.0423	0.016	-1.00	19.02	0.41

Table 1.61: Markowitz model with no short sales: performance indicators.

The portfolios have similar risk profiles except for the average drawdown, where portfolio 40 presents a lower risk.
Its dominance then takes shape with respect to the average return, which entails a growth in the Sharpe Ratio of 0.32.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.61	0.25*

Table 1.62: Markowitz model with no short sales: econometric tests.

Despite the very evident difference in performance between the two investments, the KS and AD tests struggle enormously to distinguish the data generating processes that led to the two portfolios returns.

Minimum Variance model. As in the case of the other two covariance estimation methods, the Minimum Variance portfolio performs very well for both portfolios.

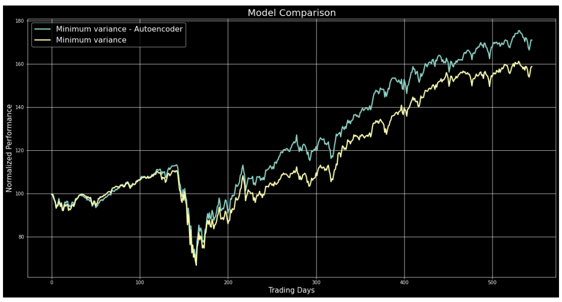

Figure 1.54: MV equity lines.

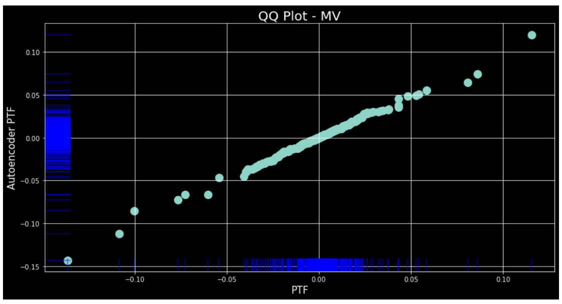

Figure 1.55: MV QQ plot.

The two portfolios grow quite similarly, with a dominance of portfolio 40 starting immediately after the pandemic and then gradually consolidating. The QQ plot shows a strong similarity as regards the returns concentrated around the mean, which then disappears as we approach to both tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mv(A)	-0.38	-0.040	-0.026	-0.045	0.0993	0.017	-1.16	15.72	0.87
mv	-0.39	-0.045	-0.025	-0.046	0.0855	0.018	-0.95	13.93	0.73

Table 1.63: Minimum Variance model: performance indicators.

The risk indicators are fairly aligned for both portfolios, with portfolio 40 being slightly preferable as for the drawdowns. In terms of performance, portfolio 40 manages to outperform its competitor, also recording a supplementar Sharpe Ratio of 0.14.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.99	0.25*

Table 1.64: Minimum Variance model: econometric tests.

The results of the econometric tests reflect the strong similarity in terms of returns of the two portfolios, and the null hypothesis cannot possibly be rejected.

Minimum Variance model, no short sales. In the case of Minimum Variance portfolio with short sales not allowed, the choice of the basket of 40 assets based on the Autoencoder generates a significant improvement over the 335 portfolio.

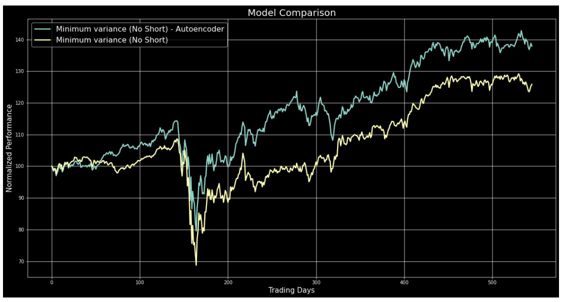

Figure 1.56: MV(NS) equity lines.

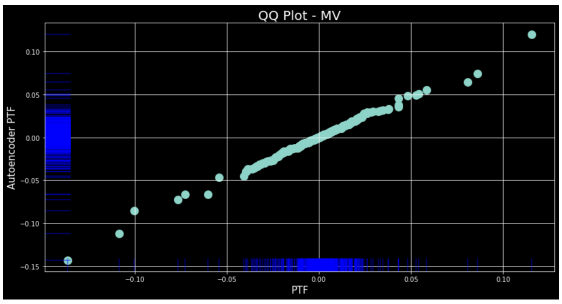

Figure 1.57: MV(NS) QQ plot.

The dominance of portfolio 40 starts already after the first quarter of trading and is consolidating over time. This also allows it to withstand the shock generated by the pandemic much better than the its competitor. In the recovery phase, the investment based on the autoencoder model is much more reactive and sensitive to the upward trend. Part of this success is also due to the ability to preserve an upward trend even when its competitor is characterized by lateral movements. After this phase, the smaller portfolio has a significant drawdown where the difference in equity lines with the 335 tapers considerably.

After this, the difference becomes significant again.
The QQ plot shows how around the beginning of the two tails, especially the right one, there are small deviations between the distributions of the yields of the two portfolios.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mvNS(A)	-0.30	-0.032	-0.017	-0.036	0.0592	0.014	-0.91	15.40	0.62
mvNS	-0.36	-0.047	-0.019	-0.040	0.0421	0.016	-0.95	18.66	0.41

Table 1.65: Minimum Variance model with no short sales: performance indicators.

It can be inferred from the performance indicators that portfolio 40 is preferable to the 335 with respect to all risk indicators, in particular for the drawdowns. Compared to the average return, the difference is less marked than in other cases, but sufficient to lead to a difference in terms of Sharpe Ratio equal to 0.21 per unit of risk.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.56	0.07

Table 1.66: Minimum Variance model with no short sales: econometric tests.

The two econometric tests are in this case evidently at odds. This is not surprising, given the structural characteristics of KS, which only examines the point of maximum divergence of the two distributions.

Maximum Sharpe Ratio model. The maximum Sharpe Ratio portfolio, on the other hand, is one of those where the superiority of the portfolio based on the Autoencoder model is more questionable.

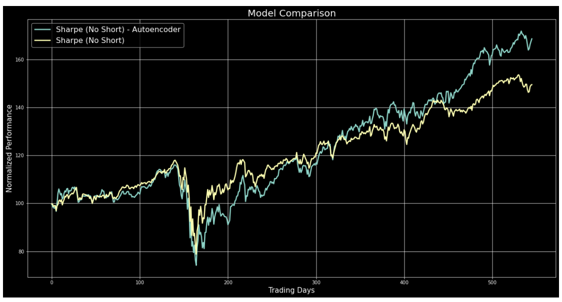

Figure 1.58: SR equity lines.

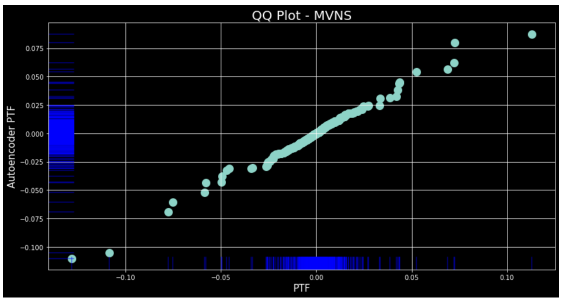

Figure 1.59: SR QQ plot.

Portfolio 335 reacts slightly better to the systemic crisis linked to the pandemic event, and clearly better to the immediate subsequent recovery. The superiority of the stock-picking procedure based portfolkio manifests only in the second half of the test set period, and is really consolidated only in the last year of trading. The QQ plot shows deviations between the two distributions starting from the beginning of both tails.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
srNS(A)	-0.36	-0.043	-0.023	-0.044	0.0966	0.018	-0.83	11.04	0.85
srNS	-0.33	-0.030	-0.021	-0.039	0.0742	0.016	-0.77	18.31	0.73

Table 1.67: Maximum Sharpe Ratio model: performance indicators.

Across all the risk measures - including the average drawdown, where the 40 portfolio has always done better - and with the exception of kurtosis, portfolio 335 overperforms its competitor. Portfolio 40, on the other hand, dominates the average yield, thus registering a Sharpe Ratio increased by 0.12.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.27	0.25*

Table 1.68: Maximum Sharpe Ratio model: econometric tests.

Econometric tests are in this case more oriented towards the hypothesis of returns sourcing from different data generating processes.

Risk Parity model. With the Risk Parity asset allocation model, the two portfolios show a similar trend, but also in this case with an undeniable dominance of the portfolio 40.

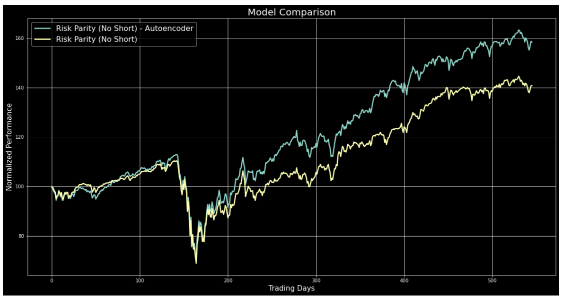

Figure 1.60: RP equity lines.

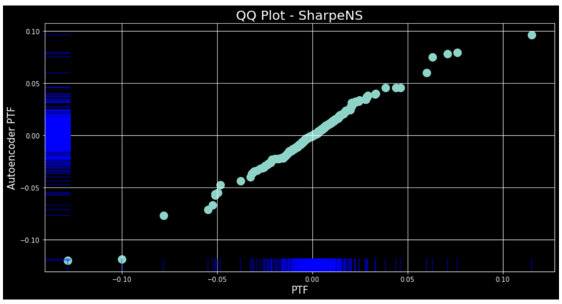

Figure 1.61: RP QQ plot.

The dominance of the autoencoded portfolio begins immediately at the start of the bullish market rally and improves more and more over time. Also in this case, part of the dominance is substantiated in some brief moments in which a side phase of portfolio 335 is contrasted by a convinced bullish phase of portfolio 40. The QQ plot shows divergences arising in the last part of the shoulders of the two distributions and starting from the tail areas.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
rp(A)	-0.37	-0.037	-0.023	-0.043	0.0848	0.016	-1.22	16.32	0.80
rp	-0.37	-0.043	-0.022	-0.042	0.0629	0.016	-1.00	15.81	0.59

Table 1.69: Risk Parity model: performance indicators.

The risk indicators are quite similar for both portfolio models, except for the average drawdown where 40 portfolio does better. The performance, on the other hand, improves considerably if we look at the average yield, and consequently at the Sharpe Ratio, where an increase of 0.21 is recorded with respect to portfolio 335.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.71	0.10

Table 1.70: Risk Parity model: econometric tests.

The econometric tests shows a marked divergence between the KS and AD. For reasons already mentioned, there is a tendency to give greater credit to the outcome of AD.

Maximum Diversification model. In the Maximum Diversification portfolio it is more difficult to distinguish a clear predominance of one portfolio against the other.

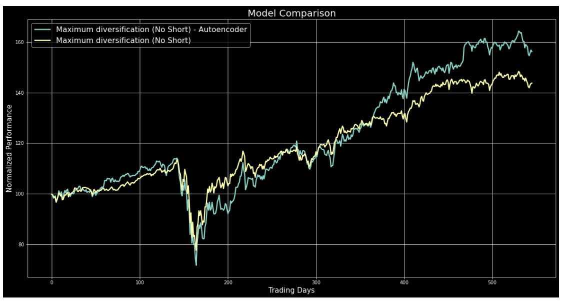

Figure 1.62: MD equity lines.

Figure 1.63: MD QQ plot.

After an initial period of slight overperformance, autoencoder-based portfolio reacts less well than its competitor to the pandemic crisis and is also less reactive in terms of recovery.

The overperfomance of portfolio 40 is substantiated above all in the last year of trading, where for short intervals the two equity lines become anti-correlated. The QQ plot shows clear deviations near both tails of the distributions, especially the right one.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
mdNS(A)	-0.37	-0.039	-0.021	-0.041	0.0825	0.016	-0.75	13.74	0.79
mdNS	-0.31	-0.025	-0.016	-0.035	0.0670	0.014	-0.59	18.97	0.73

Table 1.71: Maximum Diversification model: performance indicators.

The dominance of the larger portfolio in terms of risk is total and clear. The higher expected return brought by the 40 portfolio also struggles to compensate for the greater volatility, from which the difference in Sharpe Ratio is reduced to a narrow 0.06.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.18	0.10

Table 1.72: Maximum Diversification model: econometric tests.

Econometric tests are quite in agreement in favoring the idea of two series of returns produced by different data generating processes.

Probabilistic Sharpe Ratio model. The case of Probabilistic Sharpe Ratio strongly retraces the footsteps of the maximum Sharpe Ratio.

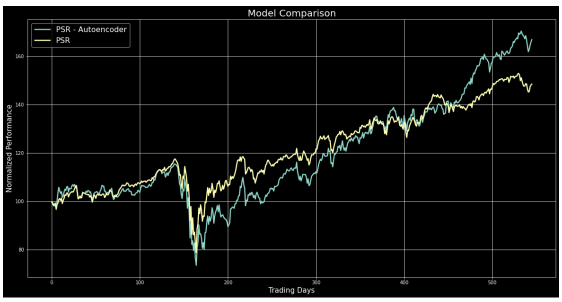

Figure 1.64: PSR equity lines.

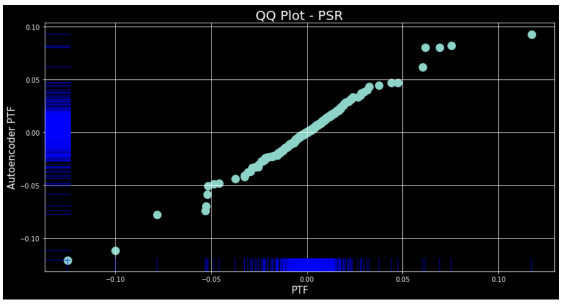

Figure 1.65: PSR QQ plot.

Also in this case, the dominance of portfolio 40 takes shape only in the last part of the data, where it consolidates markedly. For the rest, the portfolio is often dominated by 335, which also reacts a little better to the pandemic collapse. The QQ plot shows very clear deviations in the tail area, where L-shaped patterns are shown.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
psr(A)	-0.36	-0.047	-0.023	-0.045	0.0947	0.018	-0.76	10.32	0.82
psr	-0.32	-0.030	-0.021	-0.039	0.0729	0.015	-0.67	17.91	0.72

Table 1.73: Probabilistic Sharpe Ratio model: performance indicators.

The dominance of portfolio 335 is clear in terms of performance indicators, and the only truly significant parameters in which its competitor does better are the expected return and the Sharpe Ratio.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.21	0.06

Table 1.74: Probabilistic Sharpe Ratio model: econometric tests.

Econometric tests are also in this case more likely to distinguish the probability distributions of the returns for the two investment strategies.

Hierarchical Risk Parity model. The hierarchical Risk Parity model explicitly rewards portfolio 40 in this case as well.

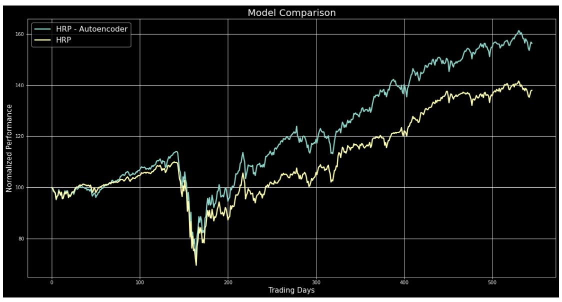

Figure 1.66: HRP equity lines.

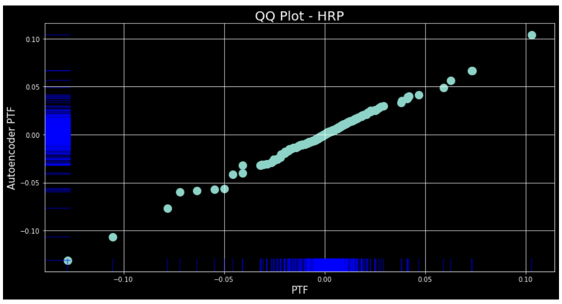

Figure 1.67: HRP QQ plot.

During the post-pandemic bullish rally in financial markets, portfolio 40 immediately begins its dominance. This is gradually widening more and more, often exploiting lateral periods of the competitor portfolio, to which it opposes markedly bullish periods. The QQ plot shows a clear distortion between the two distributions on the left tail, to the advantage of portfolio 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt	SR
hrp(A)	-0.36	-0.034	-0.022	-0.041	0.0825	0.016	-1.25	16.70	0.80
hrp	-0.36	-0.042	-0.021	-0.041	0.0593	0.016	-1.02	16.06	0.57

Table 1.75: Hierarchical Risk Parity model: performance indicators.

On the risk side, two portfolios have rather similar profiles, except for the average drawdown where there is an overperformance of autoencoder-based portfolio. Instead the supremacy of the latter strongly emerges with respect to the expected return and consequently on the Sharpe Ratio, where there is an improvement over the competitor of 0.23.

	Kolmogorov-Smirnov	Anderson–Darling
ρ-value	0.51	0.25*

Table 1.76: Hierarchical Risk Parity model: econometric tests.

Econometric tests offer hesitant results on the eventual rejection of the null hypothesis, however leaning towards the impossibility of rejecting it.

RMT Covariance: rolling portfolios analysis. Last, the results of rolling portfolio experiment carried out on Random Matrix Theory tools are presented.

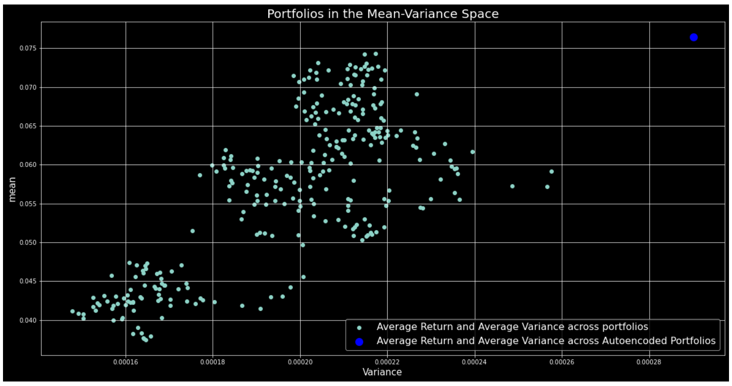

Figure 1.68: RMT based Covariance, Mean-Variance Space: portfolio 40 against competitors.

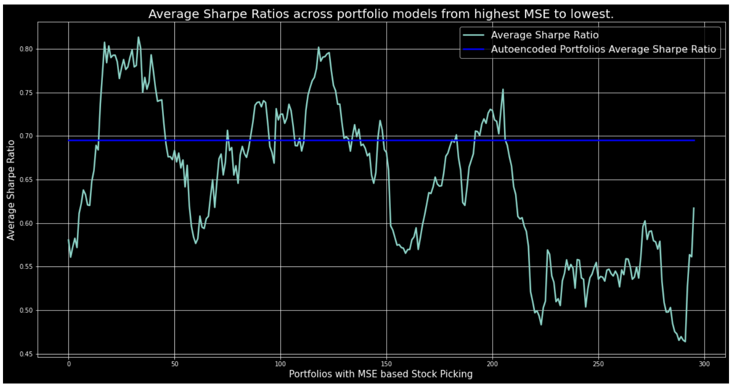

Figure 1.69: RMT based Covariance, Average Sharpe Ratio: portfolio 40 against competitors.

Again, portfolio 40 has both the highest return and the highest risk in the entire mean-variance space. However, the dominance in terms of returns is less marked than the weakness in terms of risk.

The average Sharpe Ratio obtained allows the Autoencoder-based portfolio to overperform 68% of its competitors. The result is worse than in the other two cases, but it remains significant.

The direct relationship among average Sharpe ratio and Mean Square Error is confirmed.

1.4.2 Portfolio 40 vs Portfolio 335 vs SP500

In this section the SP500 index is included in the comparison, so to briefly investigate the over or underperformance of the portfolios with respect to the index, even if the Information Sharpe Ratio partially answered to this question.
For sake of brevity, this focus has been done only with respect to historical covariance matrix estimation.
From the equity lines it can be immediately seen that, with the exception of the Markowitz and Minimum Variance models (NS) portfolio 40 always beat the SP500 in terms of overall performance, remaining dominant over the entire period considered.
On the other hand, portfolio 335 is overperformed by the SP500 with respect to Markowitz (both versions), Minimum Variance (NS) and Risk Parity, while is more or less similar as regards the Probabilistic Sharpe Ratio model and Hierarchical Risk Parity.
During the pandemic-related crash, the SP500 performed roughly in line with both portfolios.

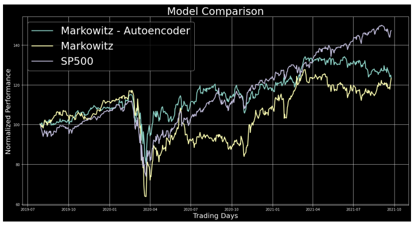

Figure 1.70: Markowitz.

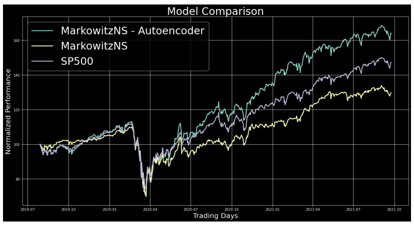

Figure 1.71: Markowitz (NS).

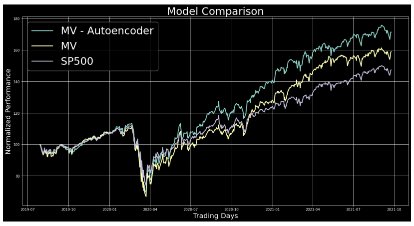

Figure 1.72: Minimum Variance.

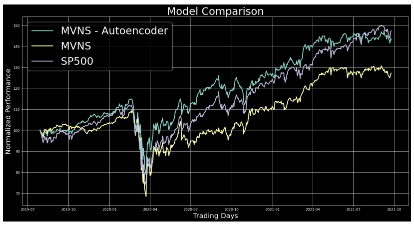

Figure 1.73: Min Variance (NS).

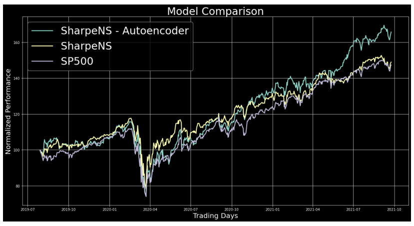

Figure 1.74: Sharpe.

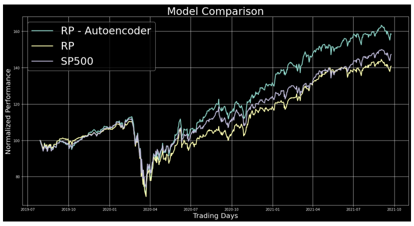

Figure 1.75: Risk Parity.

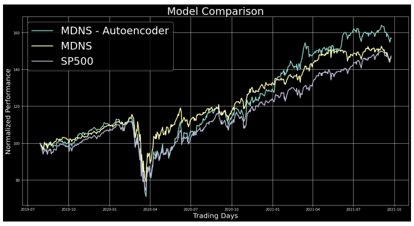

Figure 1.76: MD.

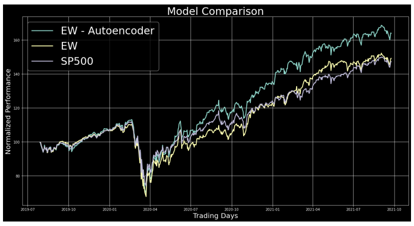

Figure 1.77: Equally Weighted.

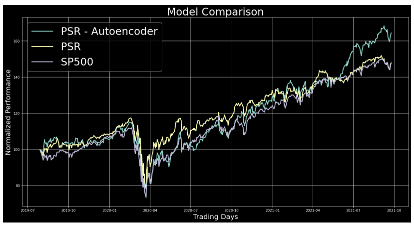

Figure 1.78: PSR.

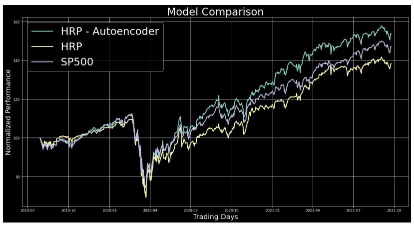

Figure 1.79: HRP.

The performance indicators confirm the strength of autoencoder-based portfolios. While for measures of extreme risk all the portfolios are roughly aligned, things dramatically change as for pure and risk-adjusted performance indicators.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurt
Markowitz	-0.30	-0.034	-0.020	-0.036	0.0381	0.015	-0.65	12.99
MarkowitzNS	-0.38	-0.039	-0.024	-0.043	0.0876	0.017	-1.19	16.05
MV	-0.38	-0.040	-0.026	-0.045	0.0996	0.018	-1.16	15.71
MVNS	-0.31	-0.030	-0.018	-0.037	0.0659	0.015	-0.89	15.72
SharpeNS	-0.35	-0.043	-0.023	-0.043	0.0931	0.017	-0.79	11.19
RP	-0.37	-0.038	-0.023	-0.043	0.0849	0.016	-1.22	16.29
MDNS	-0.37	-0.040	-0.022	-0.041	0.0832	0.016	-0.71	13.42
EW	-0.38	-0.039	-0.025	-0.044	0.0914	0.017	-1.20	16.07
PSR	-0.36	-0.046	-0.024	-0.044	0.0917	0.018	-0.75	10.55
HRP	-0.37	-0.037	-0.022	-0.042	0.0792	0.016	-1.22	16.61

Table 1.77: Historical Covariance: risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	0.401061	0.00	-0.57	0.575760	0.893176
MarkowitzNS	0.810342	1.79	0.48	1.048760	1.620289
MV	0.878966	10.03	0.75	1.148087	1.814349
MVNS	0.694543	0.01	-0.14	0.948660	1.480147
SharpeNS	0.825205	1.34	0.41	1.133741	1.834810
RP	0.803024	1.47	0.42	1.032308	1.594957
MDNS	0.795873	0.57	0.26	1.084583	1.564518
EW	0.836370	3.82	0.58	1.082632	1.687287
PSR	0.807292	0.58	0.36	1.114666	1.790694
HRP	0.757377	0.28	0.25	0.977133	1.504395

Table 1.78: Historical Covariance: performance indicators adjusted for risk for portfolios 40

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.006597	1.869464	0.735780
MarkowitzNS	1.081423	19.973684	0.930275
MV	1.083942	20.554054	0.932110
MVNS	1.027957	7.500000	0.882569
SharpeNS	1.039660	20.760000	0.954128
RP	1.106788	25.700000	0.944954
MDNS	1.082768	20.760000	0.954128
EW	1.081423	19.973684	0.930275
PSR	1.090323	15.000000	0.937615
HRP	1.106788	25.700000	0.944954

Table 1.79: Historical Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.45	-0.10	-0.025	-0.047	0.0381	0.019	-0.59	14.00
MarkowitzNS	-0.36	-0.04	-0.020	-0.040	0.0448	0.016	-1.07	17.16
MV	-0.39	-0.04	-0.025	-0.046	0.0854	0.018	-0.95	13.08
MVNS	-0.37	-0.04	-0.019	-0.041	0.0440	0.016	-0.96	17.90
SharpeNS	-0.33	-0.03	-0.021	-0.039	0.0733	0.016	-0.79	18.28
RP	-0.37	-0.04	-0.022	-0.042	0.0629	0.016	-1.09	15.81
MDNS	-0.31	-0.02	-0.016	-0.035	0.0699	0.014	-0.56	18.77
EW	-0.38	-0.04	-0.023	-0.044	0.0738	0.017	-0.98	14.87
PSR	-0.33	-0.03	-0.021	-0.039	0.0729	0.016	-0.82	18.40
HRP	-0.36	-0.04	-0.021	-0.041	0.0596	0.016	-1.04	16.30

Table 1.80: Historical Covariance: risk indicators and mean returns for portfolios 335

	SR	PSR	ISR	SoR	CR
Markowitz	0.306303	0.00	-0.37	0.432071	0.589488
MarkowitzNS	0.445047	0.00	-0.65	0.561285	0.875811
MV	0.736904	0.05	0.39	0.964208	1.523491
MVNS	0.424774	0.00	-0.56	0.544307	0.829478
SharpeNS	0.727181	0.06	0.04	0.940965	1.540999
RP	0.595316	0.00	-0.26	0.757018	1.183745
MDNS	0.761140	0.27	-0.03	0.998191	1.552187
EW	0.668725	0.00	0.07	0.852066	1.351297
PSR	0.721532	0.05	0.19	0.927101	1.522145
HRP	0.578651	0.00	-0.36	0.734309	1.142051

Table 1.81: Historical Covariance: performance indicators adjusted for risk for portfolios 335

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.007620	1.484018	0.598165
MarkowitzNS	1.051383	13.315789	0.930275
MV	1.059523	14.820000	0.908257
MVNS	1.054655	14.111111	0.933945
SharpeNS	1.051383	13.315789	0.930275
RP	1.058190	14.500000	0.906422
MDNS	1.035341	6.586667	0.906422
EW	1.062366	15.500000	0.911927
PSR	1.051383	13.315789	0.930275
HRP	1.049880	12.948718	0.928440

Table 1.82: Historical Covariance: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
ˆSP500	-0.34	-0.038	-0.024	-0.041729	0.0709	0.016	-1.09	15.87

Table 1.83: SP500: risk indicators and mean returns

	SR	PSR	ISR	SoR	CR
ˆSP500	0.704221	0.01	0.631739	0.942873	1.474809

Table 1.84: SP500: performance indicators adjusted for risk

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
ˆSP500	1.060913	15.153061	0.910092

Table 1.85: SP500: bull/bear persistence, bull/bear recovery and bull dominance

Eight out of ten asset allocation models in the Portfolio 40 group outperform the SP500 in terms of Sharpe Ratio and Probabilistic Sharpe Ratio. With regards to Information Sharpe Ratio, Sortino and Calmar ratio the outperforming models rise to nine out of ten. If the focus is shifted to portfolio 335 things are far worse. Only five out of ten models outperform the SP500 in terms of Sharpe Ratio, only four if as regards Probabilistic Sharpe Ratio, Information Sharpe Ratio and Calmar Ratio, and just two considering only the downside risk through Sortino Ratio.

1.4.3 Stock picking on validation set

Stock-picking performed on the basis of the train set MSE on the one hand provides the robustness of a very long period of time that includes different cycles, topical moments and regimes, on the other side it is anchored to excessively old information.
In this section, both previous analyzes are repeated, but the 40 assets are picked on the validation set MSE.
Experiments are still organized with respect to covariance matrix estimation techniques and across the same asset allocation models.

Historical Covariance Matrix

As can be seen from the equity lines, the dominance of the 40 portfolio has on average significantly increased.
The first thing to be noticed is that portfolio 40 overperforms the benchmark for 8 asset allocation models out of 10, while portfolio 335 only in three cases, in terms of pure performance.

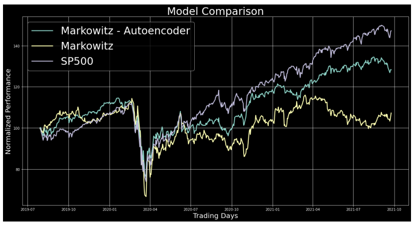

Figure 1.80: Markowitz.

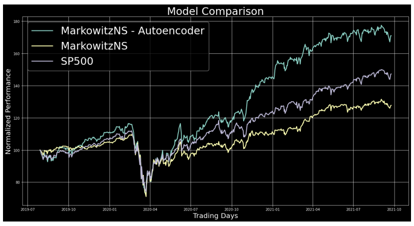

Figure 1.81: Markowitz (NS).

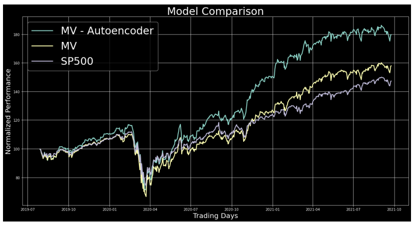

Figure 1.82: Minimum Variance.

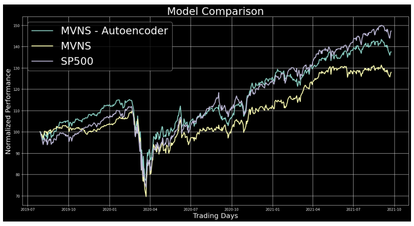

Figure 1.83: Min Variance (NS).

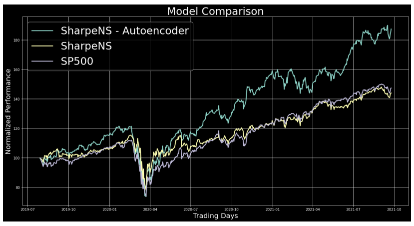

Figure 1.84: Sharpe.

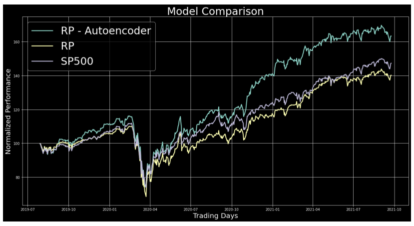

Figure 1.85: Risk Parity.

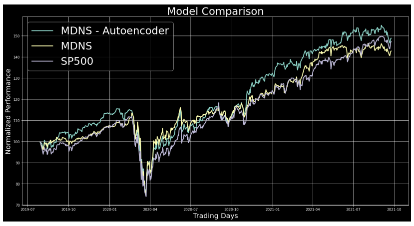

Figure 1.86: MD.

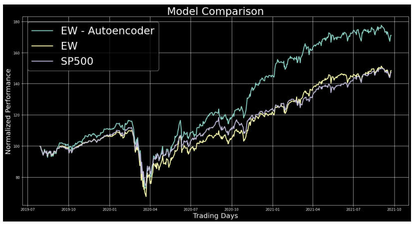

Figure 1.87: Equally Weighted.

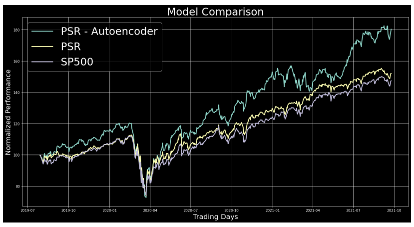

Figure 1.88: PSR.

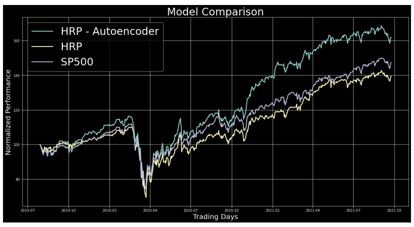

Figure 1.89: HRP.

The graphic intuition is fully confirmed by the performance metrics. The two groups of portfolios tend to be equivalent for maximum drawdown, while for the mean drawdown, portfolio 40 dominates, except for maximum Sharpe, maximum diversification and PSR. Compared to VaR a slight but not negligible prevalence of portfolio 335 is recorded, while for cVaR the two groups are on average equivalent. Things dramatically change for risk-adjusted performance, in favor of portfolio 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.34	-0.052	-0.020	-0.036026	0.0455	0.015	-0.63	11.98
MarkowitzNS	-0.38	-0.040	-0.025	-0.043186	0.0991	0.017	-0.90	13.46
MV	-0.39	-0.042	-0.027	-0.045393	0.1082	0.018	-0.90	13.68
MVNS	-0.33	-0.039	-0.020	-0.036276	0.0583	0.015	-0.81	14.16
SharpeNS	-0.39	-0.041	-0.026	-0.044261	0.1162	0.018	-0.31	15.20
RP	-0.36	-0.038	-0.024	-0.041290	0.0904	0.016	-0.90	13.34
MDNS	-0.34	-0.035	-0.020	-0.037422	0.0736	0.015	-0.76	13.65
EW	-0.38	-0.040	-0.025	-0.043186	0.0991	0.017	-0.90	13.46
PSR	-0.40	-0.043	-0.027	-0.044975	0.1091	0.018	-0.32	14.99
HRP	-0.36	-0.039	-0.024	-0.040922	0.0888	0.016	-0.86	12.79

Table 1.86: Risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	0.495529	0.00	-0.586259	0.701045	0.957634
MarkowitzNS	0.930931	22.49	0.717242	1.270438	1.860189
MV	0.965317	35.63	0.856362	1.322415	1.949844
MVNS	0.637336	0.00	-0.338095	0.863752	1.241241
SharpeNS	1.014559	56.06	0.787436	1.436505	2.087019
RP	0.888879	10.40	0.528155	1.205863	1.759570
MDNS	0.776819	0.27	0.059381	1.066253	1.525351
EW	0.930931	22.49	0.717242	1.270438	1.860189
PSR	0.947586	28.10	0.658163	1.342245	1.942064
HRP	0.879680	8.13	0.475333	1.195294	1.720440

Table 1.87: Performance indicators adjusted for risk for portfolios 40

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.006362	1.834483	0.732110
MarkowitzNS	1.070709	17.476744	0.921101
MV	1.070709	17.476744	0.921101
MVNS	1.053523	9.264151	0.902752
SharpeNS	1.022350	11.952381	0.922936
RP	1.074657	18.402439	0.924771
MDNS	1.036846	6.555556	0.867890
EW	1.070709	17.476744	0.921101
PSR	1.039660	20.760000	0.954128
HRP	1.074657	18.402439	0.924771

Table 1.88: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.42	-0.099	-0.021	-0.043566	0.0134	0.018	-0.43	11.87
MarkowitzNS	-0.35	-0.042	-0.020	-0.039022	0.0449	0.016	-0.88	16.56
MV	-0.40	-0.045	-0.025	-0.046483	0.0844	0.018	-0.99	14.23
MVNS	-0.37	-0.045	-0.020	-0.040224	0.0451	0.016	-0.85	17.38
SharpeNS	-0.32	-0.032	-0.021	-0.039681	0.0669	0.016	-0.87	17.86
RP	-0.38	-0.044	-0.022	-0.042434	0.0621	0.017	-1.04	16.11
MDNS	-0.34	-0.029	-0.018	-0.037239	0.0656	0.015	-0.68	18.72
EW	-0.39	-0.044	-0.024	-0.044337	0.0728	0.018	-1.03	15.18
PSR	-0.35	-0.031	-0.021	-0.040752	0.0774	0.016	-1.01	17.46
HRP	-0.37	-0.042	-0.022	-0.041566	0.0609	0.016	-1.10	16.57

Table 1.89: Risk indicators and mean returns for portfolios 335

	SR	PSR	ISR	SoR	CR
Markowitz	0.118370	0.00	-0.726115	0.168290	0.226532
MarkowitzNS	0.458296	0.00	-0.671223	0.572512	0.899792
MV	0.727055	0.03	0.374349	0.932410	1.505839
MVNS	0.446255	0.00	-0.585316	0.566083	0.870757
SharpeNS	0.666202	0.00	-0.107145	0.840897	1.459442
RP	0.585212	0.00	-0.285752	0.728003	1.163923
MDNS	0.685506	0.00	-0.141137	0.876941	1.345121
EW	0.657886	0.00	0.047422	0.826281	1.330298
PSR	0.758611	0.28	0.238899	0.946178	1.580979
HRP	0.588354	0.00	-0.331294	0.733006	1.163508

Table 1.90: Performance indicators adjusted for risk for portfolios 335

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.005213	1.407080	0.585321
MarkowitzNS	1.054655	14.111111	0.933945
MV	1.060913	15.153061	0.910092
MVNS	1.056439	14.542857	0.935780
SharpeNS	1.051383	13.315789	0.930275
RP	1.062366	15.500000	0.911927
MDNS	1.054655	14.111111	0.933945
EW	1.062366	15.500000	0.911927
PSR	1.072633	17.928571	0.922936
HRP	1.049880	12.948718	0.928440

Table 1.91: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335

As for Sharpe Ratio, all the models in portfolio 40 offer better performance and on average the difference between the two groups of portfolios is around 0.28 return per unit of risk. The difference in terms of PSR is evident, while as regards the Information Sharpe Ratio it can be seen how the 335 portfolio offers a negative performance compared to the benchmark in 7 cases out of 10, and in none of the remaining 3 manages to touch the barrier of 0.4. On the other hand, portfolio 40 recorded only two cases of negative excess return - Markowitz and Minimum Variance (both NS) - settling on a positive excess return in all other cases, also exceeding the threshold of 0.4 in 7 cases out of 10. In terms of Sortino and Calmar Ratio, portfolio 40 dominance becomes even more pronounced.
With respect to benchmark, the good overperformances recorded on the train set stock-picking procedure are here largerly improved.
For the rolling portfolio analysis, the results are excellent. Indeed, in the pseudo mean-variance space it can be observed that autoencoder-based portfolio overperforms almost all the competitors both in terms of variance and return.

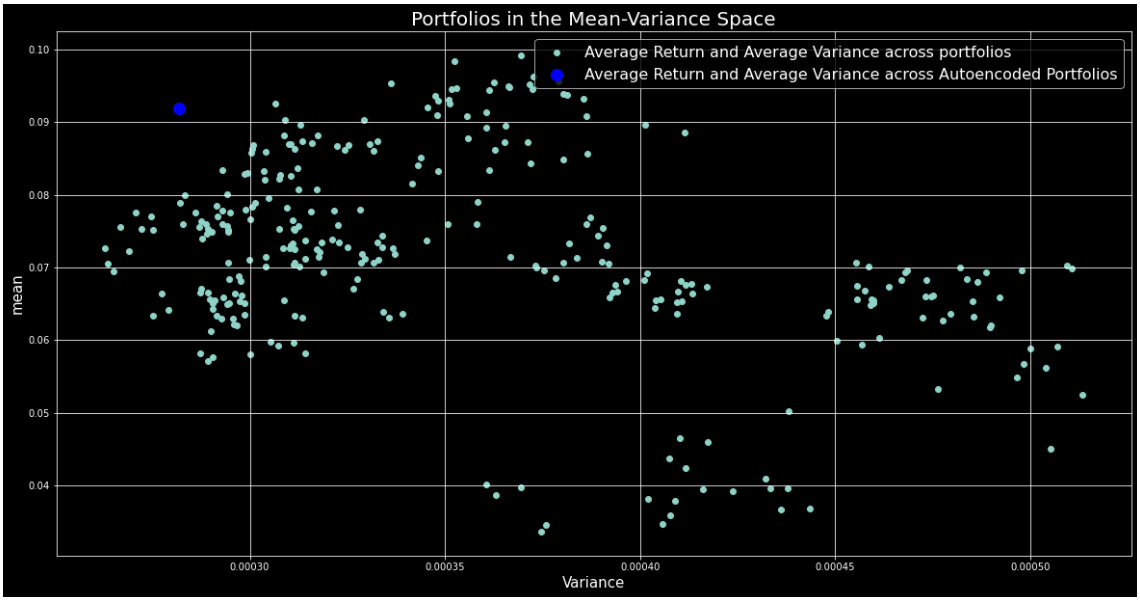

Figure 1.90: Mean-Variance Space: portfolio 40 against competitors.

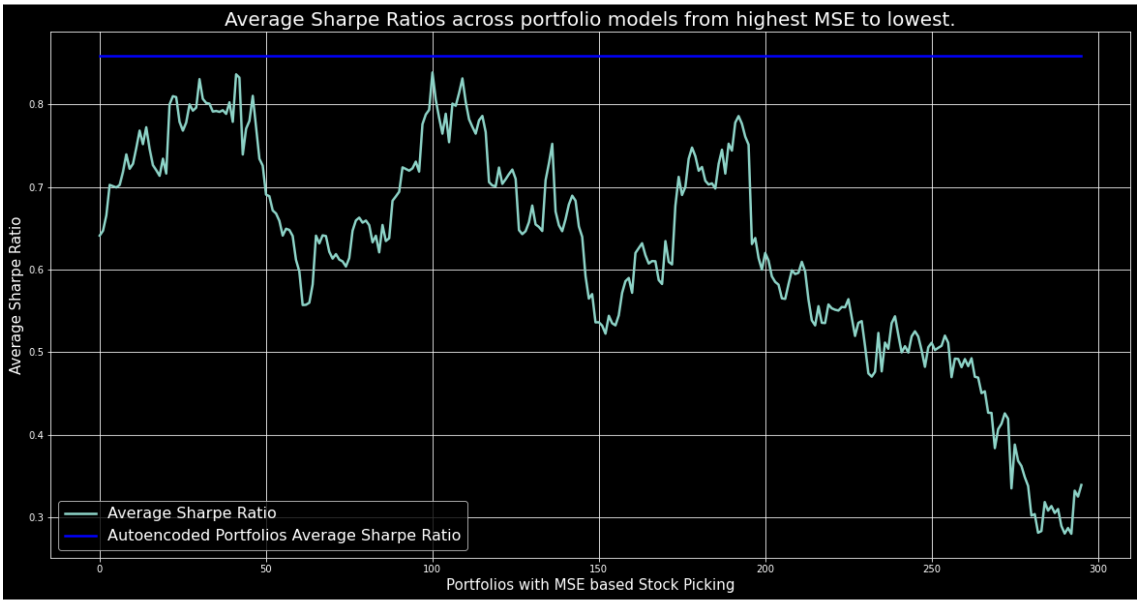

Figure 1.91: Average Sharpe Ratio: portfolio 40 against competitors.

The good balance among the average performance and risk is confirmed by the average Sharpe ratio, with respect to which portfolio 40 overperforms the 100% of its competitors. Please note that also in this case the decreasing trend of the average Sharpe ratios with respect to the MSE is strongly confirmed.

Exponentially Weighted Covariance Matrix

Even for the covariance matrix estimated by exponentially weighting deviations from mean returns, the dominance of portfolio 40 is undisputed.
Again it overpeforms the benchmark in 8 cases out of 10, against the 4 out of 10 of portfolio 335.

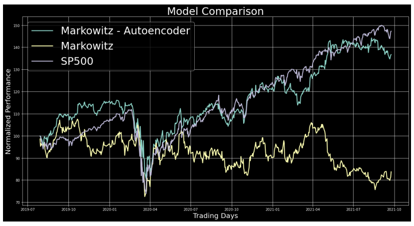

Figure 1.92: Markowitz.

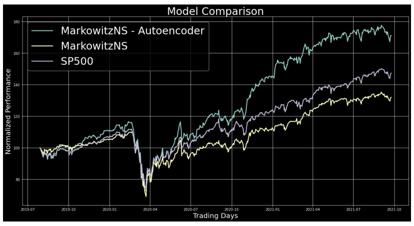

Figure 1.93: Markowitz (NS).

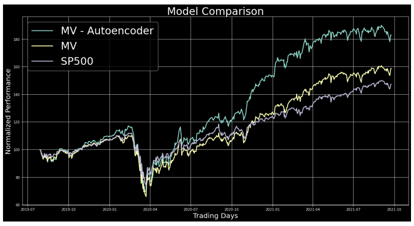

Figure 1.94: Minimum Variance.

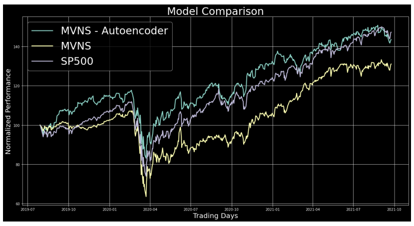

Figure 1.95: Min Variance (NS).

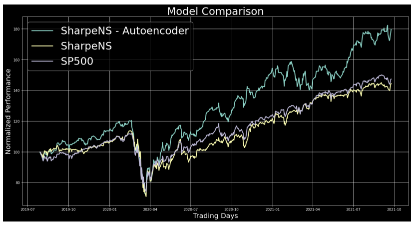

Figure 1.96: Sharpe.

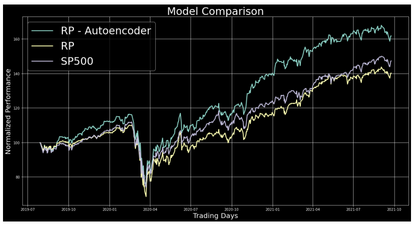

Figure 1.97: Risk Parity.

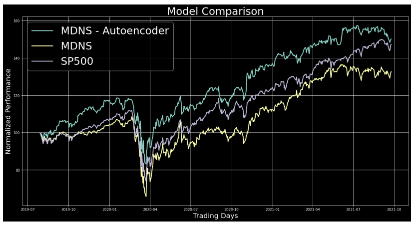

Figure 1.98: MD.

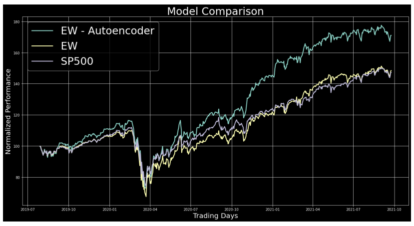

Figure 1.99: Equally Weighted.

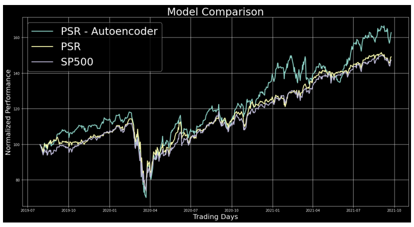

Figure 1.100: PSR.

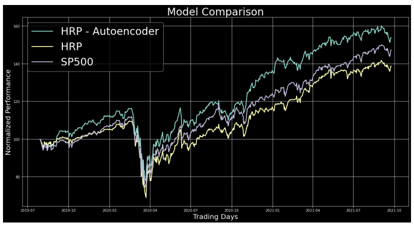

Figure 1.101: HRP.

In terms of pure risk there is however a certain balance between the two groups of portfolios. Compared to the maximum and average drawdown, a slight preference for portfolio 40 is shown, while for the Value at Risk measures the situation is totally balanced.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.31	-0.045	-0.023	-0.038321	0.0574	0.016	-0.40	8.46
MarkowitzNS	-0.38	-0.040	-0.025	-0.043186	0.0991	0.017	-0.90	13.46
MV	-0.41	-0.045	-0.028	-0.047224	0.1116	0.019	-0.87	13.78
MVNS	-0.29	-0.034	-0.020	-0.035428	0.0666	0.014	-0.91	13.38
SharpeNS	-0.40	-0.049	-0.029	-0.046134	0.1086	0.019	-0.19	13.24
RP	-0.35	-0.036	-0.024	-0.039906	0.0887	0.016	-0.93	13.29
MDNS	-0.30	-0.031	-0.019	-0.036201	0.0752	0.015	-0.72	11.87
EW	-0.38	-0.040	-0.025	-0.043186	0.0991	0.017	-0.90	13.46
PSR	-0.41	-0.051	-0.028	-0.046765	0.0900	0.019	-0.46	13.18
HRP	-0.33	-0.034	-0.023	-0.037784	0.0794	0.015	-0.88	13.06

Table 1.92: Risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	0.581885	0.0	-0.238201	0.872397	1.326639
MarkowitzNS	0.930931	22.49	0.717242	1.267843	1.860189
MV	0.953836	31.04	0.850133	1.309355	1.935063
MVNS	0.735252	0.04	-0.106779	1.027315	1.600079
SharpeNS	0.927553	19.04	0.600445	1.344564	1.898665
RP	0.897927	12.57	0.495313	1.213650	1.798490
MDNS	0.807161	0.70	0.086108	1.146729	1.774548
EW	0.930931	22.49	0.717242	1.267843	1.860189
PSR	0.758636	0.07	0.292974	1.066798	1.563707
HRP	0.839844	2.87	0.231120	1.147746	1.694192

Table 1.93: Performance indicators adjusted for risk for portfolios 40

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.012451	2.787302	0.805505
MarkowitzNS	1.070709	17.476744	0.921101
MV	1.070709	17.476744	0.921101
MVNS	1.023810	4.666667	0.873394
SharpeNS	1.036594	9.666667	0.906422
RP	1.074657	18.402439	0.924771
MDNS	1.005896	2.304124	0.820183
EW	1.070709	17.476744	0.921101
PSR	1.043557	5.800000	0.853211
HRP	1.079043	19.423077	0.928440

Table 1.94: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.33	-0.136	-0.029	-0.046325	-0.0299	0.019	-0.54	4.27
MarkowitzNS	-0.37	-0.046	-0.021	-0.041203	0.0509	0.016	-1.16	16.89
MV	-0.40	-0.047	-0.026	-0.047195	0.0853	0.019	-0.96	13.90
MVNS	-0.41	-0.066	-0.022	-0.044304	0.0496	0.018	-1.19	15.53
SharpeNS	-0.38	-0.049	-0.023	-0.043981	0.0676	0.017	-1.61	20.02
RP	-0.38	-0.044	-0.022	-0.042393	0.0622	0.017	-1.06	16.16
MDNS	-0.39	-0.053	-0.022	-0.042821	0.0523	0.017	-0.85	13.03
EW	-0.39	-0.044	-0.024	-0.044337	0.0728	0.018	-1.03	15.18
PSR	-0.35	-0.036	-0.022	-0.041583	0.0737	0.016	-1.30	18.40
HRP	-0.37	-0.042	-0.021	-0.041540	0.0602	0.016	-1.13	16.43

Table 1.95: Risk indicators and mean returns for portfolios 335

	SR	PSR	ISR	SoR	CR
Markowitz	-0.246957	0.00	-0.847837	-0.385266	-0.649424
MarkowitzNS	0.493656	0.00	-0.595970	0.609688	0.967485
MV	0.721623	0.02	0.371111	0.936753	1.499354
MVNS	0.449092	0.00	-0.438826	0.567942	0.859886
SharpeNS	0.615441	0.00	-0.083528	0.746932	1.267748
RP	0.586721	0.00	-0.287502	0.731176	1.165575
MDNS	0.479709	0.00	-0.396674	0.631881	0.935752
EW	0.657886	0.00	0.047422	0.827821	1.330298
PSR	0.712625	0.05	0.074586	0.875162	1.467682
HRP	0.583273	0.00	-0.357036	0.722682	1.147922

Table 1.96: Performance indicators adjusted for risk for portfolios 335

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.000111	1.005405	0.456881
MarkowitzNS	1.054655	14.111111	0.933945
MV	1.059523	14.820000	0.908257
MVNS	1.054655	14.111111	0.933945
SharpeNS	1.051383	13.315789	0.930275
RP	1.062366	15.500000	0.911927
MDNS	1.062366	15.500000	0.911927
EW	1.062366	15.500000	0.911927
PSR	1.051383	13.315789	0.930275
HRP	1.049880	12.948718	0.928440

Table 1.97: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335

In terms of risk-adjusted performance, the domination of portfolio 40 is clear, with an average increase in the Sharpe ratio that is around 0.30 and which is heavily reflected in the PSR. What was deduced with respect to the benchmark is confirmed by the ISR, where only three times portfolio 335 generates an excess return, and there are no cases where Information Sharpe Ratio reach the threshold of 0.4. On the other hand, portfolio 40, in addition to generating excess returns in 8 out of 10 cases as already mentioned, exceeds the 0.4 threshold in 8 of them. The other indicators furtherly confirm the preferability of portfolio 40.
Also in this case the analysis on rolling portfolios confirms the excellent positioning of portfolio 40.

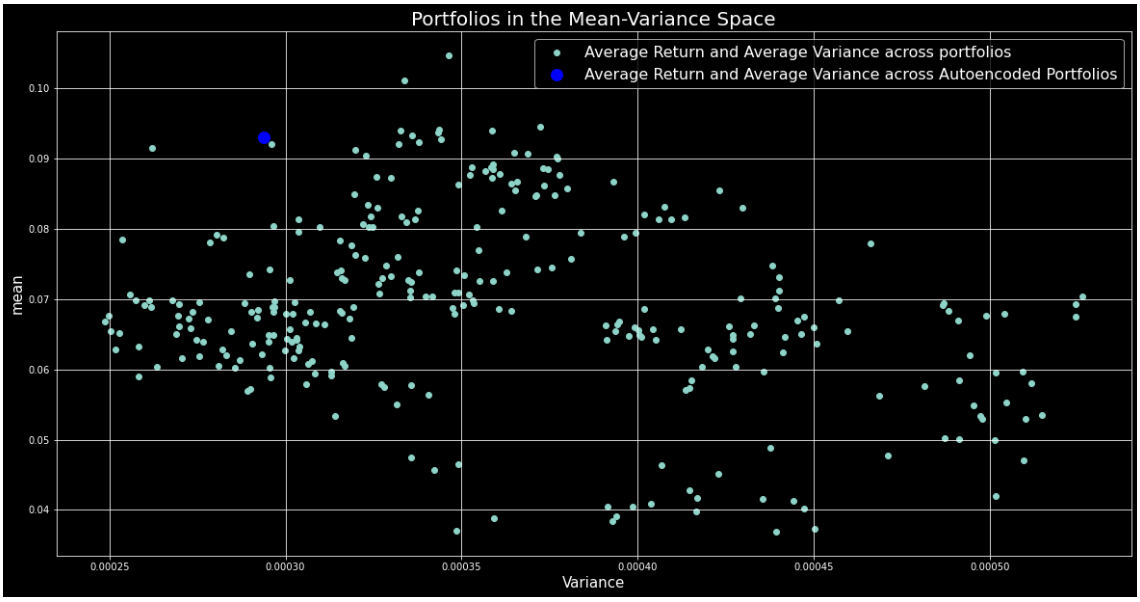

Figure 1.102: Mean-Variance Space: portfolio 40 against competitors.

In the pseudo mean-variance space portfolio 40 is highly competitive in terms of both return and risk.

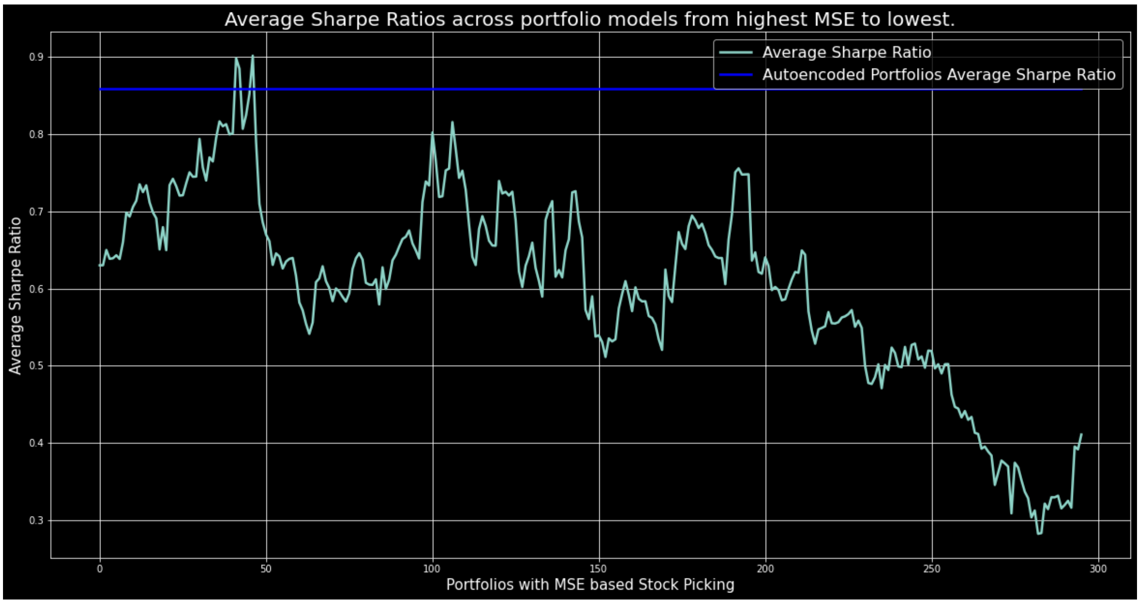

Figure 1.103: Average Sharpe Ratio: portfolio 40 against competitors.

The confirmation comes from the the average Sharpe ratios plots, where the autoencoder-based portfolio outperforms over 98% of the competitors. Also in this case the descending pattern is strongly confirmed.

RMT Covariance Matrix

The portfolios obtained through Random Matrix Theory based covariance estimation make no exception and again portfolio 40 is clearly superior on average in terms of pure performance.
In this case the benchmark is overperformed on 7 out of 10 cases, against the 3 out of 10 of portfolio 335.

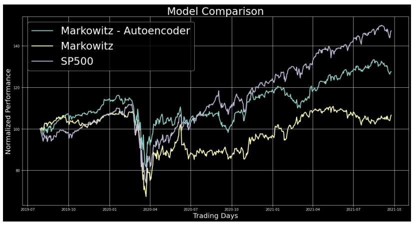

Figure 1.104: Markowitz.

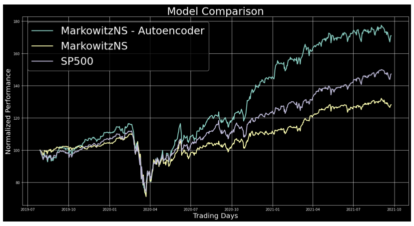

Figure 1.105: Markowitz (NS).

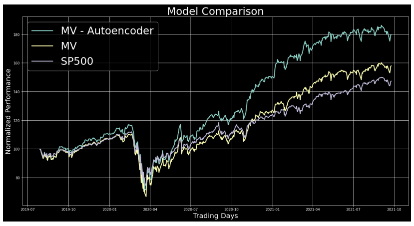

Figure 1.106: Minimum Variance.

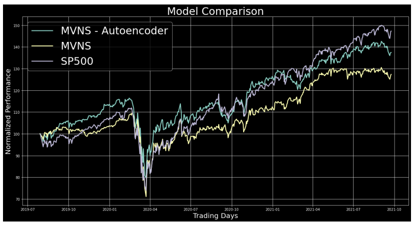

Figure 1.107: Min Variance (NS).

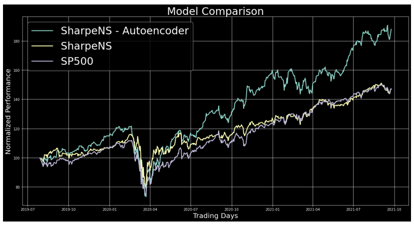

Figure 1.108: Sharpe.

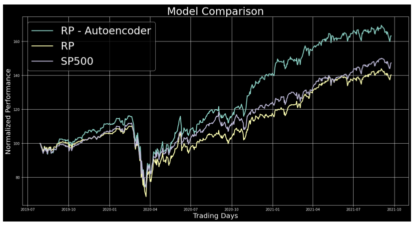

Figure 1.109: Risk Parity.

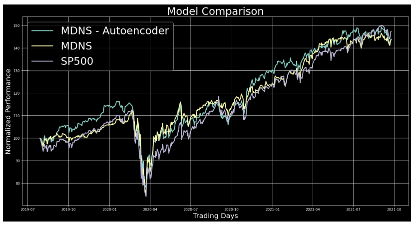

Figure 1.110: MD.

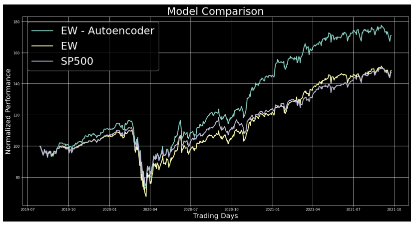

Figure 1.111: Equally Weighted.

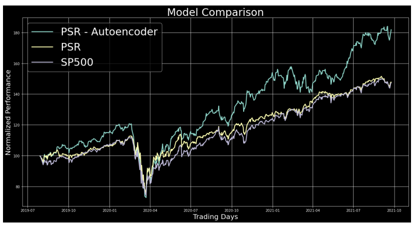

Figure 1.112: PSR.

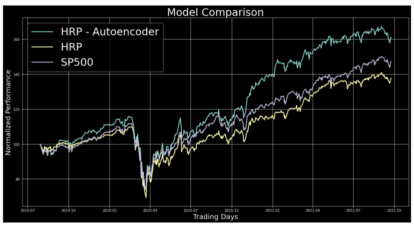

Figure 1.113: HRP.

Talking about the extreme risk measures, portfolio 335 is slightly better for what regards the maximum drawdown, while portfolio 40 presents more convenient average drawdowns. As for Value at Risk portfolio 335 is slightly better, while for the conditional Value at Risk the two portfolios are quite balanced.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.30	-0.045	-0.018	-0.032940	0.0445	0.014	-0.31	9.95
MarkowitzNS	-0.38	-0.040	-0.025	-0.043186	0.0991	0.017	-0.90	13.46
MV	-0.39	-0.042	-0.027	-0.045347	0.1082	0.018	-0.90	13.70
MVNS	-0.32	-0.034	-0.019	-0.034620	0.0581	0.014	-0.63	12.78
SharpeNS	-0.40	-0.043	-0.027	-0.044732	0.1169	0.018	-0.25	14.73
RP	-0.36	-0.038	-0.024	-0.041197	0.0903	0.016	-0.90	13.30
MDNS	-0.33	-0.035	-0.020	-0.036061	0.0672	0.015	-0.68	12.29
EW	-0.38	-0.040	-0.025	-0.043186	0.0991	0.017	-0.90	13.46
PSR	-0.40	-0.044	-0.028	-0.045191	0.1106	0.018	-0.30	14.59
HRP	-0.36	-0.039	-0.024	-0.041314	0.0878	0.016	-0.92	13.94

Table 1.98: Risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	0.520500	0.0	-0.518243	0.763414	1.053666
MarkowitzNS	0.930931	22.49	0.717242	1.257618	1.860189
MV	0.966139	35.97	0.858280	1.311625	1.951513
MVNS	0.658451	0.0	-0.310531	0.908933	1.299892
SharpeNS	1.013090	55.55	0.772592	1.436000	2.082105
RP	0.889744	10.57	0.526049	1.192010	1.760154
MDNS	0.734239	0.02	-0.098782	1.010515	1.435836
EW	0.930931	22.49	0.717242	1.257618	1.860189
PSR	0.958758	32.34	0.672497	1.345111	1.962699
HRP	0.861037	5.67	0.464368	1.149241	1.702705

Table 1.99: Performance indicators adjusted for risk for portfolios 40

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.006283	1.568182	0.675229
MarkowitzNS	1.070709	17.476744	0.921101
MV	1.070709	17.476744	0.921101
MVNS	1.070487	11.952381	0.922936
SharpeNS	1.054655	14.111111	0.933945
RP	1.074657	18.402439	0.924771
MDNS	1.045192	7.918033	0.888073
EW	1.070709	17.476744	0.921101
PSR	1.034983	9.264151	0.902752
HRP	1.074657	18.402439	0.924771

Table 1.100: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.38	-0.087	-0.019	-0.040144	0.0120	0.016	-1.16	12.94
MarkowitzNS	-0.35	-0.041	-0.019	-0.039203	0.0454	0.016	-0.98	17.27
MV	-0.40	-0.045	-0.025	-0.046478	0.0844	0.018	-0.99	14.23
MVNS	-0.35	-0.040	-0.019	-0.039051	0.0443	0.016	-0.80	17.90
SharpeNS	-0.32	-0.030	-0.020	-0.039696	0.0707	0.016	-0.88	17.70
RP	-0.38	-0.044	-0.022	-0.042441	0.0622	0.017	-1.04	16.11
MDNS	-0.34	-0.027	-0.018	-0.036997	0.0666	0.015	-0.68	19.48
EW	-0.39	-0.044	-0.024	-0.044337	0.0728	0.018	-1.03	15.18
PSR	-0.33	-0.029	-0.021	-0.039304	0.0721	0.016	-1.00	18.54
HRP	-0.37	-0.042	-0.022	-0.041456	0.0586	0.016	-1.12	16.45

Table 1.101: Risk indicators and mean returns for portfolios 335

	SR	PSR	ISR	SoR	CR
Markowitz	0.120019	0.00	-0.945205	0.160721	0.223593
MarkowitzNS	0.461001	0.00	-0.663707	0.580239	0.904801
MV	0.727194	0.03	0.374682	0.939592	1.506150
MVNS	0.449648	0.00	-0.600588	0.579713	0.893303
SharpeNS	0.702960	0.02	-0.014318	0.894596	1.547341
RP	0.585347	0.00	-0.284927	0.733663	1.164179
MDNS	0.700994	0.02	-0.114231	0.902457	1.403435
EW	0.657886	0.00	0.047422	0.832494	1.330298
PSR	0.726345	0.07	0.027112	0.910988	1.527840
HRP	0.568319	0.00	-0.402500	0.712209	1.123053

Table 1.102: Performance indicators adjusted for risk for portfolios 335

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.028493	7.634921	0.884404
MarkowitzNS	1.054655	14.111111	0.933945
MV	1.060913	15.153061	0.910092
MVNS	1.054655	14.111111	0.933945
SharpeNS	1.051383	13.315789	0.930275
RP	1.062366	15.500000	0.911927
MDNS	1.054655	14.111111	0.933945
EW	1.062366	15.500000	0.911927
PSR	1.041472	21.666667	0.955963
HRP	1.049880	12.948718	0.928440

Table 1.103: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 335

In terms of Sharpe Ratio portfolio 40 averaged more than 0.25 units higher than its competitor. This dominance is reflected unchanged in terms of PSR. For the Information Sharpe Ratio, the larger portfolio shows results similar to those obtained with previous covariance estimation methods. For portfolio 40, on the other hand, all 7 positive values are also greater than 0.4. Sortino and Calmar Ratio just make the difference even sharper. The rolling analysis strongly confirms the goodness of the autoencoder based portfolio.

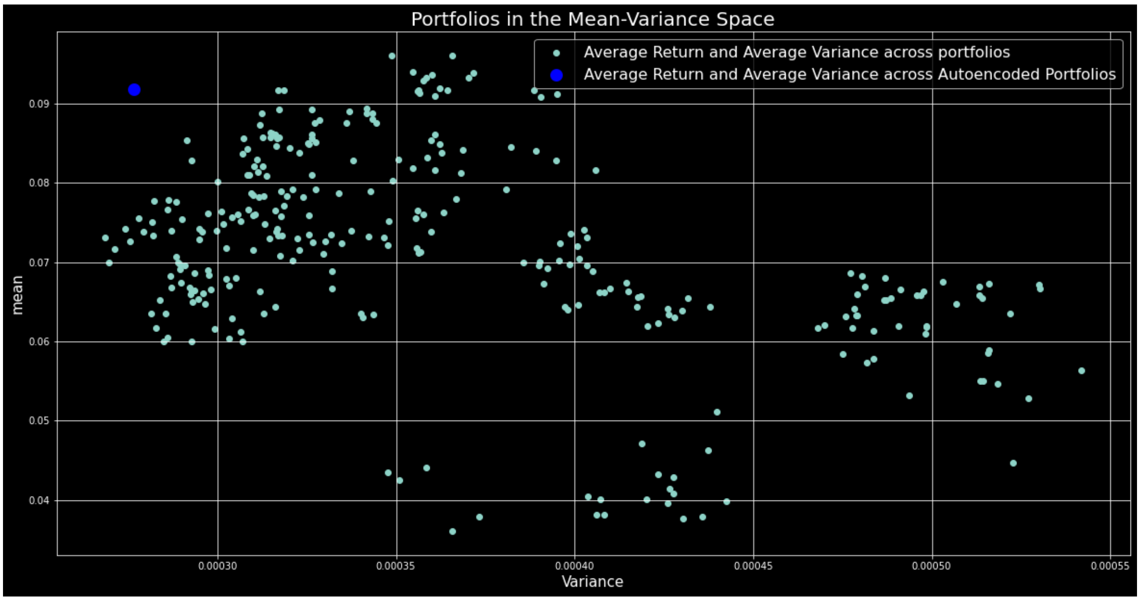

Figure 1.114: Mean-Variance Space: portfolio 40 against its competitors.

The position in the mean-variance space is even better than in the other cases, with portfolio 40 recording one of the highest average returns and one of the lowest average variance.

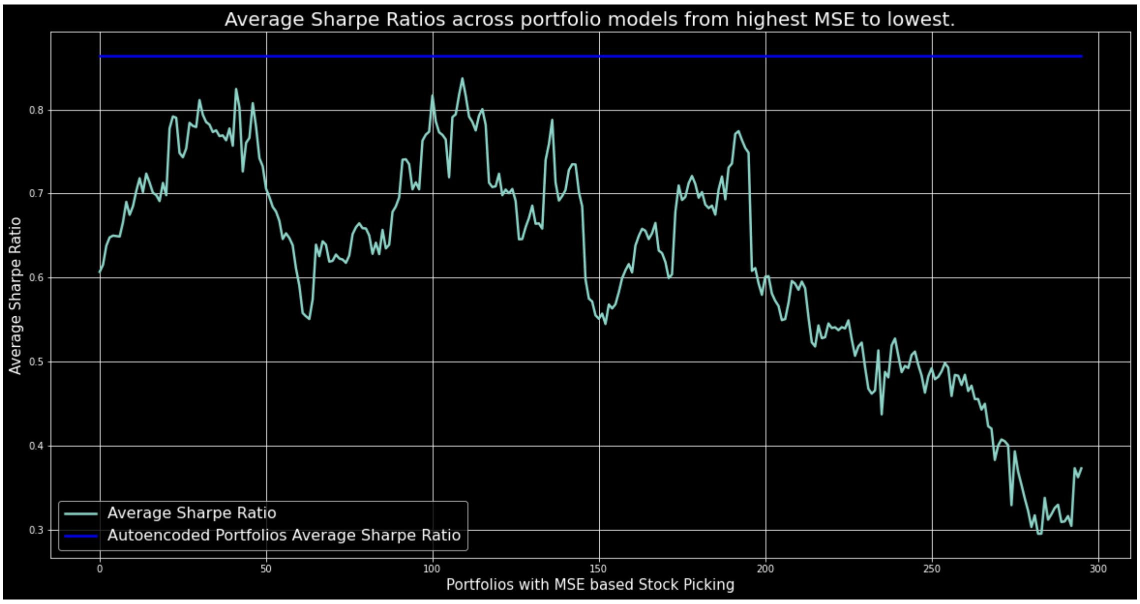

Figure 1.115: Average Sharpe Ratio: portfolio 40 against its competitors.

The average Sharpe ratio confirms the excellent return-risk combination, and portfolio 40 outperforms by a margin all the other portfolios. Also this time the decreasing trend is clearly observed.

1.4.4 Autoencoder implied investment style.

In this section some factors impacted on the autoencoder-based stock-picking. Specifically, first of all is investigated whether the proposed model isolates poorly correlated assets as well as the relationship with β, broken down into its components as on [3]. Secondly, it will be explored whether there are factors related to past performance in the autoencoder cryterium.

Ideally, the Mean Square Error must capture at least part of the information contained in the correlation matrix, possibly enriching it with non-linear components.

Something similar should happen for the β that connects the yield of the individual stocks to that of the reference market: it would be natural that the stock picking naturally exploits the classic β anomaly, but also in this case enriched by non-linear components and implicit factors.

It is interesting to compare the correlation matrix of the 40 stocks selected by the model with the 40 assets that instead have the lowest Mean Square Error. This experiment is run both for train set picking and for validation set picking procedures.

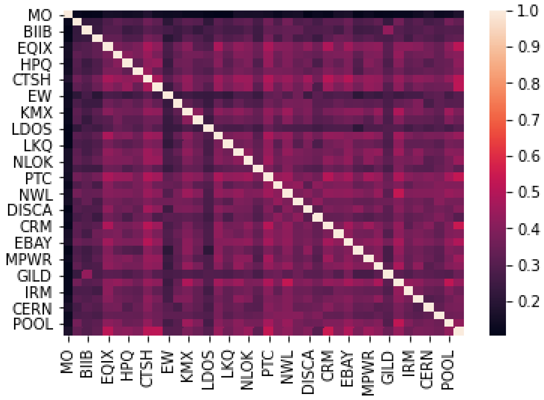

Figure 1.116: Train picking, Train dataset: correlation heatmap of 40 stocks with highest Mean Square Error.

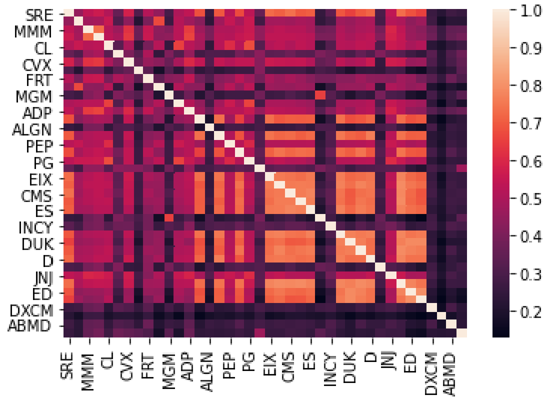

Figure 1.117: Train Picking, Train dataset: correlation heatmap of 40 stocks with lowest Mean Square Error.

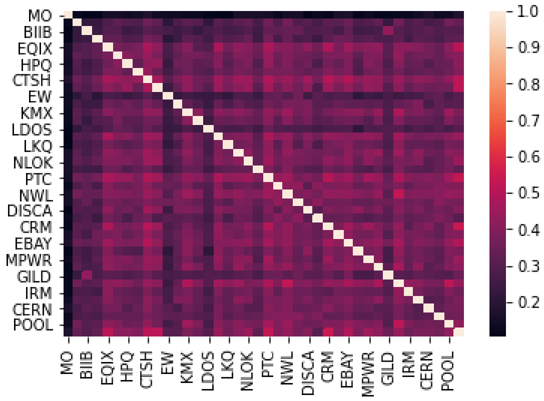

Figure 1.118: Train picking, Validation dataset: correlation heatmap of 40 stocks with highest Mean Square Error.

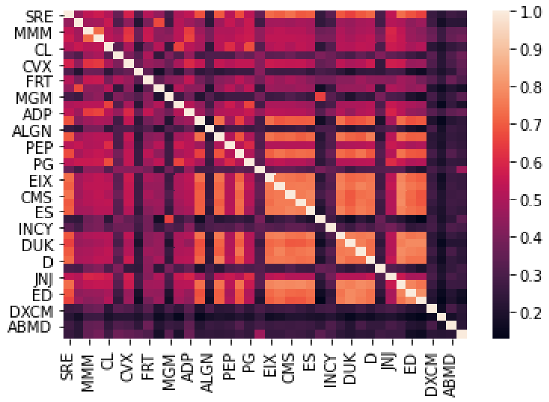

Figure 1.119: Train picking, Validation dataset: correlation heatmap of 40 stocks with lowest Mean Square Error.

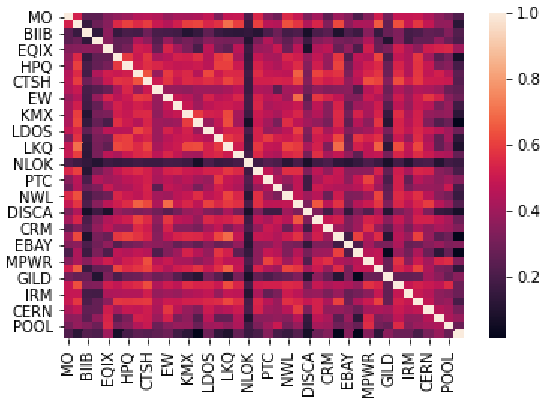

Figure 1.120: Train picking, Test dataset: correlation heatmap of 40 stocks with highest Mean Square Error.

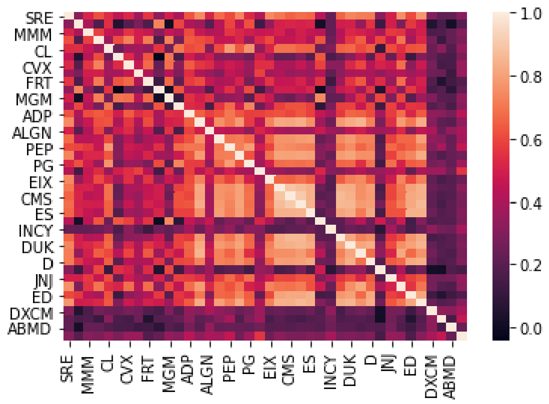

Figure 1.121: Train picking, Test dataset: correlation heatmap of 40 stocks with lowest Mean Square Error.

It clearly emerges that the group of 40 securities that recorded the highest Mean Square Error have on average much weaker correlations against the 40 assets that are in the lowest decile. This behavior remains unchanged over the various periods and datasets, also showing a certain stability about the relationship extracted by the Autoencoder.
It is also worthy of attention how the correlations increase on average a lot for both groups of stocks in the test set, where the pandemic crisis occurred, testifying that during systemic collapses correlations rise strongly by construction.
The experiment replicated for the Validation set picking shows very similar results.

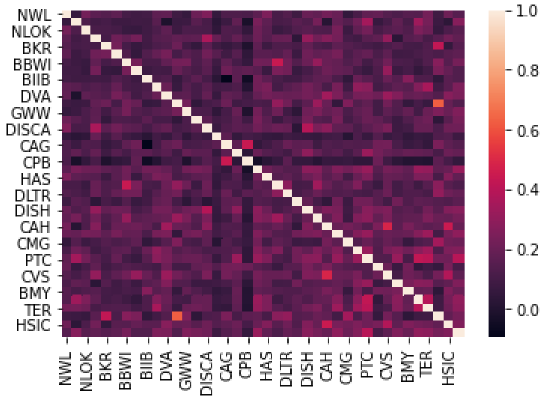

Figure 1.122: Validation picking, Validation dataset: correlation heatmap of 40 stocks with highest Mean Square Error.

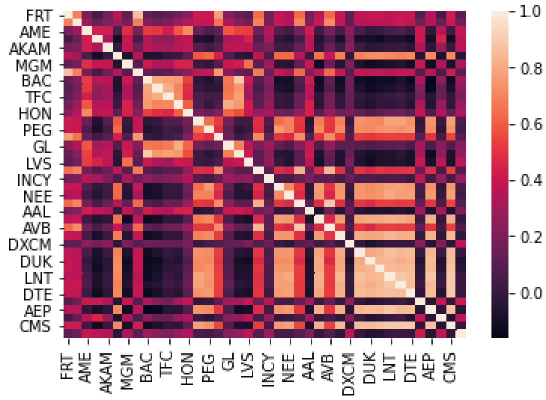

Figure 1.123: Validation picking, Validation dataset: correlation heatmap of 40 stocks with lowest Mean Square Error.

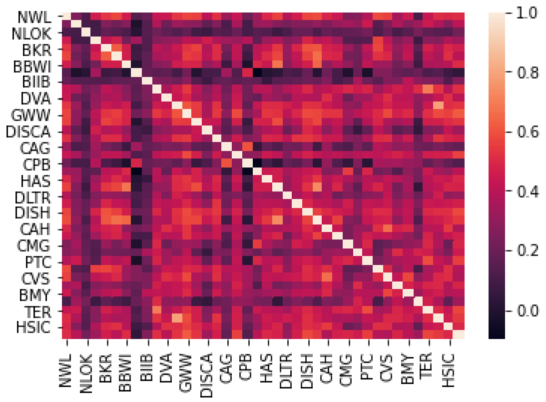

Figure 1.124: Validation picking, Test dataset: correlation heatmap of 40 stocks with highest Mean Square Error.

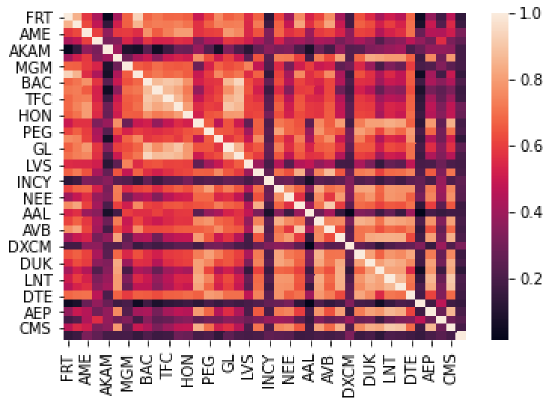

Figure 1.125: Validation, Test dataset: correlation heatmap of 40 stocks with lowest Mean Square Error.

For β and its components as in [3] it has been analyzed the ranking of the stocks selected by the autoencoder among the whole universe of assets.
In particular, stocks have been ranked on the basis of the percentile recorded with respect to the β, the correlation between the yield of the stock and that of the SP500 index ρ_{r,r_m}

and with respect to the ratio between the standard deviation of the single stock and the standard deviation of the SP550 $\frac{\sigma_r}{\sigma_{r_m}}$.
Ideally, we expect the stocks in portfolio 40 to register on average on first half of percentiles for the three measurements. Even in this case, the experiment is repeated for both the stock-picking procedures.
As for the Train set stock-picking, the average ranking and the standard deviation ranking (in parenthesis) obtained by the 40 securities with regard to β is 45%(21%), 60%(14%) and 47%(24%) respectively for Train, Validation and Test datasets.
Compared to ρ_{r,r_m} the average ranking and standard deviations amount to 27%(21%), 47%(24%) and 39%(25%), while in reference to $\frac{\sigma_r}{\sigma_{r_m}}$ values of 60%(14%), 65%(23%) and 56%(26%) were recorded.
For the stock-picking made on the Validation set, we have for β average and standard deviation rankings 50%(24%) and 44%(30%) respectively for the Validation and the Test set. For ρ_{r,r_m} we recorded 24%(14%) and 27%(22%), while for $\frac{\sigma_r}{\sigma_{r_m}}$ we have 86%(9%) and 59%(28%).
Please notice that it has been found a partially different evidence from [3]: the autoencoder choose stocks which are in a low ranking in terms of index correlation and in a high ranking as it regards volatility. It has to be said that the β computation adopted here is much more straightforward than that provided in the aforementioned paper, and that high standard deviations in our basket of secuirities in terms of ranking suggest that the autoencoder relies on this strategy just for a small part.
That said, also some insights on the selection logic implied by the autoencoder have been investigated. For both Train and Validation stock-picking, the rank of the basket of 40 asset has been computed, in both in-sample and out-of-sample perspective.

	mean	std	min	25%	50%	75%	max
maxD	0.425667	0.247777	0.069333	0.201333	0.372000	0.640000	0.893333
meanD	0.401533	0.231107	0.042667	0.190667	0.406667	0.576667	0.917333
VaR	0.453533	0.194261	0.098667	0.308667	0.465333	0.590000	0.968000
cVaR	0.443800	0.152003	0.101333	0.365333	0.446667	0.535333	0.770667
Mean	0.545000	0.301794	0.018667	0.259333	0.597333	0.800667	0.968000
Std	0.603533	0.146420	0.320000	0.492000	0.593333	0.708667	0.938667
Skew	0.407867	0.343937	0.002667	0.086000	0.312000	0.674000	1.000000
Kurt	0.641600	0.279518	0.037333	0.475333	0.641333	0.908667	1.000000
SR	0.503467	0.286112	0.016000	0.273333	0.478667	0.762000	0.986667
ISR	0.525200	0.285062	0.021333	0.278667	0.546667	0.756667	0.968000
SoR	0.501667	0.286010	0.013333	0.272000	0.489333	0.751333	0.978667
CR	0.515733	0.294630	0.018667	0.256000	0.481333	0.809333	0.997333

Table 1.104: Train set stock-picking: characteristics of the 40 stocks during the Train period.

There doesn't seem to be an obvious criterion in the selection made by the autoencoder. In fact, the average ranking across all the measures almost always remains in the range of 40-60, with the sole exception of kurtosis. Furthermore, for all measures with the exception of VaR, cVaR and standard deviation, the volatility of portfolio rankings is close to 30%. Risk-adjusted performance indicators in particular are all anchored around 50%. However, it should be noted that, cVaR aside, there is for each measure at least one stock in the portfolio that reaches 90% ranking, and many are particularly close to 100%.

	mean	std	min	25%	50%	75%	max
maxD	0.461667	0.270094	0.002667	0.278667	0.444000	0.692000	0.901333
meanD	0.450800	0.232135	0.002667	0.278667	0.436000	0.600667	0.938667
VaR	0.433467	0.239153	0.005333	0.199333	0.510667	0.582667	0.893333
cVaR	0.425933	0.255562	0.008000	0.226667	0.417333	0.612667	0.981333
Mean	0.555800	0.282585	0.010667	0.357333	0.574667	0.785333	0.997333
Std	0.569133	0.269313	0.013333	0.362000	0.573333	0.792667	0.997333
Skew	0.429000	0.292450	0.002667	0.209333	0.368000	0.656000	1.000000
Kurt	0.516400	0.291008	0.010667	0.297333	0.530667	0.746667	1.000000
SR	0.532333	0.268799	0.021333	0.317333	0.574667	0.759333	0.960000
ISR	0.548467	0.275269	0.018667	0.327333	0.572000	0.777333	0.978667
SoR	0.529200	0.267625	0.018667	0.322000	0.576000	0.739333	0.941333
CR	0.533067	0.269108	0.024000	0.336667	0.574667	0.740667	0.997333

Table 1.105: Train set stock-picking: characteristics of the 40 stocks during the Test period.

The ranking of the securities selected on the Train set seems to substantially preserve the same properties on the Test set. Notable the reduction in average kurtosis. Therefore, a precise pattern does not seem to emerge as regards the stock picking produced on the Train set.

On the opposite, looking at the selection done on the Validation set, the investment style becomes clearer.

	mean	std	min	25%	50%	75%	max
maxD	0.207800	0.185432	0.002667	0.068667	0.153333	0.343333	0.709333
meanD	0.218133	0.227116	0.002667	0.039333	0.154667	0.272667	0.858667
VaR	0.268200	0.220540	0.013333	0.080667	0.184000	0.412000	0.733333
cVaR	0.166867	0.165031	0.008000	0.065333	0.108000	0.206667	0.669333
Mean	0.331600	0.351003	0.002667	0.044667	0.152000	0.614000	1.000000
Std	0.863267	0.093332	0.661333	0.817333	0.884000	0.930000	0.992000
Skew	0.349267	0.373444	0.002667	0.036000	0.154667	0.634667	1.000000
Kurt	0.798000	0.241497	0.240000	0.730000	0.914667	0.963333	1.000000
SR	0.298600	0.304279	0.002667	0.046667	0.165333	0.492667	0.997333
ISR	0.344667	0.315299	0.010667	0.058000	0.240000	0.592000	0.992000
SoR	0.301267	0.303192	0.002667	0.052000	0.169333	0.482667	0.989333
CR	0.300133	0.310270	0.002667	0.042000	0.164000	0.524667	0.968000

Table 1.106: Validation set stock-picking: characteristics of the 40 stocks.

The autoencoder selected stocks with a very high standard deviation (an average ranking of 86%, with a variability of 9%, very low if compared to the other measures), a very high kurtosis (79%) and pronounced extreme risk measures. On the opposite, the selected assets are on low levels in terms of performance and risk-adjusted performance, being aroung 31% rank.

	mean	std	min	25%	50%	75%	max
maxD	0.465867	0.305050	0.002667	0.182667	0.452000	0.722667	1.000000
meanD	0.444667	0.294621	0.037333	0.158667	0.441333	0.723333	0.933333
VaR	0.414067	0.288154	0.005333	0.172667	0.361333	0.638000	0.997333
cVaR	0.427133	0.292076	0.008000	0.160667	0.396000	0.688667	0.989333
Mean	0.561467	0.313856	0.021333	0.340000	0.574667	0.846000	0.997333
Std	0.593867	0.285591	0.008000	0.359333	0.606667	0.863333	0.997333
Skew	0.529067	0.296112	0.002667	0.285333	0.580000	0.746000	1.000000
Kurt	0.420133	0.324698	0.008000	0.105333	0.374667	0.714667	1.000000
SR	0.541800	0.304162	0.024000	0.317333	0.586667	0.792000	0.968000
ISR	0.572867	0.283611	0.021333	0.350000	0.586667	0.841333	0.944000
SoR	0.549467	0.308981	0.024000	0.314667	0.580000	0.849333	0.968000
Calmar Ratio	0.554600	0.305451	0.032000	0.336667	0.548000	0.831333	0.994667

Table 1.107: Validation set stock-picking: characteristics of the 40 stocks during the Test period.

During the test period, the average rankings settle down to more ambiguous values, and the reduction of kurtosis is particularly striking. At the same time, the variability of the rankings themselves increases, and the portfolio rationale becomes again more difficult to identify.
Noteworthy, however, is the fact that even on the Test set there are no metrics related to individual stocks that justify the excellent portfolio performance achieved. This means that the goodness of the results obtained is to be ascribed not so much to the individual stocks, but to their combination in the portfolio. This suggests that the autoencoder has actually enhanced diversification possibilities.

1.4.5 Other comparisons

In addition to the experiments proposed up to now, other analyzes have been carried out. For sake of space they are not reported but briefly mentioned in this section. All the experiments have been compiled with stock-picking made on the train set (because it is the worse version of the proposed model), and, if not differently specified, across the three covariance matrix estimation techniques, and they are available upon request.

- **Portfolio 40 vs Portfolio 375:** The entire main analysis was reproposed, but contrasting portfolio 40 with portfolio 375, that is by including in the competitor portfolio also the 40 stocks in portfolio 40. The results showed an improvement in portfolio 375 compared to 335, but the dominance of portfolio 40 was nevertheless undisputed. This result also confirms the applicative goodness of the model, since reinserting the 40 securities with the greatest Mean Square Error has benefited the portfolio.

- **Validation and test set unified:**Validation and test datasets have been unified to have a longer and richer test dataset, calculating the weights of the allocation models on the Training set where the Autoencoder model was estimated. Such a choice ignores the fact that the validation dataset was used to ensure the model generalization power, so that is affected by a bias given by the certainty that the Autoencoder adapts sufficiently good to the validation dataset, which now flows into the test one. Anyway, the analysis was performed to provide further experiments possibilities. Portfolio 335 dominates portfolio 40 substantially in all the risk components. It is interesting to observe that in this case, portfolio 40 almost always reacts much worse than its competitor to the pandemic crisis, while in the main analysis the two portfolios were roughly similar in this aspect. Since the stocks in the portfolio are the same as in the train stock-picking case, it can be understood that the reason for this greater sensitivity is to be attributed to very different covariances estimation, performed over an extremely long period of time and with historical method, thus strongly mitigating the effects of a very large number of assets, and making greater use of the data snooping implicit in the choice of the dataset.
  The only parameter against which portfolio 40 models almost always dominates is the expected return. This overperformance is sufficient for the portfolio to record a higher Sharpe Ratio for each model, with the exception of the Maximum Sharpe Ratio, the Maximum Diversification and the Probabilistic Sharpe Ratio allocation models.
  This analysis shows that over a broader period and without portfolio rebalancing, having the entire basket of securities effectively serves to reduce risk, but returns are affected. However, the behavior of the Equally Weight portfolio, where no further parameter estimates are required, is interesting, because it shows how the choice of 40 stocks using the Autoencoder was more effective than the 335 remaining stocks. As for the 40 stocks rolling portfolios, portfolio 40 outperforms a large number of competitors in terms of average expected return, and as many in terms of average variance. It is also significant that no competitor portfolio is able to overperform it at the same time in terms of average risk and return. This result then produces a very competitive average Sharpe Ratio: portfolio 40 record a Sharpe Ratio greater than 96% of all other competing portfolios. The decreasing trend of the average Sharpe Ratio is again verified.

- **Inverse Proportional Beta portfolio:** Portfolio 40 and 335 have been compared against a stratey which allocates to an asset i the weight given by $\frac{1/\beta_i}{\sum_{i=1}^{375} 1/\beta_i}$, i.e. assets with a lower Beta had a proportionately greater weight in the investment. Betas have been estimated on daily log-returns and with respect to the SP500 index. This experiment served to show whether the autoencoder-based model improved performance beyond the exploitation of a standard Beta anomaly strategy. In all asset allocation models, portfolio 40 outperformed the β portfolio. The results of the comparison between the 335 portfolio and the β portfolio instead gave mixed results, with a slight advantage of the β portfolio.

- **Beta stock-picking:** Portfolio 40-based asset allocation models were tested against those obtained on the 40 stocks with lowest Beta, which has been calculated as in the previous experiment. Autoencoder-based portfolio performs better in terms of return than that based on the classic Beta, with the only exception represented by the Markowitz portfolio. The difference becomes enormous especially looking at the case of Markowitz (NS), Minimum Variance, Risk Parity, Equally Weighted and Hierarchical Risk Parity. On the other hand, in most cases - Markowitz, Minimum Variance (NS), Risk Parity, Maximum Diversification, Equally Weighted and Hierarchical Risk Parity - the portfolio based on classic Beta is better able to withstand the impact of the pandemic crisis than its antagonist. With the exception of kurtosis, classic Beta-based portfolios quite clearly dominate their competitors in terms of extreme risk measures. Apart from the Markowitz portfolio case, in terms of Drawdowns (maximum and average), VaR and cVaR, the portfolios based on classic Beta are decidedly more attractive.
  The situation is completely reversed as for performance. With the exception also in this case of the Markowitz model, the Autoencoder-based portfolios are strongly preferable to their competitors. Considering the Sharpe Ratio, on average Autoencoder-based portfolios provide for an incremental 0.18 return for a single unit of risk. The difference is then further increased in all cases by looking at the Probabilistic Sharpe Ratio and the Sortino Ratio. On the other hand, the improvement in Calmar Ratio was slightly milder, thanks to the contained Drawdowns of the classic Beta-based portfolios.

1.5 Analyzes on different datasets

In this section similar analyzes have been collected on other datasets. These ones present different problems which affect the reliability of the results obtained.
Anyway, they were included in the analysis to extend the empirical proposal of the model as far as possible.
The first dataset is again on SP500 universe but data were collected in a previous period. The second one is on the FTSE350 universe, in a period similar to that chosen for the

main analysis.
For each dataset, the criticalities are reported at the opening of the proper section, along with how this can overshadow the reliability of the results obtained.
Theese analyzes are more condensed: pairs comparisons are not included and only the historical covariance matrix results has been included for sake of brevity. Anyway, the other two estimation methods provided very similar results which are available upon request.

1.5.1 SP500: 08/07/1992 to 22/09/2006

The dataset runs from 08/07/1992 to 22/09/2006, where the train dataset records data from the first day until 10/07/2002, the validation dataset from 11/07/2002 to 18/08/2004 and finally the test set from 19/08/2004 until the end of the overall period.
There are only 237 stocks surviving within the index over the period - as well as to date, by construction -, that is, less than half of those required. So the stocks considered are those able to enter the most capitalized index in the world and remain there continuously for about 30 years, given that the list of tickers considered has to be the current one. The data snooping problem is obvious and huge, and from one side it undermines the performance reliability but from the other side makes stock-picking much more difficult, in the sense that it is very difficult to correctly select the best securities in a small subset that already collects by construction the best stocks in the world for the last thirty years.

Stock picking on Train set.

In fact, it can be observed from the performance indicators how from a risk profile portfolio 197 dominates portfolio 40 with respect to all asset allocation models.
In terms of return, the situation is partly reversed in favor of portfolio 40, except for three very notable cases: maximum Sharpe Ratio, Probabilistic Sharpe Ratio and Maximum Diversification. These models built on the 197 residual stocks are in fact the ones that result in the highest possible performance, in particular the Maximum Diversification portfolio which also shows a very good risk profile.
This result should not come as a surprise: knowing a priori that the stocks on which the portfolio has been built will perform well, it makes sense to choose a model that maximizes diversification.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.10	-0.028	-0.013	-0.015839	0.0689	0.007	-0.12	0.53
MarkowitzNS	-0.10	-0.024	-0.011	-0.014847	0.0596	0.008	0.06	0.34
MV	-0.11	-0.028	-0.012	-0.015833	0.0551	0.008	0.07	0.33
MVNS	-0.09	-0.021	-0.012	-0.015793	0.0849	0.008	-0.16	0.51
SharpeNS	-0.17	-0.036	-0.015	-0.018730	0.0891	0.010	0.08	0.68
RP	-0.09	-0.021	-0.010	-0.014046	0.0641	0.007	0.05	0.34
MDNS	-0.08	-0.014	-0.010	-0.013240	0.0821	0.007	0.15	0.60
EW	-0.10	-0.024	-0.011	-0.014847	0.0596	0.008	0.06	0.34
PSR	-0.12	-0.027	-0.012	-0.015441	0.0721	0.008	0.03	0.28
HRP	-0.09	-0.020	-0.010	-0.013491	0.0681	0.007	0.04	0.25

Table 1.108: SP500 92-2006: risk indicators and mean returns for portfolios 40.

	SR	PSR	ISR	SoR	CR
Markowitz	1.467783	100.00	1.142131	2.817996	5.022597
MarkowitzNS	1.262553	100.00	0.931830	2.632235	4.226776
MV	1.098535	99.63	0.789339	2.293688	3.505228
MVNS	1.778693	100.00	1.452687	3.518666	6.821322
SharpeNS	1.477820	100.00	1.221764	3.096844	3.624100
RP	1.430275	100.00	1.079711	2.991970	4.929661
MDNS	1.881277	100.00	1.517560	4.232669	7.379091
EW	1.262553	100.00	0.931830	2.632235	4.226776
PSR	1.466049	100.00	1.148003	3.071426	4.317418
HRP	1.571101	100.00	1.207177	3.365372	5.634195

Table 1.109: SP500 92-2006: performance indicators adjusted for risk for portfolios 40.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.015854	2.584034	0.773585
MarkowitzNS	1.018827	2.778571	0.735849
MV	1.022246	3.165354	0.760377
MVNS	1.032652	8.446429	0.894340
SharpeNS	1.017208	2.598639	0.722642
RP	1.025346	3.521368	0.779245
MDNS	1.037948	6.557143	0.867925
EW	1.018827	2.778571	0.735849
PSR	1.017884	3.371901	0.771698
HRP	1.021242	3.898148	0.796226

Table 1.110: SP500 92-2006: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.06	-0.018	-0.009	-0.012081	0.0431	0.006	0.12	0.41
MarkowitzNS	-0.07	-0.017	-0.010	-0.013760	0.0631	0.007	0.05	0.21
MV	-0.09	-0.021	-0.012	-0.014703	0.0640	0.007	0.06	0.17
MVNS	-0.05	-0.013	-0.010	-0.012517	0.0685	0.006	0.09	0.30
SharpeNS	-0.14	-0.028	-0.016	-0.020532	0.1250	0.011	0.06	0.50
RP	-0.06	-0.014	-0.010	-0.013199	0.0662	0.007	0.04	0.20
MDNS	-0.08	-0.015	-0.013	-0.015934	0.1027	0.008	0.06	0.39
EW	-0.07	-0.017	-0.010	-0.013760	0.0631	0.007	0.05	0.21
PSR	-0.13	-0.027	-0.016	-0.019670	0.1191	0.010	0.07	0.47
HRP	-0.06	-0.013	-0.010	-0.012384	0.0596	0.006	0.06	0.27

Table 1.111: SP500 92-2006: risk indicators and mean returns for portfolios 197.

	SR	PSR	ISR	SoR	CR
Markowitz	1.100893	99.74	0.677138	2.249783	4.960098
MarkowitzNS	1.443325	100.00	1.063785	2.876687	6.152128
MV	1.362011	100.00	1.010214	2.720955	5.176207
MVNS	1.694353	100.00	1.281562	3.313615	9.456874
SharpeNS	1.860806	100.00	1.611583	3.740149	6.296796
RP	1.576464	100.00	1.180167	3.107065	7.285725
MDNS	1.991670	100.00	1.674715	3.966957	9.180567
EW	1.443325	100.00	1.063785	2.876687	6.152128
PSR	1.842203	100.00	1.583481	3.741657	6.389969
HRP	1.503295	100.00	1.083996	2.975798	7.406497

Table 1.112: SP500 92-2006: performance indicators adjusted for risk for portfolios 197.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.022128	4.038095	0.801887
MarkowitzNS	1.035357	4.688172	0.824528
MV	1.037618	4.038095	0.801887
MVNS	1.053319	8.981132	0.900000
SharpeNS	1.028595	3.898148	0.796226
RP	1.033082	5.782051	0.852830
MDNS	1.035639	9.173077	0.901887
EW	1.035357	4.688172	0.824528
PSR	1.028595	3.898148	0.796226
HRP	1.033634	5.870130	0.854717

Table 1.113: SP500 92-2006: bull/bear persistence, null/bear recovery and bull dominance for portfolios 197.

However, the analysis on *rolling portfolios 40* reveals a certain effectiveness of the proposed stock-picking model.

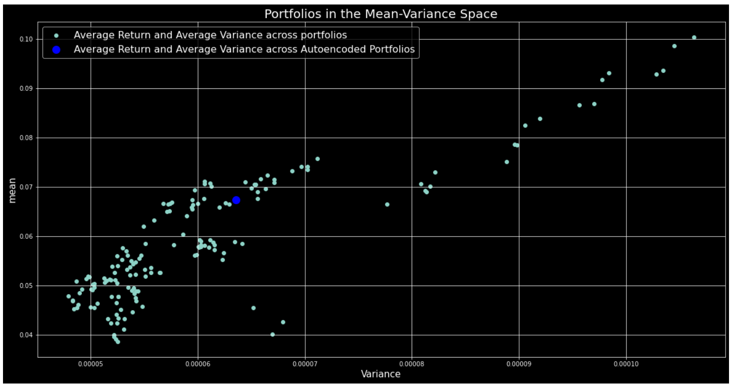

Figure 1.126: SP500 92-2006, Mean-Variance Space: portfolio 40 against competitors.

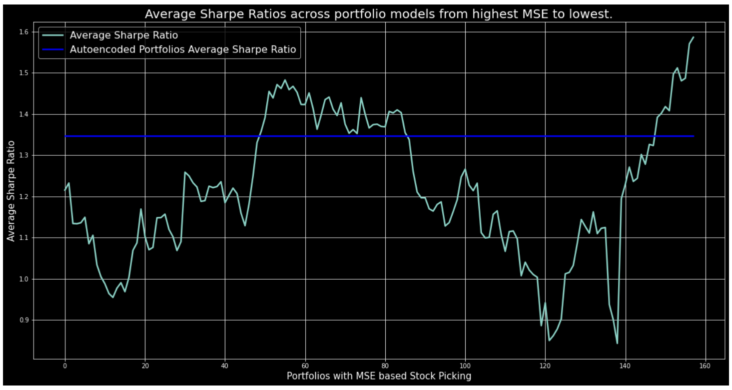

Figure 1.127: SP500 92-2006, Average Sharpe Ratio: portfolio 40 against competitors.

In fact, there are many competitors on which the 40 portfolio dominates in terms of expected return, while maintaining an average competitive risk profile.
In terms of average Sharpe Ratio, the 40 portfolio performs better than 71% of its potential competitors. It should be noted that in this case the downward trend observed so far does not emerge clearly, and indeed the maximum average Sharpe Ratio in absolute is obtained on average on stocks with the lowest MSE. The problem of data snooping may have produced this result.

Stock picking on Validation set.

By operating stock picking on the Validation set, the performance of portfolio 40 improves considerably. In terms of pure returns, portfolio 40 tends to be superior to portfolio 197, with the mild exceptions represented by maximum Sharpe and PSR, and the notable exception of Maximum Diversification, which in the 197 version is the best portfolio in

terms of performance among all those analysed.
That said, please notice how the data snooping mechanism crearly triggers in this example: both the portfolios overperform the benchmark by a large amount. This is obviously due to the fact that by construction the best possible securities have been pre-selected.

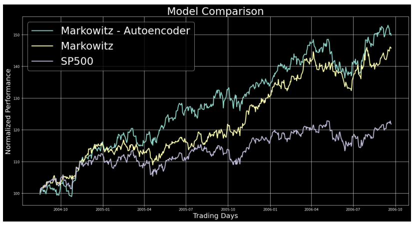

Figure 1.128: Markowitz.

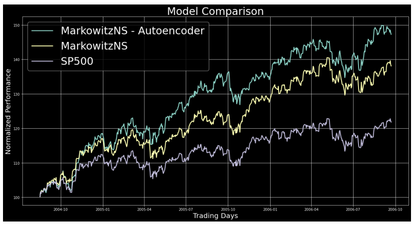

Figure 1.129: Markowitz (NS).

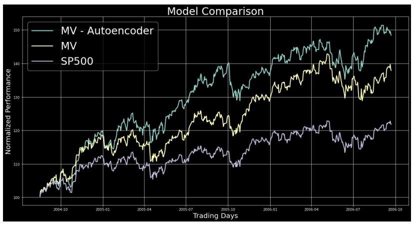

Figure 1.130: Minimum Variance.

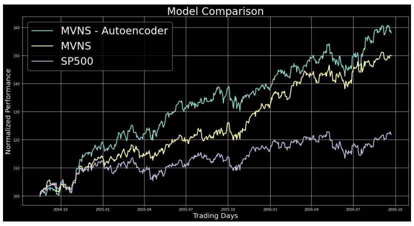

Figure 1.131: Min Variance (NS).

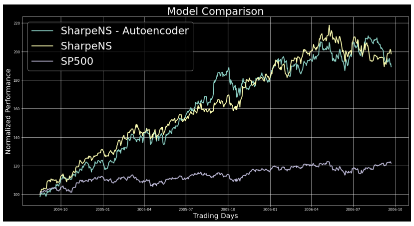

Figure 1.132: Sharpe.

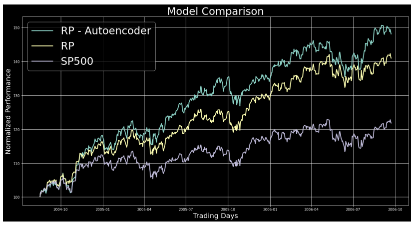

Figure 1.133: Risk Parity.

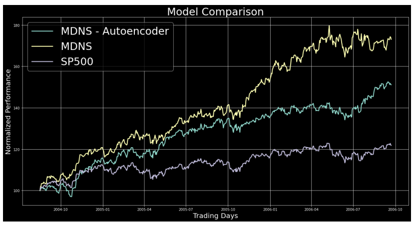

Figure 1.134: MD.

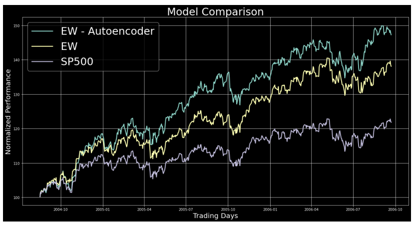

Figure 1.135: Equally Weighted.

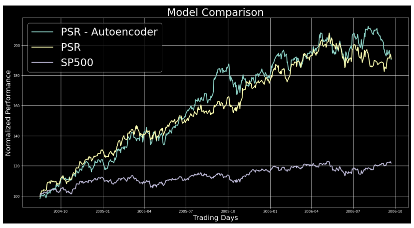

Figure 1.136: PSR.

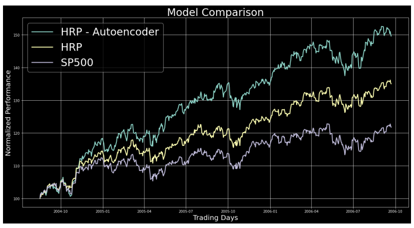

Figure 1.137: HRP.

With respect to the extreme risk measures the two portfolios are pretty aligned, preserving very good and balanced values. On the risk-adjusted performance perspective, portfolio 40 on average does better, with some notable exceptions.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.08	-0.016	-0.011	-0.014717	0.0769	0.007	0.07	0.87
MarkowitzNS	-0.08	-0.017	-0.012	-0.013910	0.0725	0.007	0.10	0.33
MV	-0.08	-0.019	-0.013	-0.014981	0.0738	0.008	0.05	0.33
MVNS	-0.08	-0.014	-0.011	-0.014872	0.0868	0.007	0.08	0.95
SharpeNS	-0.11	-0.025	-0.017	-0.023789	0.1237	0.011	-0.06	0.42
RP	-0.07	-0.016	-0.011	-0.013520	0.0738	0.007	0.11	0.39
MDNS	-0.06	-0.013	-0.011	-0.013933	0.0780	0.007	0.12	0.63
EW	-0.08	-0.017	-0.012	-0.013910	0.0725	0.007	0.10	0.33
PSR	-0.11	-0.025	-0.017	-0.023656	0.1248	0.012	-0.05	0.39
HRP	-0.07	-0.015	-0.011	-0.013758	0.0759	0.007	0.12	0.54

Table 1.114: Risk indicators and mean returns for portfolios 40

	SR	PSR	ISR	SoR	CR
Markowitz	1.638575	99.98	1.373699	3.351668	6.614666
MarkowitzNS	1.576885	99.66	2.066152	3.353291	6.737503
MV	1.535148	86.55	1.860167	3.189944	6.445923
MVNS	1.856229	100.00	2.039027	3.802252	8.157040
SharpeNS	1.708971	100.00	1.653447	3.528423	7.944421
RP	1.636964	100.00	2.181378	3.509638	7.142288
MDNS	1.715270	100.00	1.798833	3.629992	8.785716
EW	1.576885	99.66	2.066152	3.353291	6.737503
PSR	1.721712	100.00	1.678670	3.578090	8.191081
HRP	1.676024	100.00	2.150789	3.567443	7.341950

Table 1.115: Performance indicators adjusted for risk for portfolios 40

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.050446	6.450704	0.866038
MarkowitzNS	1.029389	3.990566	0.800000
MV	1.034334	4.568421	0.820755
MVNS	1.046974	11.902439	0.922642
SharpeNS	1.021836	2.726667	0.771698
RP	1.029389	3.990566	0.800000
MDNS	1.066533	11.022727	0.916981
EW	1.029389	3.990566	0.800000
PSR	1.021836	2.726667	0.771698
HRP	1.039929	5.223529	0.839623

Table 1.116: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.08	-0.018	-0.010	-0.012803	0.0704	0.007	0.08	0.37
MarkowitzNS	-0.08	-0.019	-0.011	-0.013999	0.0604	0.007	0.05	0.23
MV	-0.10	-0.024	-0.011	-0.015052	0.0599	0.008	0.06	0.21
MVNS	-0.06	-0.013	-0.011	-0.012716	0.0759	0.007	0.03	0.21
SharpeNS	-0.13	-0.026	-0.016	-0.020015	0.1299	0.010	0.02	0.37
RP	-0.07	-0.016	-0.011	-0.013302	0.0645	0.007	0.03	0.21
MDNS	-0.08	-0.017	-0.012	-0.016117	0.1033	0.008	-0.02	0.43
EW	-0.08	-0.019	-0.011	-0.013999	0.0604	0.007	0.05	0.23
PSR	-0.12	-0.025	-0.016	-0.019118	0.1226	0.010	0.02	0.25
HRP	-0.06	-0.015	-0.010	-0.012440	0.0566	0.006	0.03	0.26

Table 1.117: Risk indicators and mean returns for portfolios 197

	SR	PSR	ISR	SoR	CR
Markowitz	1.655807	100.0	0.851964	3.565143	5.873334
MarkowitzNS	1.374715	100.0	2.491214	2.798940	5.529950
MV	1.259001	100.0	1.808931	2.598472	4.317272
MVNS	1.836550	100.0	1.582818	3.722501	9.327454
SharpeNS	1.999811	100.0	2.127472	4.170880	6.919232
RP	1.537985	100.0	2.670559	3.125837	6.653851
MDNS	2.025581	100.0	2.158964	4.060154	8.655417
EW	1.374715	100.0	2.491214	2.798940	5.529950
PSR	1.973347	100.0	2.093503	4.137112	7.088051
HRP	1.440157	100.0	1.806861	2.929890	6.624504

Table 1.118: Performance indicators adjusted for risk for portfolios 197

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.019603	3.640351	0.784906
MarkowitzNS	1.021242	3.898148	0.796226
MV	1.024960	2.889706	0.743396
MVNS	1.045460	5.870130	0.854717
SharpeNS	1.029536	3.300813	0.767925
RP	1.022434	4.086538	0.803774
MDNS	1.043970	5.696203	0.850943
EW	1.021242	3.898148	0.796226
PSR	1.029152	3.266129	0.766038
HRP	1.021532	3.943925	0.798113

Table 1.119: Bull/bear persistence, bull/bear recovery and bull dominance for portfolios 197

In terms of Sharpe ratio, portfolio 40 overperforms its competitor in 6 cases out of 10, but 2 of the 4 remaining models - Sharpe and MD - are also the best ones overall in terms of Sharpe ratio. The situation is basically the same across the other risk-adjusted performance indicators.
Rolling 40 portfolios analysis furtherly proves the goodness of our approach.

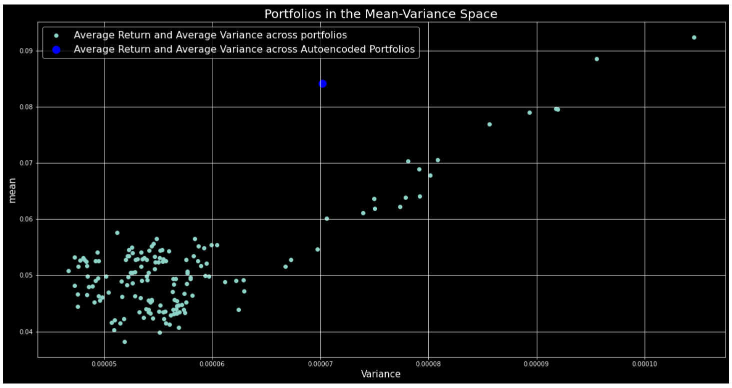

Figure 1.138: SP500 92-2006, Mean-Variance Space: portfolio 40 against competitors.

In the pseudo mean-variance space, the autoencoder portfolio has an incredibly good positioning in terms of return but low-medium one in terms of variance. Despite that, as for average Sharpe ratio, the portfolio performs very good, exceeding the 100% of competing portfolios.

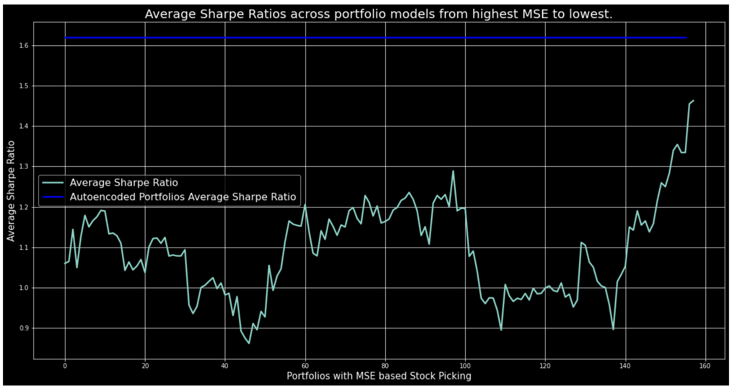

Figure 1.139: SP500 92-2006, Average Sharpe Ratio: portfolio 40 against competitors.

1.5.2 FTSE350: 27/08/2009 to 08/12/2021

This further analysis was carried out on the securities of the FTSE350, in the period from 27/08/2009 to 08/12/2021. Of the total of 350 stocks that currently make up the index, only 185 were continuously listed over the entire period. This leads to another strong data snooping problem. Furthermore, it has been chosen to not include the subprime mortgage period because the number of assets would have dropped below a hundred in that case. That said, the train dataset runs from the beginning of the period to 29/05/2009, the validation set from 30/05/2019 to 02/09/2020, finally the test set from 03/09/2020 until the end of the period.

Stock picking on Train set

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.13	-0.040	-0.013	-0.018	0.0176	0.008	0.01	0.90
MarkowitzNS	-0.12	-0.023	-0.013	-0.020	0.1056	0.009	-0.02	3.47
MV	-0.12	-0.025	-0.014	-0.022	0.1258	0.011312	0.12	5.26
MVNS	-0.15	-0.03	-0.013	-0.016	0.0339	0.007	-0.07	0.60
SharpeNS	-0.22	-0.049	-0.020	-0.026	0.0251	0.013	0.26	0.85
RP	-0.12	-0.022	-0.012	-0.018	0.0892	0.008	-0.10	2.05
MDNS	-0.14	-0.024	-0.015	-0.017	0.0773	0.008	-0.13	0.39
EW	-0.12	-0.02	-0.013	-0.020	0.1056	0.009	-0.026	3.47
PSR	-0.19	-0.038	-0.016	-0.021	0.0455	0.011	0.17	0.47
HRP	-0.12	-0.022	-0.012	-0.017	0.0745	0.008	-0.131	1.23

Table 1.120: FTSE350: risk indicators and mean returns for portfolios 40.

	SR	PSR	ISR	SoR	CR
Markowitz	0.311395	0.00	0.108836	0.659431	0.908293
MarkowitzNS	1.698178	98.26	1.503859	3.377253	6.119748
MV	1.767603	99.10	1.596817	3.499250	7.287153
MVNS	0.689973	0.00	0.454096	1.405436	1.587445
SharpeNS	0.301862	0.00	0.164205	0.674175	0.790426
RP	1.599462	90.53	1.384187	3.216284	5.173521
MDNS	1.454447	19.11	1.233701	2.993505	3.756131
EW	1.698178	98.26	1.503859	3.377253	6.119748
PSR	0.634420	0.00	0.472901	1.439142	1.640240
HRP	1.418669	10.10	1.191672	2.898762	4.239710

Table 1.121: FTSE350: performance indicators adjusted for risk for portfolios 40.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	0.997366	0.887755	0.540373
MarkowitzNS	1.032921	3.422222	0.720497
MV	1.023405	2.593272	0.661491
MVNS	1.014292	1.697479	0.630435
SharpeNS	1.020071	1.587010	0.577640
RP	1.029042	3.079038	0.698758
MDNS	1.018251	3.168831	0.760870
EW	1.032921	3.422222	0.720497
PSR	1.020408	1.960000	0.611801
HRP	1.029042	3.079038	0.698758

Table 1.122: FTSE350: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.11	-0.031	-0.009	-0.013776	0.0223	0.007	-0.06	2.55
MarkowitzNS	-0.06	-0.015	-0.011	-0.017019	0.0766	0.008	0.02	4.04
MV	-0.08	-0.021	-0.015	-0.023850	0.1094	0.012	1.06	11.40
MVNS	-0.04	-0.011	-0.008	-0.011219	0.0420	0.005	-0.27	3.06
SharpeNS	-0.16	-0.066	-0.022	-0.035011	-0.0148	0.013	-1.01	3.04
RP	-0.06	-0.015	-0.012	-0.018246	0.0837	0.009	0.29	5.90
MDNS	-0.09	-0.024	-0.013	-0.017339	0.0605	0.009	0.15	0.92
EW	-0.06	-0.017	-0.013	-0.020800	0.0965	0.010	0.66	8.55
PSR	-0.12	-0.029	-0.013	-0.021710	0.0293	0.009	-0.84	3.61
HRP	-0.06	-0.014	-0.011	-0.016460	0.0770	0.008	0.05	4.25

Table 1.123: FTSE350: risk indicators and mean returns for portfolios 135.

	SR	PSR	ISR	SoR	CR
Markowitz	0.540759	0.00	0.226550	1.124112	1.453398
MarkowitzNS	1.552997	71.47	1.275733	2.959242	8.511888
MV	1.463482	38.52	1.282022	3.077307	9.567263
MVNS	1.292679	0.63	0.877043	2.456068	7.254291
SharpeNS	-0.180690	0.00	-0.335513	-0.286902	-0.642646
RP	1.552725	69.33	1.298750	3.011718	9.549168
MDNS	1.107067	0.00	0.862019	2.487509	4.886124
EW	1.525709	58.71	1.310291	3.069603	10.583928
PSR	0.505914	0.00	0.283848	0.869902	1.708347
HRP	1.627650	90.38	1.336746	3.072718	9.318329

Table 1.124: FTSE350: performance indicators adjusted for risk for portfolios 135.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.010386	1.914894	0.562112
MarkowitzNS	1.057296	3.041199	0.723602
MV	1.030688	2.208850	0.649068
MVNS	1.013896	1.675000	0.627329
SharpeNS	0.989564	0.725806	0.419255
RP	1.053154	3.331765	0.736025
MDNS	1.004710	1.276596	0.559006
EW	1.032629	2.301818	0.658385
PSR	1.007603	1.213793	0.549689
HRP	1.051901	2.817376	0.708075

Table 1.125: FTSE350: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 135.

Portfolio 135 beats Portfolio 40 in terms of extreme risk, i.e. with respect to drawdowns and Value at Risks. As regards the standard risk represented by the standard deviation, the comparison is balanced, with a slight dominance of portfolio 40. Then portfolio 40 outperforms its competitor with respect to the return. This dominance is reflected in the Sharpe Ratio, where with the exception of the Markowitz model and Minimum Variance (NS), the stock-picked portfolio shows increasingly better results. The predominance in this sense is confirmed by the Probabilistic Sharpe Ratio, with respect to which portfolio 40 records significantly higher values. The differences are then further amplified if looking at the Sortino Ratio.

The performance of the Equally Weight model is always interesting, particularly useful in evaluating pure stock picking, and it reveals a dominance of portfolio 40.

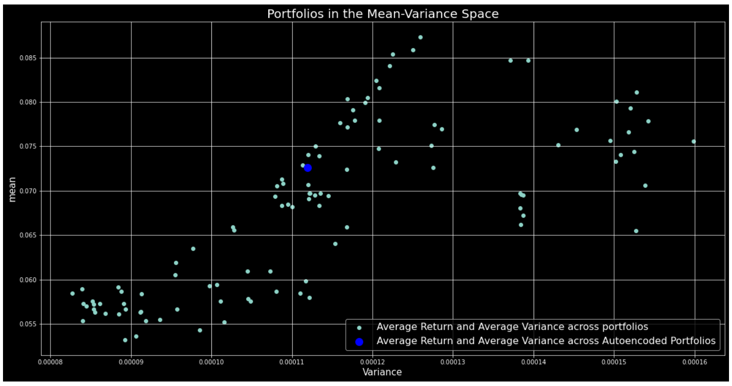

Figure 1.140: FTSE350, Mean-Variance space: portfolio 40 against competitors.

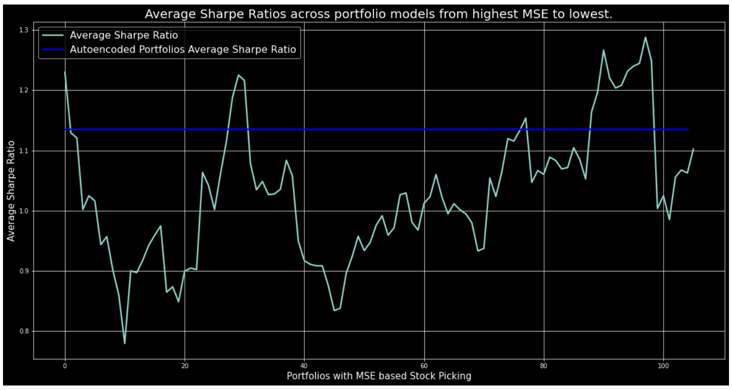

Figure 1.141: FTSE350, Average Sharpe Ratio: portfolio 40 against competitors.

As for the *rolling portfolios* analysis, portfolio 40 is better than half of its competitors in terms of return and outperforms another half in terms of risk.

Its average Sharpe Ratio is then better than 85% of the othe 40 assets portfolios.

Even in this case, there is no clear downtrend in the average Sharpe ratios with respect to the decreasing Mean Square Error. Again, the dataset reduced from 350 to 185 total assets may have affected the results.

Stock picking on Validation set

Also in this case, performing the stock selection procedure on the validation setstrongly improves portfolio 40, as it can be seen immediately from the equity lines. In fact, in terms of pure performance, the different among the two grops of portfolios is huge.

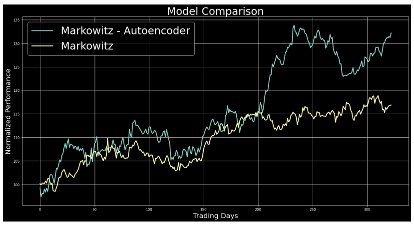

Figure 1.142: Markowitz.

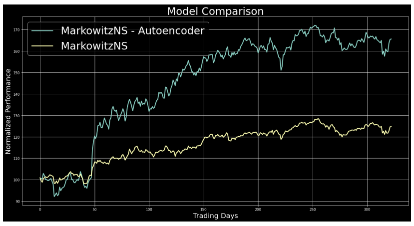

Figure 1.143: Markowitz (NS).

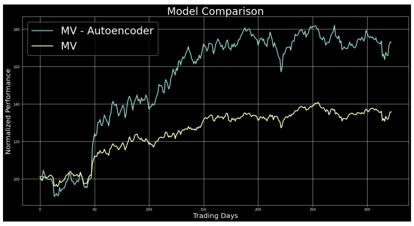

Figure 1.144: Minimum Variance.

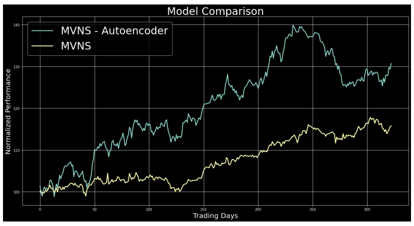

Figure 1.145: Min Variance (NS).

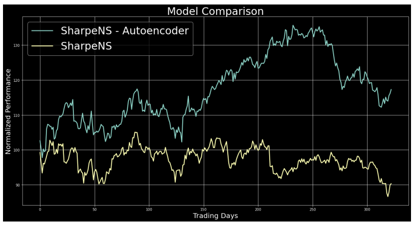

Figure 1.146: Sharpe.

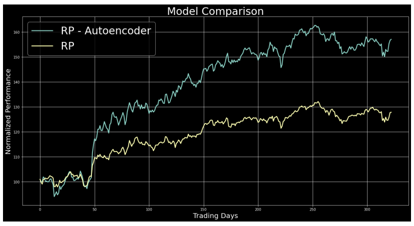

Figure 1.147: Risk Parity.

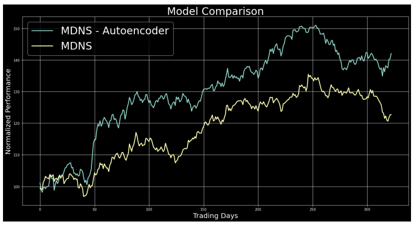

Figure 1.148: MD.

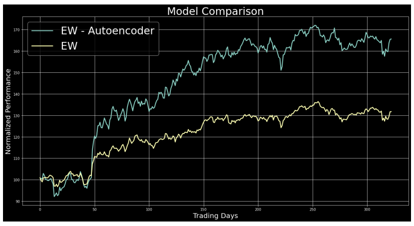

Figure 1.149: Equally Weighted.

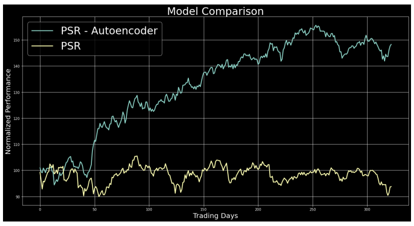

Figure 1.150: PSR.

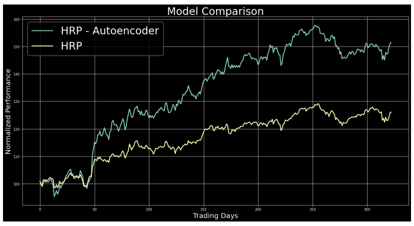

Figure 1.151: HRP.

From the other side, portfolio 135 provides a more controlled extreme risk profile, dominating portfolios 40 for most of the asset allocation models. But putting together risk and return, the results are again extremely skewed in favor of portfolio 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.08	-0.026	-0.013	-0.018165	0.0890	0.009	-0.03	1.41
MarkowitzNS	-0.10	-0.025	-0.018	-0.026769	0.1538	0.014	1.52	14.78
MV	-0.13	-0.033	-0.022	-0.032290	0.1668	0.018	1.82	17.29
MVNS	-0.11	-0.027	-0.014	-0.017103	0.0789	0.008	-0.05	0.37
SharpeNS	-0.17	-0.048	-0.023	-0.030119	0.0415	0.013	0.09	3.78
RP	-0.08	-0.021	-0.015	-0.022354	0.1378	0.012	1.09	10.73
MDNS	-0.11	-0.024	-0.014	-0.018270	0.1057	0.010	0.11	1.37
EW	-0.10	-0.025	-0.018	-0.026769	0.1538	0.014	1.52	14.78
PSR	-0.09	-0.021	-0.016	-0.021641	0.1193	0.011	-0.09	2.26
HRP	-0.08	-0.021	-0.015	-0.020254	0.1258	0.010	0.73	7.14

Table 1.126: FTSE350: risk indicators and mean returns for portfolios 40.

	SR	PSR	ISR	SoR	CR
Markowitz	1.610935	95.84	-	3.412764	7.714492
MarkowitzNS	1.697368	88.95	-	3.916708	10.357184
MV	1.514414	53.73	-	3.588232	8.817139
MVNS	1.506986	55.86	-	3.296977	5.271374
SharpeNS	0.500060	0.00	-	0.967146	1.691766
RP	1.854617	99.04	-	4.192744	12.542093
MDNS	1.767909	100.00	-	3.982542	7.009792
EW	1.697368	88.95	-	3.916708	10.357184
PSR	1.793635	99.98	-	3.772392	9.563285
HRP	1.922756	99.95	-	4.296215	11.088478

Table 1.127: FTSE350: performance indicators adjusted for risk for portfolios 40.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.019881	2.028302	0.670807
MarkowitzNS	1.031195	3.268817	0.711180
MV	1.028528	3.034014	0.695652
MVNS	1.039000	3.111413	0.714286
SharpeNS	1.008224	1.445887	0.521739
RP	1.031195	3.268817	0.711180
MDNS	1.017794	1.812977	0.593168
EW	1.031195	3.268817	0.711180
PSR	1.044654	2.897872	0.708075
HRP	1.026559	2.862745	0.683230

Table 1.128: FTSE350: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 40.

	maxD	meanD	VaR	cVaR	mean	std	skew	kurtosis
Markowitz	-0.06	-0.017	-0.009	-0.013398	0.0480	0.006	0.02	2.36
MarkowitzNS	-0.07	-0.016	-0.011	-0.016323	0.0660	0.007	-0.16	2.64
MV	-0.07	-0.020	-0.014	-0.021143	0.0920	0.010	0.21	5.49
MVNS	-0.04	-0.011	-0.008	-0.011195	0.0449	0.005	-0.24	3.17
SharpeNS	-0.18	-0.068	-0.020	-0.033776	-0.0293	0.013	-0.69	2.63
RP	-0.07	-0.017	-0.011	-0.017377	0.0733	0.008	-0.08	3.48
MDNS	-0.11	-0.024	-0.012	-0.016761	0.0653	0.009	0.16	0.71
EW	-0.07	-0.018	-0.012	-0.019252	0.0827	0.009	0.07	4.56
PSR	-0.14	-0.060	-0.020	-0.033751	-0.0175	0.013	-0.69	2.16
HRP	-0.06	-0.015	-0.010	-0.015837	0.0692	0.007	-0.19	2.68

Table 1.129: FTSE350: risk indicators and mean returns for portfolios 135.

	SR	PSR	ISR	SoR	CR
Markowitz	1.189279	0.00	-	2.496175	5.392221
MarkowitzNS	1.400349	10.52	-	2.732904	6.949571
MV	1.469170	37.71	-	2.955808	9.094857
MVNS	1.379201	7.93	-	2.684742	7.340206
SharpeNS	-0.347151	0.00	-	-0.624375	-1.176076
RP	1.454195	29.90	-	2.855941	7.745495
MDNS	1.206689	0.00	-	2.899220	4.212077
EW	1.465946	35.80	-	2.909125	8.410411
PSR	-0.210459	0.00	-	-0.378057	-0.864151
HRP	1.525144	61.76	-	2.956681	7.874297

Table 1.130: FTSE350: performance indicators adjusted for risk for portfolios 135.

	BullToBull/BearToBear	BearToBull/BullToBear	Bull Periods
Markowitz	1.028172	2.178218	0.683230
MarkowitzNS	1.059615	3.137931	0.729814
MV	1.032629	2.301818	0.658385
MVNS	1.035137	2.527473	0.717391
SharpeNS	1.007310	1.188679	0.506211
RP	1.051901	2.817376	0.708075
MDNS	1.000915	1.059140	0.611801
EW	1.035349	2.433962	0.670807
PSR	1.008908	1.233974	0.515528
HRP	1.059615	3.137931	0.729814

Table 1.131: FTSE350: bull/bear persistence, bull/bear recovery and bull dominance for portfolios 135.

Indeed, portfolio 40 dominates portfolio 135 by a large average amount across all risk-adjusted performance measures and with respect to all the portfolio models which have been implemented. For example, as regarding the Sharpe ratios, portfolios 40 models overperforms on average its competitor by more than 0.50 units of return for one unit of risk. The PSR, which is this time computed with respect to the 1.5 treshold, underlines the dominance of portfolio 40. The ISR has not been computed because our data provider incredibly does not have the historical data for the FTSE350 index.
As for the rolling portfolios analysis, the model performs not bad. The positioning in the mean-space is very extreme: portfolio 40 has the highest return but also the highest variance, and in both cases by a large margin.

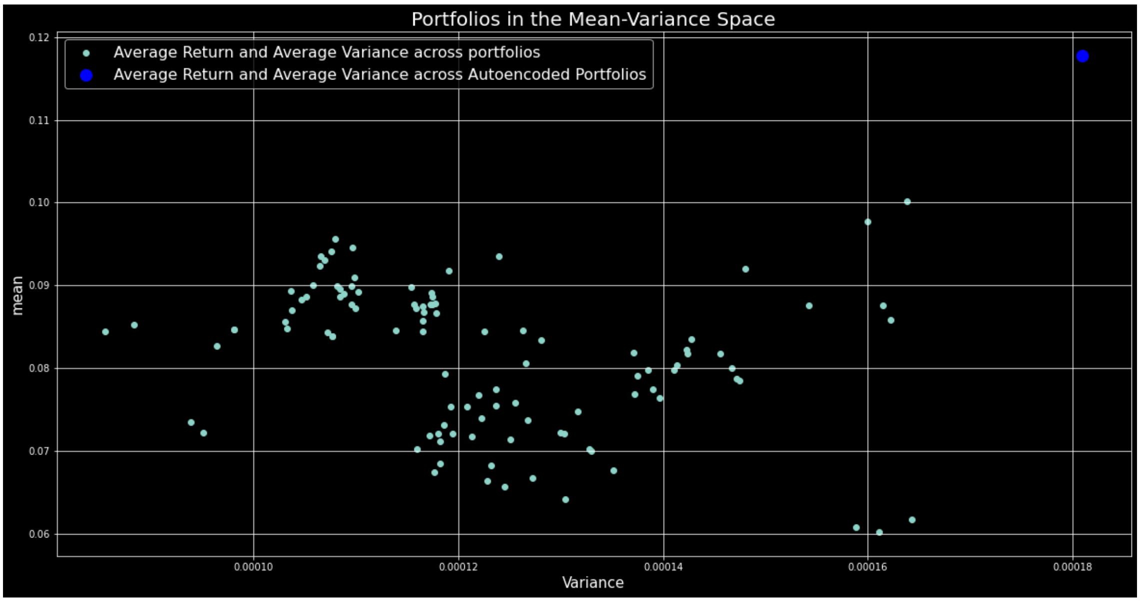

Figure 1.152: FTSE350, Mean-Variance Space: portfolio 40 against competitors.

In the compromise among performance and risk, the overall positioning is very good. Indeed, in terms of average Sharpe Ratio, portfolio 40 outperforms 93% of its competitors.

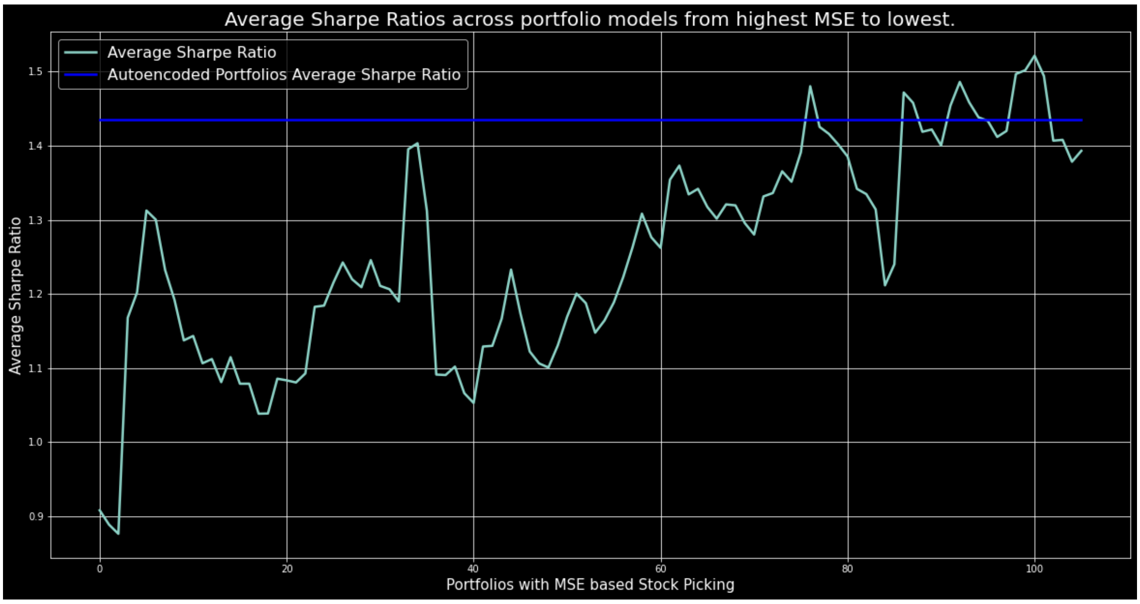

Figure 1.153: FTSE350, Average Sharpe Ratio: portfolio 40 against competitors.

1.6 Conclusions

In this paper the *curse of dimensionality* problem have been adressed.
The proposed solution seeks to select a small basket of assets from an investment universe, such those have two characteristics: they must be poorly correlated with each other and poorly correlated with the market. In this way, true diversification can be achieved and a generalized Beta anomaly strategy can be exploited.
For this purpose, an Autoencoding neural network was produced, modified to leave room for features that capture the momentum of the individual assets. The Autoencoder is therefore configured as a market model, which, starting from a reduced number of latent factors, tries to reconstruct the series of returns of the individual assets. The less accurately reconstructed stocks are by construction poorly correlated to each other and poorly correlated to the market as a whole, where the *correlation* concept is to be understood in this case in a more general sense, given the non-linearity of the Autoencoder.
To confirm the validity of the intuitions, various empirical analyzes and backtests have been proposed, supported by performance indicators, applying several asset allocation models crossed with various methods of estimating the covariance matrix. In almost all the many experiments proposed, it was shown how the portfolio built on the securities selected by the Autoencoder has strongly overperformed its competitors, and therefore the benefits provided by the model are real and none of the asset allocation models or covariance estimation methods are able to absorb them.
The numerous empirical tests brought in favor of the goodness of the model are however partially affected by some problems in the datasets, in particular by data snooping

mechanics and survivor biases. The proposed framework essentially looks for deviant behavior in a universe of securities, given the hypotheses of the experiment, i.e. the chosen model, network structure, input data, etc. At this point, a doubt can be raised about the reliability of the results, based on the following reasoning. Deviant behavior with respect to the market can be essentially of two types: virtuous, whereby the securities tend to do much better than the reference market, or negative, whereby the securities tend instead to do strongly worse. Knowing that securities in the datasets have been listed in the SP500 - for instance - continuously in a certain range, it is clear that part of the problematic deviant behaviors are discarded a priori, because securities that have performed too badly over time have evidently left the SP500. Then the residual deviant behaviors are necessarily virtuous.

To mitigate in part this legit critique, the stocks selected through the neural network have been selected in order to find some patterns in the selection logic. At this prpose, the average ranking of the basket of 40 stocks was analyzed across several performance measures. At the same time the relationships between the generalized correlations and Betas implied in the autoencoding neural network and the classic linear correlation and CAPM Beta have been empirically investigated, proving how the model implicitly captures a large part of them, but enriched with non-linear and complex relationships that allow for an additional investment edge. It has been found that the model tended to select for high risk stocks, and indeed this is somewhat understandable as we are effectively looking for outliers. However, it is also true that these stocks do not stand out with respect to any of the performance metrics in the strict sense, and that they show very low rankings in terms of correlation with the reference index. This suggests that the model did not necessarily select the best stocks, but the best stocks to fit together in a portfolio, thus suggesting that the outcome is affected by data snooping and survivor biases more tangentially than one might think.

Chapter 2

Invoices default forecasting for credit factoring.

In the present work we propose a machine learning based model in order to forecast invoices default. The context is the business to business Italian market, and the dataset has been provided by Credit Service, an Italian company whose business is based on and invoice trading. Given the commitment of the company, our model performs very well, and largerly outperforms the benchmark model used by the company and several credit institutions in Italy, allowing then for arbitrage opportunities.

2.1 Invoices default prediction.

Credit Service is an Italian company whose business is focused on invoice trading. During the last years of activity, the company collected a very large dataset invoices between SMEs in the Italian context. Credit Service relies on the credit rating calculated by Mode Finance for its activities. This represents a standard for the pricing policies in credit factoring and invoice trading of most of the Italian credit institutions. Mode Finance's company rating is used as a proxy in order to measure the credit risk for single invoices as well.

In literature, a lot of Machine Learning and Deep Learning prediction models and feature selection techniques have been employed for companies or individuals credit scoring.

Tslking about data filtering and selection, among the others deserve attention the F-Score which measures some kind of distance among two sets of real numbers [?], [110]; rough set theory based measures [95], [119], [123].

In the class of so called *wrapper methods*, we have for example the stepwise feature selection which is composed from two steps: forward feature selection and backward feature elimination and is based on linear regression and the use of the $p - value$ as filter mea-

sure [108], [116]; then a more moderne class of wrapper methods is given by the genetic algorithms [88], [70], [37], [76].
Then, in the set of embedding methods, we find penalized regression based models such as the LASSO method [114].
Other more invasive methods work on the provided variables to form new features. These ones are the feature engineering techniques. They are really powerful, but at the same time some loss of interpretability is produced. The most famous one is the PCA, which linearly combines initial variables in new orthogonal features ranked according to a degree of information explained by each feature [105]; then we have Autoencoders derived from neural networks structures, which basically respects the PCA idea but also accounts for any kind of non linearity and Linear Discriminant Analysis introduced by Fisher [50], [**?**].
Talking instead of prediction models, a large plethora of ML and DL techniques have been proposed. Starting from logistic regression, LDA [97], the kNN classifier [63] and the naive Bayes classifier [59] to Deep Learning models as Artificial Neural Networks [112], [89], [16], [116], [117], [91], passing through Support Vector Machines [35] and Trees based models, which we will see in details.
That said, nothing standing to our knowledge has been proposed in order to predict invoice payment default in a Business to Business context.
Credit Service's idea is in fact to exploit their internal database in order to gain better insight into the credit rating of single invoices. Indeed, a more accurate result could lead to statistical arbitrage opportunities by pricing invoices based on Mode Finance rating while composing a more competitive portfolio by exploiting more accurately estimated probabilities.
In this document we propose a simple Random Forest approach which provides good results. This work is organized as follows: in section 2 the dataset is presented and accurately analyzed, in order to capture any apparent relation (e.g. the pandemic impact); in section 3 the preprocessing steps are detailed; section 4 shows the model construction, estimation and fine tuning; finally, section 5 presents the results, which are accurately tested in terms of robustness.
Please notice that, standing to our current knowledge, there are not literature proposals about this precise task.

2.2 Credit Service database analysis.

The initial database is composed of over 800000 invoices, for which amount, issue date, expiry date, payment method and the registry of debtors and creditors are recorded among other data. However, all these invoices were issued by just 18 creditors. In particular: there are on average 47743 invoices and 2522 debtors per creditor. On the contrary, the number of debtors is very large (45403) and the credit history for each of them is very limited.

Creditor	Invoices per creditor	% of total
A	169456	19.72%
B	152870	17.79%
C	118364	13.77%
D	69584	8.10%
E	64727	7.53%
F	50047	5.82%
G	43961	5.12%
H	43609	5.07%
I	30992	3.61%
J	30918	3.60%
K	24880	2.90%
L	17125	1.99%
M	16201	1.89%
N	9103	1.06%
O	8267	0.96%
P	3903	0.45%
Q	3395	0.40%
R	1965	0.23%

Table 2.1: Invoices number for each creditor.

There are on average 19 invoices per debtor and in most cases debtors have invoices with respect to only one creditor. This unbalance makes it impossible to face the problem through a time series approach.

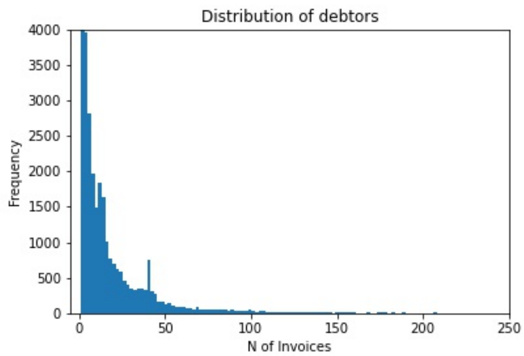

Figure 2.1: Invoices distribution through debtors. The y axis has been truncated at 4000 in order to make the plot readable.

Then, it's interesting to see if there is some kind of evolution in the payment behaviour during the years. In particular, we want to investigate the impact of time-related phenomena on the dataset such as the pandemic or economic cycles. In our model, we will distinguish among invoices paid *On time*, *Default* and *Strong Default*. The default of an invoice is defined as those cases in which the invoice has not been paid or has been paid with more than 7 days of delay, while the *Strong Default* is defined as the scenario with 90 or more days of dealy. In the following analysis we focused on *Default* case, because the *Strong Default* provides as it will be seen a very low number of observations. The percentage of default has been computed as the number of defaults over the total number of invoices recorded during the given year. Lastly, we calculated the percentage of *realized delay*, a measure strictly related to the composition of the database. This has been computed as the number of days of delay the invoice has recorded divided by the total time elapsed since emission. This aims at taking into account the fact that by construction older invoices can realize larger delays than those recorded more recently.

Year	**Defaults**	**% default**	**Average delay**	**% realized delay**
2021	33588	20%	24	21%
2020	65871	30%	62	14%
2019	71174	25%	126	16%
2018	65149	44%	82	7%
2017	21648	74%	329	22%
2016	4836	70%	972	51%
2015	7598	85%	1396	62%

Table 2.2: Defaults analysis

Apparently, there seems to be no major impact on defaults directly caused by time-related phenomena.

Anyway, this kind of information is not very reliable because the data collection during the first years of Credit Service's activities are prone to several errors, and later maintenance of the database also resulted in potential loss of older data. This was confirmed directly from the company representatives.

Another interesting analysis related to the database concerns the correlation between the amount of the invoice and the payment's delay. The following table shows that the delay seems not to be linearly affected by the invoice amount.

Finally, the last kind of information which seems interesting to investigate is related to the payment methods. Indeed, some of them naturally discourage delays in payments and defaults, as for example the RiBa or the RID method. We provide some analysis on the contribution of this feature to potentially discriminate payment behaviours. We disclose the analysis for two different datasets: the first one collects all the invoices while the second one includes only those invoices starting from 2018. This is to mitigate the

Average delay—Invoice amount			
Amount	**Count**	**Mean**	**Max**
[0, 250)	110008	62	364
[250, 500)	36022	55	364
[500, 750)	20443	56	364
[750, 1000)	11877	55	363
[1'000, 10'000)	48090	55	364
[10'000, 100'000)	9089	59	364
[100'000, 1'000'000)	272	47	349

Table 2.3: Average delay conditioned on the amount of money due.

potential unreliability of older data and to furtherly check this hypothesis.

Payment Method	% Default	Count	n° Creditors	Avg Delay
Legale	100%	33	1	199.50
Assegno Bancario	78%	4038	3	62.85
Contrassegno / contante	72%	6877	10	33.18
Rimessa Diretta	71%	53607	8	263.05
Bollettino postale	57%	37337	1	96.99
Bonifico bancario	48%	324598	17	31.12
Esattore	42%	63457	1	60.89
Bonifico estero	35%	13770	1	7.85
Altro tipo	34%	78606	7	63.30
RiBa	16%	190760	13	-30.32
RID	8%	69274	2	-13.12
Pagamento SDD	7%	17010	2	-6.36

Table 2.4: Relationship between payment delay and payment method over the entire dataset.

Payment Method	% Default	Count	n° Creditors	Avg Delay
Altro tipo	87%	87	5	88.72
Assegno Bancario	81%	289	2	79.45
Bollettino postale	62%	303	1	88.37
Rimessa Diretta	43%	9964	8	0.96
Bonifico bancario	38%	102738	17	13.59
Esattore	36%	3436	1	15.72
Contrassegno / contante	18%	1481	7	-10.93
RiBa	10%	91812	13	-33.90
RID	5%	22546	2	-18.85
Pagamento SDD	4%	6456	2	-9.54

Table 2.5: Relationship between payment delay and payment method for the invoices collected starting from 2018.

Our hypotheses have been confirmed by this analysis.
Lastly, the comparison emphasizes and confirms the differences among the invoices collected before 2018 and the recent ones. This is especially clear looking at the average delay statistics.

2.3 Data preprocessing and final datasets.

In order to construct our final dataset, we enriched the set of variables through the balance sheet and financial data of the debtors, using the Aida database. In particular, we associated the financial data of year N to each invoice emitted from June 1^{st} of year N to May 31^{st} of year $N + 1$. This is due to the fact that balance sheets are usually published with some months of delay with respect to the period which they are referring to. Furthermore, revisions are also provided in the following months.
Long story short, we ensured not to use data which in a live forecasting application would not be in fact available.
Then we constructed three further variables: the invoice emission month, the invoice payment month and the number of interactions. Given a particular invoice, the number of interactions is the number of past invoices between that particular debtor and that particular creditor (standing to the database). Please note that for the emission and payment months, we represented them as ordinal values ranging from 0 to 11. Indeed, the one-hot-encoding representation suggested by literature is not very suitable for the Random Forest algorithm we chose to implement, even if it does not represent a strong obstacle.
So, the final list of features for our problem modelling is the following, organized by sources:

- Database data: balance divided by total cost, (i.e. the amount of money to be paid for the single invoice divided by the total cost computed from the balance sheet data), maturity, emission month, payment month, ateco code (Classification of Economic Activity), payment method, number of interactions
- Balance sheet data: EBITDA, EBITDA/sales, Liquidity Index, number of employees, debt, net income, ROE, Working capital, net worth, short term debt, current assets, Total value of production.

We then selected for our final dataset those invoices whose debtors' balance sheet data was available in the Aida database, as well as filtering these invoices by emission date from 2018 onwards. This allowed us to to clean the dataset from missing data and from the old unreliable invoices as per the informations provided by the company.
We emphasize the fact that the features provided don't imply any information about the ID of the companies involved in each invoice, so there is no way for the algorithm to deduce any temporal structure or history among the invoices. Furthermore, the vast majority of the debtors in our dataset only have a few invoices recorded, with most of these consisting of one-off transactions.
This fact allows us to safely shuffle the invoices during the dataset split into train and test datasets - with a ratio 80/20 - providing different economic periods, cycles and regimes to the algorithm. This is particularly important in order to provide the model with a sufficient plethora of scenarios, especially given the pandemic situation and its impact on companies' balance sheets, also due to government choices in order to mitigate the negative pandemic effects. Then the *shuffle* option provides the required informativity to the dataset but does not affect the performance in any fraudulent way. Further experiments in order to rule out this hypothesis have been computed and are provided upon request.
As a result, our train set presents 146560 invoices correctly paid and 44729 invoices labeled as *Default*, while the test set is composed by 35365 invoices paid on time and 10887 *Default* occurrences.
Then, for the *Default* vs *Strong Default* model we have for the test set respectively 8837 and 2029 observations.

2.4 The Random Forest model

In our work, we tried to face a main classification task and a derived regression task. For the classification part, we have in particular three interconnected problems:

- Classify invoices between those paid with at most 7 days of delay and those which are paid after 7 days of delay (*On time vs Default*), i.e. estimate $P(D > 7)$. We call this one *Default Model*.
- Classify invoices which have been paid with more than 7 days of delay between those

paid within 90 days of delay and those paid after (*Default vs Strong Default*), i.e. estimate $P(D \geq 90|D > 7)$. We call this one *Default vs Strong Default Model*.

- Classify invoices between those paid with at most 89 days of delay and those which are paid after (*On Time & Default vs Strong Default*). This last one is made on the basis of the previous tasks. In particular, by exploiting the fact that $P(D \geq 90) = P(D > 7, D \geq 90)$, then we used the Bayes formula $P(D > 7, D \geq 90) = P(D \geq 90|D > 7)P(D > 7)$. This choice was due to the fact that the direct estimation of $P(D \geq 90)$ was made really difficult by a very unbalanced dataset. In this way, we reduce the unbalance among by running the training on a model focused on the second problem. We call this last one *Strong Default Model*.

Instead the regression part just tries to predict in how many days the invoice will be paid after it is send.
The model we chose for both tasks is the Random Forest. We also tried different models - Bagging, Gradient Boosting and Support Vector Machines - but the performance of the first was not as good as the Random Forest and the second one was computationally too heavy for our virtual machine. [61], [66], [21].
The Random Forest is an ensemble learning algorithm, based on Decision Trees. Given a space $S \subseteq \mathcal{R}^n$ where $n \in \mathcal{N}$ is the number of features provided to the model, a Decision Tree splits up this space S in different sub-regions, trying to maximize the purity of each sub-region with respect to the label, which in our case is a binary one corresponding to *On time* and *Default*.
Tree-based models partition the feature space into a set of subspaces, in particular rectangles, and fit a trivial model - usually a constant - in each subregion. They are conceptually intuitive but really powerful, allowing to decipher any kind of pattern relationship among the features and the output variable. The method we chose for the tree-based regression and classification called CART. X_i, each taking values in the unit interval. Then we split the space through recursive binary partitions. First we split it into two regions, and get the response by the mean of Y in each region. We choose the feature and split-point to achieve the best fit. Then one or both of these regions are split into two more regions, and this process is iterated, until some stopping rule is applied.
Now suppose than after the iterations M subregions R_m of the original subspace S have been produced, then the prediction is given by:

$$\hat{f}(X) = \sum_{m=1}^{N} c_m I\{(X_1, ..., X_n) \in R_m\} \tag{2.1}$$

where c_m is the Y mean or mode - depending on the task - in the subspace R_m. Observations satisfying the condition at each junction are assigned to the left branch, and the others to the right one. The terminal nodes (or leaves) of the tree correspond to the regions R_1, R_2, ..., R_M.

A great advantage of the recursive binary tree - in particular for our circumstances - is its interpretability. The feature space partition is fully described by a single tree. Given this general idea, let's deep dive in the Regression Trees.

2.4.1 Regression Trees

We start with the regression task, where we assume a continuous response Y and n features. Our data consists of n inputs and an output variable, for each of T observations. So we have (x_i, y_i) for $i = 1, 2, ..., T$, with $x_i = (x_{i_1}, x_{i_2}, ..., x_{i_p})$. The algorithm has to automatically produce a decision on the splitting variables as well as regarding the split points, and also which topology the tree should take. As already said, given the M subregions R_i, we model the output as a constant c_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{2.2}$$

With a sum of squares as loss function, i.e. $\sum_i (y_i - f(x_i))^2$, it is immediate to see that the best c_m is the average of y_i in region R_m:

$$\hat{c}_m = \mu(y_i | x_i \in R_m) \tag{2.3}$$

The problem is that finding the best binary partition under a minimum sum of squares cryterion is generally computationally devastating. thus the choice is to proceed with a greedy algorithm that, starting with the entire dataset, considers a splitting feature j and a split point s that is $s \in X_j$, and extracts the pair of half-planes:

$$R_1(j, s) = \{X | X_j \leq s\} \ and \ R_2(j, s) = \{X | X_j > s\} \tag{2.4}$$

Then we are looking for the splitting variable j and split point s that solve:

$$\min_{j,s} \left[min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + min_{c_1} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right] \tag{2.5}$$

where clearly c_1 and c_2 which solves the inner minimization are given by:

$$\hat{c}_m = \mu(y_i | x_i \in R_m) \tag{2.6}$$

with $i = 1, 2$. For each possible splitting feature, the determination of the split point s can be produced very fast, so that by scanning through all of the variables, determination of the best pair (j, s) is doable.

Once found the best split, we partition the data into the two resulting regions and iterate the splitting process on each of the two regions. Then this process is repeated again on all of the resulting regions.

Then the next problem is related to how large should we grow the tree. Obviously a very large tree is prone to overfit the data, and of course a tree that is left free to grow can produce a single prediction for each observation, in the more extreme case. On the other side, a small tree might not capture some important patterns in the structure.

Tree size is a tuning parameter - or hyperparameter - which governs the model's complexity, so that the optimal tree size should be adaptively chosen from the data. One possible approach is to split tree nodes only if the decrease in sum-of-squares due to the split exceeds some threshold. However, this strategy is too short-sighted because a seemingly worthless split might allow for a very good split in the next steps.

So, the preferred strategy is to first grow a large tree T_0, stopping the splitting process only when a given minimum node size is reached. Then ex-post this large tree is pruned through cost-complexity techniques.

We define a subtree $\tau \subset T_0$ to be any tree that we can get by pruning T_0, i.e. collapsing any number of its internal - that is non-terminal - nodes. Then we index the terminal nodes by m, with node m that represents the region R_m. Let $|\tau|$ denote the number of terminal nodes in τ. Letting:

$$N_m = \#\{x_i \in R_m\}$$

$$c_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(\tau) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - c_m)^2$$

then we can finally define the so called cost complexity criteryon:

$$C_\alpha(\tau) = \sum_{m=1}^{|\tau|} N_m Q_m(\tau) + \alpha|\tau| \tag{2.7}$$

The idea is to find, for each α, that subtree $\tau_\alpha \subseteq T_0$ which minimizes $C\alpha(\tau)$. The hyperparameter $\alpha \geq 0$ rules the tradeoff between tree size and its goodness of fit to the data. Large values of α result of course in smaller trees τ_α, and conversely for small values of α. With $\alpha = 0$ the solution is the entire tree T_0.

For each α it can be shown that there exists a unique smallest subtree τ_α that minimizes $C_alpha(\tau)$. In order to find $\tau\alpha$ weakest link pruning is used: we collapse in succession the internal node that produces the smallest per-node increase in $\sum_m N_m Q_m(\tau)$, and continue until the single-node (root) tree is produced. This provides a sequence of subtrees, and it can be shown this sequence must contain $\tau\alpha$. See [?] or [98] for details. Then the estimation of α is achieved by k-fold cross-validation, that is we choose the value $\hat{\alpha}$ that minimizes the cross-validated sum of squares. Our final tree is $\tau_{\hat{\alpha}}$.

2.4.2 Classification Trees

Consider a node m, representing the region R_m which collects N_m observations, let:

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \tag{2.8}$$

where k represents the classes, that in our cases are just two. Then p_{mk} is the proportion of observations pertaining to class k in node m. So the observations in node m are classified to class:

$$k(m) = argmax_k(\hat{p}_{mk} \tag{2.9}$$

that is the majority class in node m. For the $Q_m(T)$ function, which in the classification tasks represent the node impurity, we have different measures. The most used are the following:

$$\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)} \tag{2.10}$$

$$\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.11}$$

$$-\sum_{k=1}^{K} \hat{p}_{mk} \ln(\hat{p}_{mk}) \tag{2.12}$$

which are respectively the Missclassification Error, Gini Index and the Crossentropy.
In a two classes task such those here proposed, consider p to be the proportion of observations pertaining to the second class, then these measures are respectively:

$$1 - \max(p, 1 - p) \tag{2.13}$$

$$2p(1 - p) \tag{2.14}$$

$$-p \ln p - (1 - p) \ln(1 - p) \tag{2.15}$$

Those three measures are of course similar, but crossentropy and the Gini index are differentiable and this makes them more suited for numerical optimization. Comparing them, we note the need to weight the node impurity measures by the quantities N_{mL} and N_{mR} of observations in the two child nodes created by splitting node m. Furhermore, crossentropy and the Gini index are more sensitive to variations in the node probabilities than the misclassification rate. A classic example in this sense: we have a two-class problem with 400 observations in each class - call it $(400, 400)$, suppose a split generated the two child nodes $(300, 100)$ and $(100, 300)$, while the other created nodes $(200, 400)$ and $(200, 0)$. Both splits will produce a misclassification error of 0.25, but the second one generates a pure node and for this reason is preferable. Gini index and crossentropy are both lower for the second split. This is the reason why either the Gini index or crossentropy should be used when growing the tree. For the cost-complexity pruning, any of the three measures can be used, but typically the misclassification error is preferred.

2.4.3 Other details on Trees Algorithms

In this subsection we provide some other details about the Decision Trees based algorithms.

How to treat missing values

Often data has some missing predictor values in some feature. A possible way is to discard observations with missing values, but depending on the case this could bring serious depauperation of the training set. Other approaches includes filling the missing values with the mean of that predictor over the nonmissing observations.
But for tree-based models, there are two better ways. The first one is good for categorical predictors: simply a new category for *missing* is created. Notice that with this choice we might discover that observations which presentmissing values for some variable behave differently than those with nonmissing values.
A second more general approach is the creation of surrogate features. When a predictor is considered for a split, only the observations for which that predictor is not missing are used. After we choose the best (primary) predictor and split point, a list of surrogate predictors and split points is produced. The first surrogate is the predictor and corresponding split point that best fits the split of the training data achieved through the primary split. The second surrogate is the predictor and corresponding split point that provides second best, and so forth.
When sending observations down the tree both in the training phase or during prediction, the surrogate splits are used in order, provided that the primary splitting predictor is missing. Surrogate splits use correlations among predictors in order lighten the effect of missing data. The higher the correlation between the missing predictor and the other predictors, the smaller the loss of information due to the missing value.

Binary splits

Instead of splitting each node into just two groups at each step, we may consider multiway splits into more than two groups. In general this is not a good idea, even if in some cases it could be useful. Multiway splits divide the data too fast, then leaving insufficient data for the next level down. Furthermore multiway splits can be achieved by a series of binary splits, the latter are preferred.

Linear combinations of features

Rather than constraining the splits to be only in the form $X_j \leq s$, one can allow splits with respect to linear combinations $\sum_j a_j X_j \leq s$. The weights a_j and split point s are then optimized to minimize the provided relevant criterion. On one side this could improve the predictive power of the tree but on the other hand it can cause loss of interpretability. Furthermore, because of the discreteness of the split point search, it is difficult to optimize for the weights.
Notice that a similar result, even if not identical, can be achieved by applying a PCA

decomposition on the original feature space. The difference is that the Decision Tree process and the linear combination of the features are decoupled.

Instability

As it can be deduced, trees present high variance. Even a small change in the data can produce a very different series of splits, then making interpretation precarious. One of the reasons for this issue is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it. This can be lighten to some level by trying to use a more stable split criterion, but the physiological instability still persists. Random Forest averages many trees to reduce the variance.

Additive structure problem

Another problem with trees is that it is difficult for them to model additive structure. In a regression, suppose for instance that

$$Y = c_1 I(X_1 < t1) + c_2 I(X_2 < t2) + \epsilon$$

with ϵ as noise. Then what a binary tree can do is to first split on X_1 near t_1. Then at the next level it has to split both nodes on X_2 at t_2 in order to detect and reflect the additive structure. This might happen with a sufficiently large dataset, but there is no guarantees in general that the model will find such a structure. This problem is clearly amplified when the number of addictive effects increases, and in that case it would require many fortuitous splits to recreate the structure, also making it difficult to recognize ex-post. This problem is related to the binary tre recursive algorithm.

2.4.4 Random Forest

Random forests [27] is similar to bagging given that a large group of uncorrelated decision trees results are averaged, but with some important differences. The idea in bagging is to average many noisy but approximately unbiased models so to reduce the variance. Trees are very suited candidates for bagging, since they can detect complex interaction structures in the data, and with a potentially low bias. At the same time they are really noisy, and that makes them benefit a lot from the averaging mechanism.
Furthermore, provided that each tree produced in bagging is identically distributed, the expectation of an average of the trees is the same as the that of each of the trees. So the only room for improvement is related to variance reduction.
Now let B to be the number of some i.i.d. random variables, each of them with variance σ^2, then the variance of the average of the B r.v.'s is given of course vy:

$$\frac{1}{B}\sigma^2$$

This is due to the independent condition. But if the independency is removed, and we suppose a correlation ρ, let's say a positive one, then it is easy to see that the variance will be:

$$\rho\sigma^2 - \frac{1-\rho}{B}\sigma^2 \tag{2.16}$$

Now if we take the limit for $B \to \infty$ of the two different cases, we get respectively:

$$\lim_{B\to\infty} \frac{1}{B}\sigma^2 = 0 \tag{2.17}$$

$$\lim_{B\to\infty} \rho\sigma^2 - \frac{1-\rho}{B}\sigma^2 = \rho\sigma^2 \tag{2.18}$$

So it is evident that correlation of pairs of bagged trees limits the benefits of the average mechanism. The intuition behind the random forests is precisely to remove as possible the correlation among the trees, without icreasing too much the variance.

In order to pursue this purpose, the idea is to random select a subset of the features at each step of the tree-growing process.

Then, the prediction is obtained through a voting mechanism, inducing a natural probability measure: each Decision Tree provides its own prediction, and the label which collects the majority of votes becomes the predicted one.

Talking about the features importance, at each split and for each tree, the improvement in the split-criterion provides the importance measure related to the splitting feature, cumulated over all the trees in the forest, for each variable.

The random forest seems to fit well with the required tasks and the available dataset.

The hierarchical structure of trees in the first place is a welcome feature in the problems we face.

To better understand the importance of a hierarchical structure, keep in mind the classic example of the dataset of people on the Titanic. In this basic ML problem, the predictors are the characteristics of the people on board, such as gender, social class, etc., and the variable to be predicted is whether the given member survived or not.

In such a context - as applied modeling later demonstrated - the social class was a feature of capital importance in the modeling of the problem, and the rest of the variables showed predictive utility only once the contribution of the social class had been filtered.

Furthermore, given this hierarchical nature combined with the ensembling mechanism of random forests, the models are robust to collinearity issues.

Talking about the missing data, we just eliminated the rows and did not use any metdhod to handle the problem. We made this choice because our original dataset was really large, and the observations with missing data presented a lot of missing data.

That said, we consider an invoice as *Default* when the payment happens after 7 or more days from the expiration date. The functions which capture purity are the Gini index and the Entropy function.

On the train set, we selected the best hyperparameters for the Random Forest through a k-fold validation with $k = 5$. In particular, we got 460 decision trees and a max depth of

25 for each tree. The chosen objective function is the entropy function. Then we computed the feature importance, which allows some degree of transparency and interpretation:

Feature	**Impact**	**Aggregated**
Interactions	14%	**56%**
Payment method	11%	
Maturity	7%	
Expiration month	7%	
Ateco code	6%	
Emission month	6%	
Balance/costs	5%	
EBITDA	4%	**44%**
Short term credit	4%	
Short term debt	4%	
Total value of production	4%	
Working capital	4%	
Net income	3%	
EBITDA/sales	3%	
Number of employees	3%	
Current assets	3%	
Net worth	3%	
Net income/losses	3%	
Liquidity index	3%	
ROE	3%	
Long term debt	2%	

Table 2.6: Feature importance for the model *On Time vs Default*.

As it can be seen, the *number of interactions* was the most important single feature in our model, accounting for 14% in the entire decision process. This is expected and easily interpretable: a company which has a long business history with a partner is far more prone to pay the related invoices on time.
In second place we find the *payment method*, which affects payment behaviour in the aforementioned ways with an impact measured at 11% on the decision process.
Then we have time related features, such as the *maturity* and the *emission* and *expiration month*, which account respectively for 7%, 7% and 6% of the prediction. These features work well together, allowing for example to capture payment behaviours in particular periods of the year, such as those close to holidays or vacations. Furthermore, these features can partially detect the periods in which the pandemic situation was particularly impacting.
Next the *Ateco code* comes in place, whose influence has been assessed at 6%. The Ateco

code is an indicator of the main area of business of the debtor company, and it may be particularly useful when some kind of shock has affected a specific business area as a whole. During the pandemic period its importance is easily interpreted.
Then, we have the *balance over costs* feature with 5% impact, which measures how much the particular invoice is affecting the debtor total costs.
Finally we have the balance sheet features, which partly influence the balance over costs feature as well, given that the costs were computed from the balance sheet data. Even if there is not any *balance sheet item* which clearly dominates the others as importance and impact, it has to be said that their aggregate impacts for 44% on the decision process.
The feature importance analysis for the case *Default* vs *Strong Default* is quite similar and available upon request.
Please notice that also different approaches were proposed. In particular, we applied dimensionality reduction techniques such as PCA and Autoencoder Neural Networks, in order to extract a compact feature space representation. These attempts - especially the autoencoders - provided a slight improvement in the performance measures (about a 3% accuracy increment) but we discarded them because the loss of interpretability was too strong for such a small improvement.

2.5 Model performance analysis

In this section we report and deeply analyze the performance recorded by the models on the test set. The following table summarizes some important metrics used to formally capture how good the prediction is. In particular we have:

- **Accuracy:** the percentage of correct predictions. Even if it helps to get a picture of the whole performance, the accuracy metric is far from being sufficient when the dataset is unbalanced with respect to the labels. This is exactly our case, because the ratio between the invoices paid on time and those considered in default is 3/1 circa. To better understand why the accuracy is a misleading measure in these cases, consider a toy scenario where we have 200 invoices. 160 of them are labeled as correctly paid while the remaining 40 are defaulted. A trivial model which always predicts a correct payment would meet an 80% accuracy, despite not capturing any of the invoices in default. This is only an apparently a good result. For this reason, the next metrics are very important to get the full picture.

- **Precision:** given a target class of the problem and taking into account all the occurences in which that class is predicted, precision measures the percentage of cases in which the prediction was correct. The more strict the prediction cryterium for a given class - which in our case is just the probability treshold 0.5 - the higher the precision.

- **Recall:** given a target class of the problem, recall measures the percentage of occurrences of that class that are captured by the model. The less strict the prediction cryterium for a given class, the higher the recall.

It is then clear that precision and recall are antagonist measures. That said, there are some ways in which these measures can be combined in order to provide a unique value which captures the total performance, as well as basic averaging techniques which try to merge the measures across the different classes.

- **F1-score:** let *tp*, *fp* and *fn* to be respectively the number of true positives, false positives and false negatives, than the F1-score is given by

$$\frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

- **Macro average:** this measure is just the simple average of the different measures across the two classes

- **Weighted average:** it's the same as the previous measure but the average is weighted with respect to the support, i.e. the number of occurences for each class.

2.5.1 Default Model

Here are presented the results for the *Default Model*:

	Precision	**Recall**	**F1 - Score**	**Support**
On time	95%	96%	96%	35365
Default	87%	84%	85%	10887
Accuracy		93%		46252
Macro Average	91%	90%	91%	46252
Weighted Average	93%	93%	93%	46252

Table 2.7: Test set performance measures

As it can be seen from the table, the model has good precision measures. When the model predicts an invoice will be paid on time, the prediction is correct in 95% of the cases. This is also because the model captures 84% of the invoices in default.
Since the model shows a very high recall for the *On Time* class, we are prone to sacrifice some of this recall in favor of more precision, considering that in general a missed opportunity is better than a realized loss. So, after the train phase, we chose to lower the treshold of our model for the default prediction from a 50% to a 40% probability. In the next table the new results are summarized.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	96%	94%	95%	35365
Default	83%	89%	86%	10887
Accuracy		93%		46252
Macro Average	90%	91%	90%	46252
Weighted Average	93%	93%	93%	46252

In addition to these metrics, it's interesting to check how the probabilities predicted by the Random Forest through the voting mechanism are distributed, in order to investigate the model's confidence in its predictions. In the ideal scenario, we want the following: when the model correctly predicts on time payments, we want the default probabilities of the output to be particularly concentrated on very low values; when the model correctly predicts a default, we want the probabilities to be concentrated on very high values; for the wrong predictions we want the probabilities to be more oriented on values around 50% (or 40%, depending on the chosen treshold) from one side or the other, depending on whether the errors are false positives or false negatives.

This kind of distribution would ensure that the model is very confident with respect to its correct predictions but doubtful when it comes to being wrong. In the next figures these distributions are plotted, with respect to the 40% treshold, but of course with the natural treshold of 50% the sense of the results is the same.

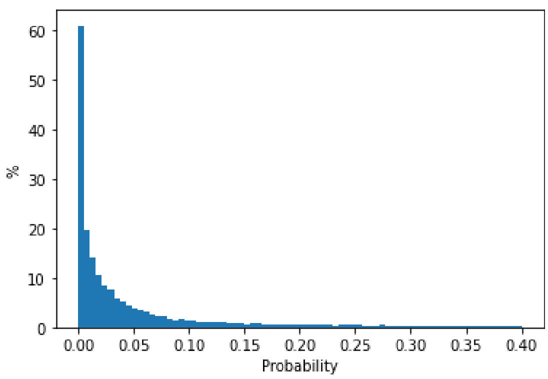

Figure 2.2: True negatives

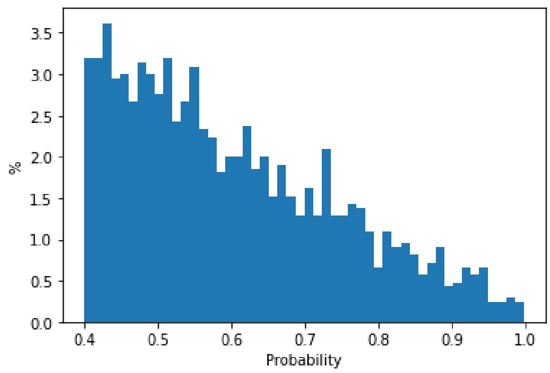

Figure 2.3: False positives

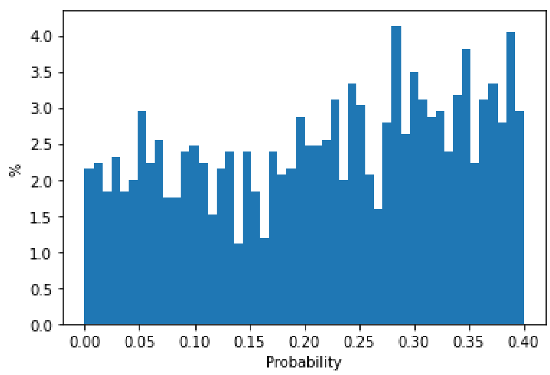

Figure 2.4: False negatives

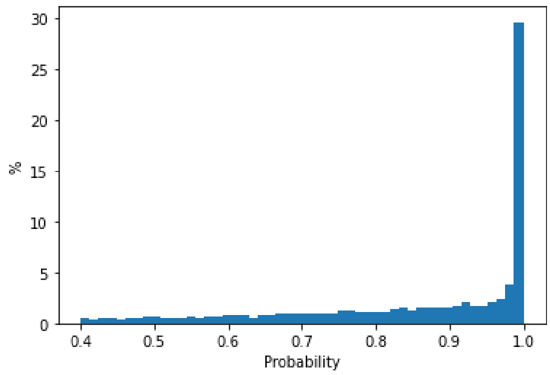

Figure 2.5: True positives

These plots show that the model behaves in the correct way: it is really confident for correct predictions and tends to be doubtful for wrong decisions. This behaviour is even more pronounced when the true label is *On time*. This is probably because the dataset presents more invoices paid before due date than after it, but could also be related to a more general fact: paraphrasing Tolstoj, all the solvent debtors are similar to each other, while each insolvent debtor is insolvent in a peculiar way.

In conclusion, given the final application area of the proposed model, it's worth to analyze the risk associated with the prediction of false positives and, even more, the prediction of false negatives. In our case, the risk is given by the balance and delay. Here we show some statistics with respect to these two measures of risk for the 40% treshold, but the results are quite similar for the natural treshold as well. 1 is the label for *Default*, 0 is the label for *On time*.

T = *True*	**Balance**			**Delay**		
P = *Predicted*	**Mean**	**Std. Dev**	**Median**	**Mean**	**Std. Dev**	**Median**
$T = 1$	3078	32972	366	66	107	26
$T = 1,\ P = 1$	2786	13699	359	65.49	106	26
$T = 1,\ P = 0$	5376	90322	417	68	114	24
$T = 0$	1736	7860	354	—	—	—
$T = 0,\ P = 0$	1683	7345	349	—	—	—
$T = 0,\ P = 1$	2601	13762	441	—	—	—

Table 2.8: Risk analysis for 40% probability treshold

As it can be seen, the model could be improved from this perspective. In particular, the true positives captured by the model ($T = 1,\ P = 1$) show a lower balance with respect both to the sub-dataset of true defaults ($T = 1$) and to the one of false negatives ($T = 1,\ P = 0$). At the same time, the true positives show a lower payment delay with respect the other subdatasets. This means that our model is more prone to capture insolvent behaviours which have to do with lower balance and payment delays, i.e. the less risky ones.

Something similar happens with the invoices correctly paid ($T = 0, P = 0$) and their balance. Future improvements can be oriented at trying to penalize the incorrect classifications which are related to higher delay and higher balance.

Anyway, the weakness in terms of risk seems to be related to the presence of outliers. In fact, the risk associated problems of the model can be mitigated if we consider the median values instead of the mean ones, which are more prone to the distortive action of outliers.

Then, in order to check for the robustness of the proposed model and its ability to generalize, two experiments have been performed:

1. We excluded from the test set each row which was identical in terms of balance sheet data to any row in the train set.

2. We excluded from the test set each row which was related to a debtor which was in the train set.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	84%	81%	82%%	511
Default	75%	76%	75%	385
Accuracy		79%		881
Macro Average	82%	75%	79%	881
Weighted Average	81%	78%	80%	881

Table 2.9: Test set results after excluding rows with identical balance sheet with respect to any observation in the train set.

As it can be seen, the results get obviously worse with respect to the base case. Still the obtained results guarantees that the model has in fact learnt robust patterns in order to recognize the *Defaults*.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	85%	89%	87%%	1004
Default	77%	70%	73%	526
Accuracy		82%		1530
Macro Average	81%	79%	80%	1530
Weighted Average	82%	82%	82%	1530

Table 2.10: Test set results after excluding rows related to debtors which already appeared in the train set.

Of course even in this case the results get worse with respect to the base analysis, but at the same time we can see that the model is good enough to generalize its predictions with respect to the debtors which were not in the training set.
Further analysis with respect to these particular subsets of the test set such for example distribution of predicted probabilities are available upon request.
At last, we compared our results with those obtained by exploiting Mode Finance ratings. We did that by predicting as *Default* an invoice which debtor rating was lower than *CCC* an *On Time* for the residual ratings.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	77%	95%	84%%	20934
Default	29%	7%	73%	6417
Accuracy	73%			27351
Macro Average	53%	51%	53%	27351
Weighted Average	65%	68%	65%	27351

Table 2.11: Mode Finance *Default Model*.

As it can be seen the model extracted from Mode Finance ratings is far from good in predicting invoices *Defaults*. In fact, this kind of prediction is really close to a naive prediction, i.e. the model based on Mode tend to predict an on time payment in almost any case. Furthermore, it is interesting to check if there is any correlation among the predictions of the models - the Random Forest and the Mode Finance one - and the percentage of realized *Defaults*:

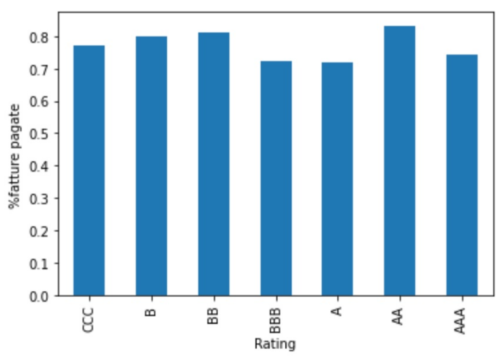

Figure 2.6: Mode Finance

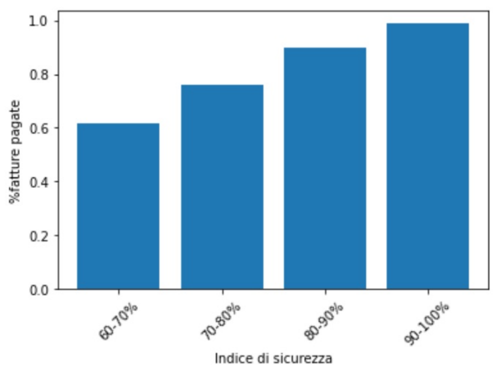

Figure 2.7: Random Forest

It is then clear that there is no correlation among the ratings predicted by Mode and the realized *Defaults* while there is a clear one among the probabilities predicted by the Random Forest and the *Defaults*. Please notice that the test dataset employed for the

Random Forest performance assesment and for the Mode Finance model are not identical, even if we tried to preserve the most similarity we could. This was due to economical reasons, which constrained us to preserve the Mode Finance test set as compact as we could.

2.5.2 Default vs Strong Default Model

In this section we analyze the results for the *Default vs Strong Default* Model, where the aim is to classify if an invoice have been paid in the $(7, 90)$ interval of delay or in the $[90, ...)$ one. The results on the test set are the following:

	Precision	**Recall**	**F1 - Score**	**Support**
On time	96%	97%	96%%	8837
Default	87%	80%	83%	1897
Accuracy	95%			10734
Macro Average	92%	88%	90%	10734
Weighted Average	94%	94%	94%	10734

Table 2.12: *Default vs Strong Default Model.*

In order to check for the model robustness, we excluded again from the test set all those rows which were identical in terms of balance sheet data to any observation in the train set.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	86%	97%	91%	689
Default	36%	20%	28%	118
Accuracy	85%			807
Macro Average	61%	58%	60%	807
Weighted Average	68%	65%	68%	807

Table 2.13: Test set results after excluding rows with identical balance sheet with respect to any observation in the train set.

In that case, not only the results get worse but at the same time there seems to be evidences of the difficulty of the model to generalize its findings. In our opinion, even if the result is a warning signal which suggests to deepens the approach for this kind of prediction, the output produced is not very reliable given a very low number of observations for the *Strong Default* class.
Further analysis are available upon request.

2.5.3 Strong Default Model.

In this section we analyze the results for the *Strong Default* Model, where the aim is to classify if an invoice have been paid in the $(..., 90)$ interval of delay or in the $[90, ...)$ one. The results on the test set are the following:

	Precision	**Recall**	**F1 - Score**	**Support**
On time	99%	99%	99%%	44337
Default	80%	73%	77%	2047
Accuracy	98%			46384
Macro Average	90%	86%	88%	46384
Weighted Average	94%	94%	94%	46384

Table 2.14: *Strong Default Model.*

In order to check for the model robustness, we excluded again from the test set all those rows which were identical in terms of balance sheet data to any observation in the train set.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	93%	99%	97%	812
Default	59%	10%	40%	69
Accuracy	96%			881
Macro Average	76%	55%	69%	881
Weighted Average	81%	65%	77%	881

Table 2.15: Test set results after excluding rows with identical balance sheet with respect to any observation in the train set.

In that case, not only the results get worse but at the same time there seems to be evidences of the difficulty of the model to generalize its findings. In our opinion, even if the result is a warning signal which suggests to deepens the approach for this kind of prediction, the output produced is not very reliable given a very low number of observations for the *Strong Default* class.
Further analysis are available upon request.
Even in this case, we present here the results we get by basing the prediction on Mode Finance ratings.

	Precision	**Recall**	**F1 - Score**	**Support**
On time	96%	93%	95%%	44377
Default	8%	13%	10%	2047
Accuracy		90%		46424
Macro Average	52%	53%	53%	46424
Weighted Average	65%	68%	65%	46424

Table 2.16: Mode Finance *Strong Default Model*.

As it can be seen the model extracted from Mode Finance ratings is far from good in predicting invoices *Defaults*. In fact, also in this case the prediction is really close to a naive prediction, i.e. the model based on Mode tend to predict an on time payment in almost any case. Furthermore, it is interesting to check if there is any correlation among the predictions of the models - the Random Forest and the Mode Finance one - and the percentage of realized *StrongDefaults*:

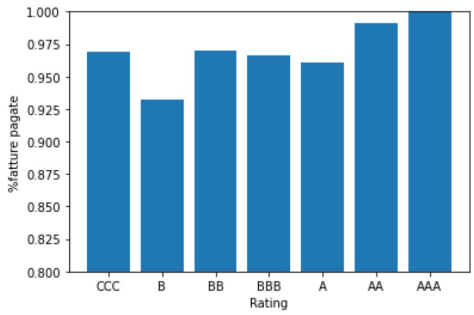

Figure 2.8: Mode Finance

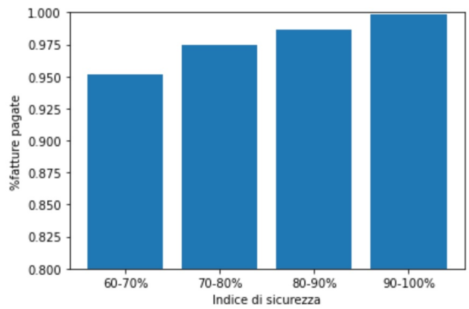

Figure 2.9: Random Forest

It is then clear that even in this case there is no correlation among the ratings predicted by Mode and the realized *Strong Defaults* while there is a clear one among the probabilities predicted by the Random Forest and the *Strong Defaults*.

2.5.4 Regression Model

Given that our classification tasks were based indirectly - through labeling - to the payment delay, we also tried to directly regress the payment delay.

The model structure and hyperparameters are identical to the classification model. On the response variable we applied two cut-offs in order to preserve a suitable range. Indeed delays in the order of thousands days which are not so rare in our dataset, just tends to polarize the predictions and to fictiously increase error. Even if they cannot be considered outliers in a strict statistical sense, they just distort the predictions in an unuseful way.

The same happens for the negative delays, which are those cases on which the company paid before the expiration date.

That said, we set a low cut-off to 0 - that is: all negative delays are set to be 0 - and the high cut-off to 480 - that is: all values higher than 480 are set to be equal to 480.

Here we present the results for the regression prediction on the entire datasets. In the following plots are the real delay vs predicted delay for the first 800 observations. Being the the test dataset very large - about 50000 observations - a complete plot would have been infeasible.

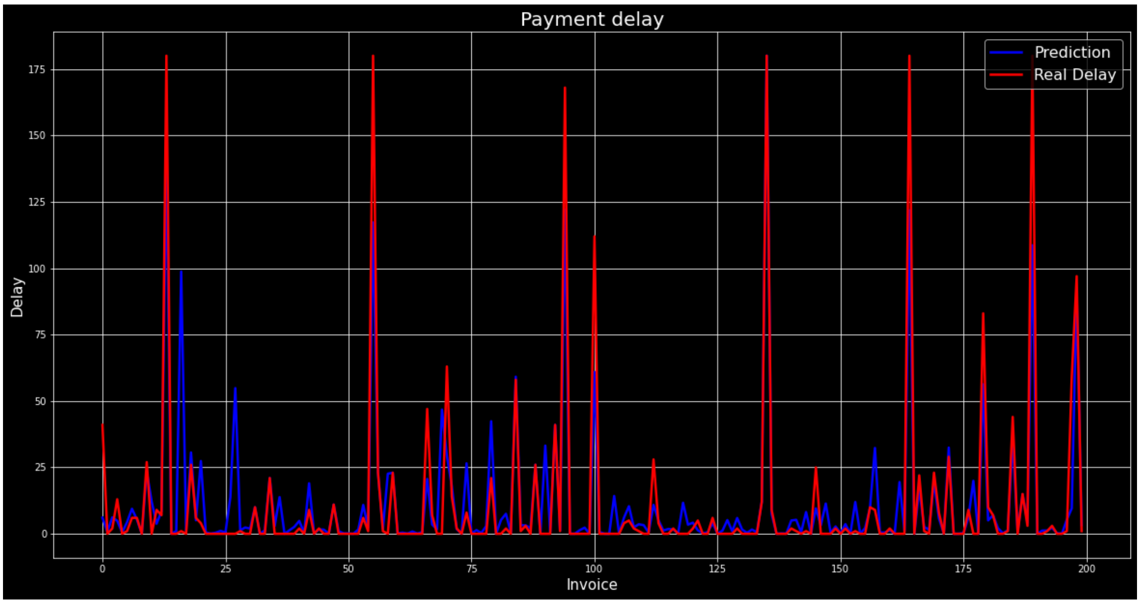

Figure 2.10: Payment Delay: Prediction vs Real - first 200 invoices

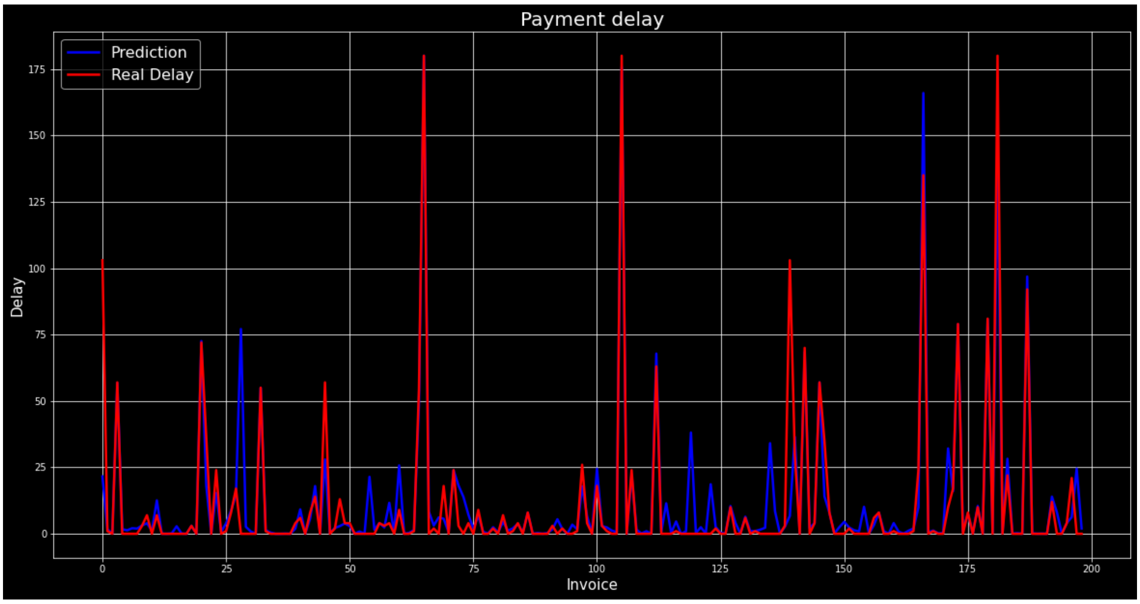

Figure 2.11: Payment Delay: Prediction vs Real - from 201 to 400 invoice

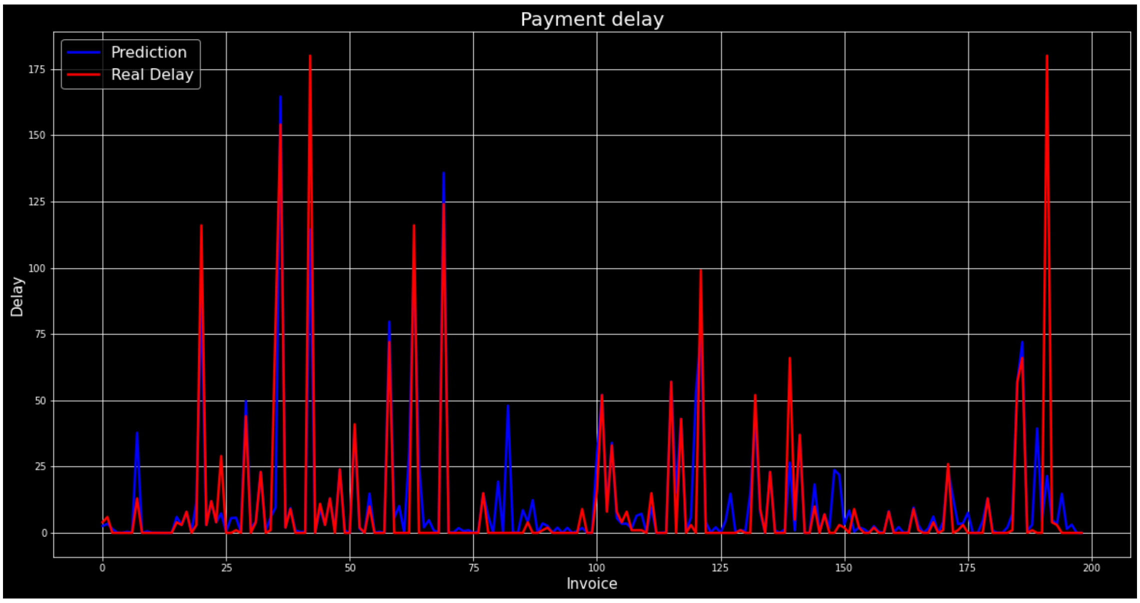

Figure 2.12: Payment Delay: Prediction vs Real - from 401 to 600 invoice

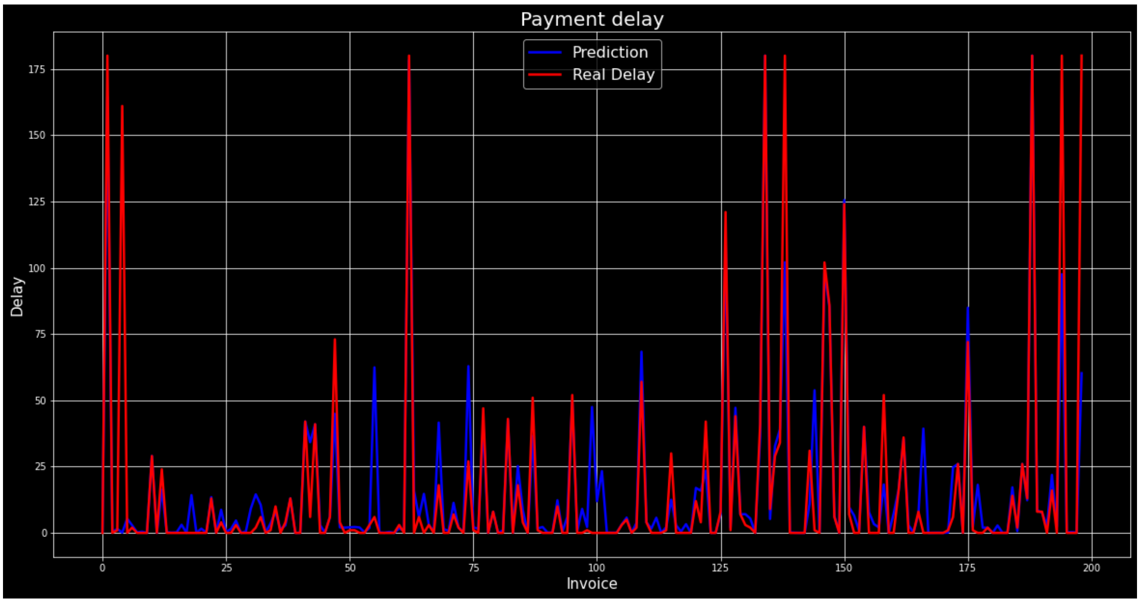

Figure 2.13: Payment Delay: Prediction vs Real - from 601 to 800 invoice

In metrics terms:

- MAE: 6.45
- MSE: 333
- RMSE: 18
- Pseudo R^2 score: 77%

As it can be seen, the results are really good, especially looking at MAE and to R^2. The problem is that things dramatically change if we include on the test set only those observations which balance sheet data are quite different from anything seen in the training set.

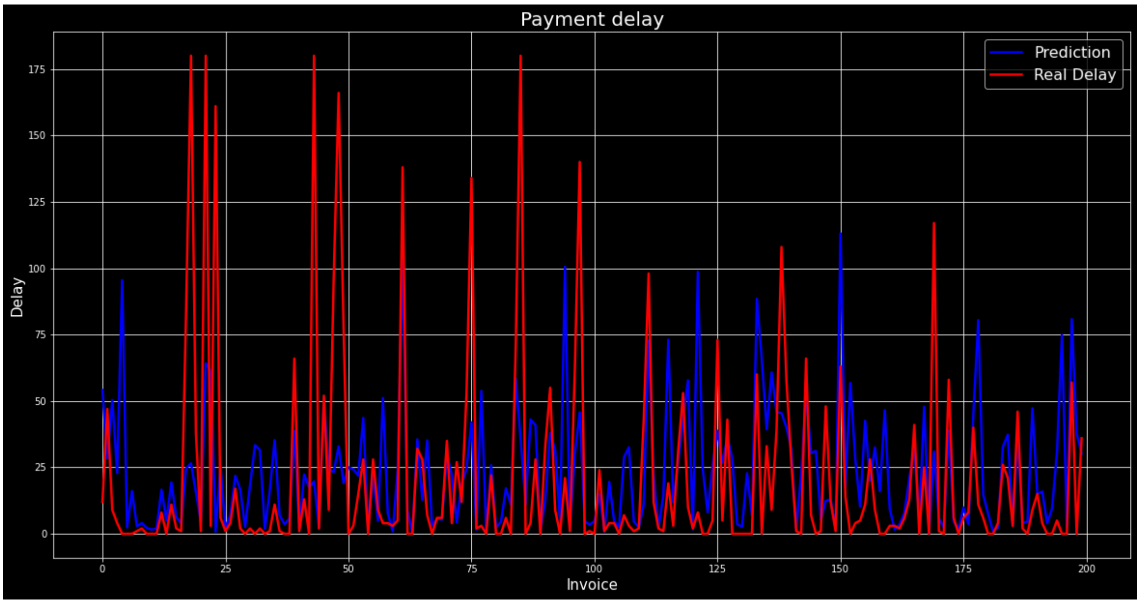

Figure 2.14: Payment Delay: Prediction vs Real - first 200 invoices

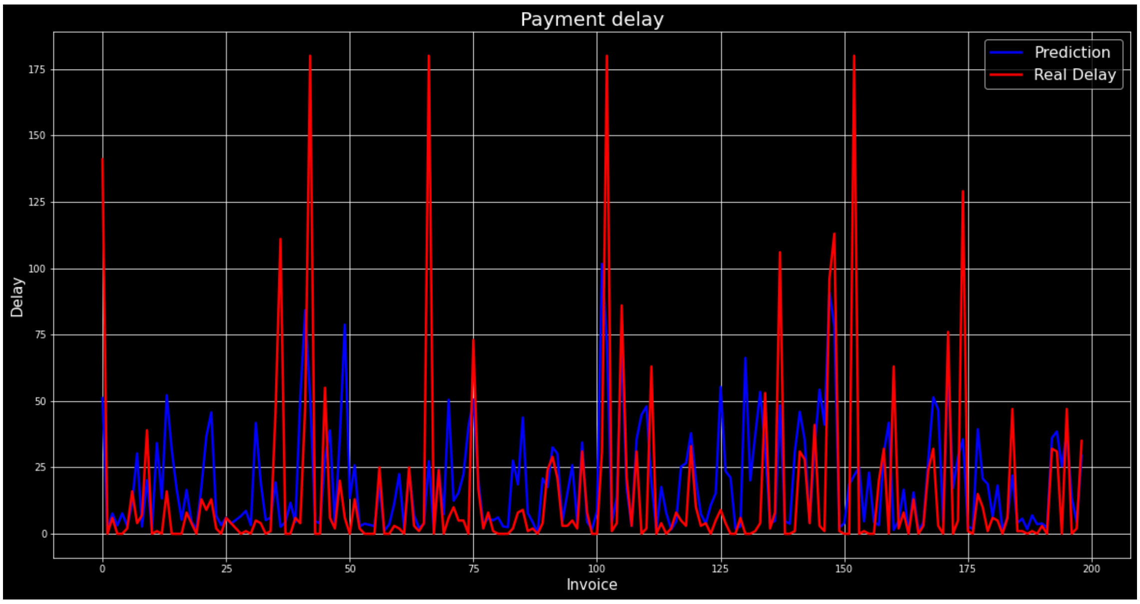

Figure 2.15: Payment Delay: Prediction vs Real - from 201 to 400 invoice

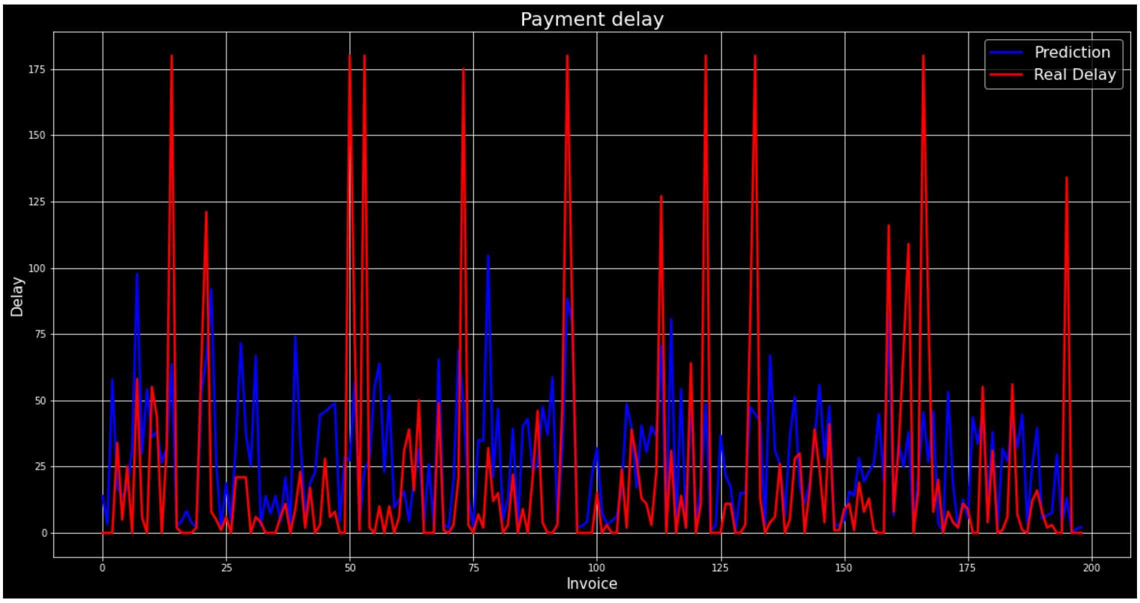

Figure 2.16: Payment Delay: Prediction vs Real - from 401 to 600 invoice

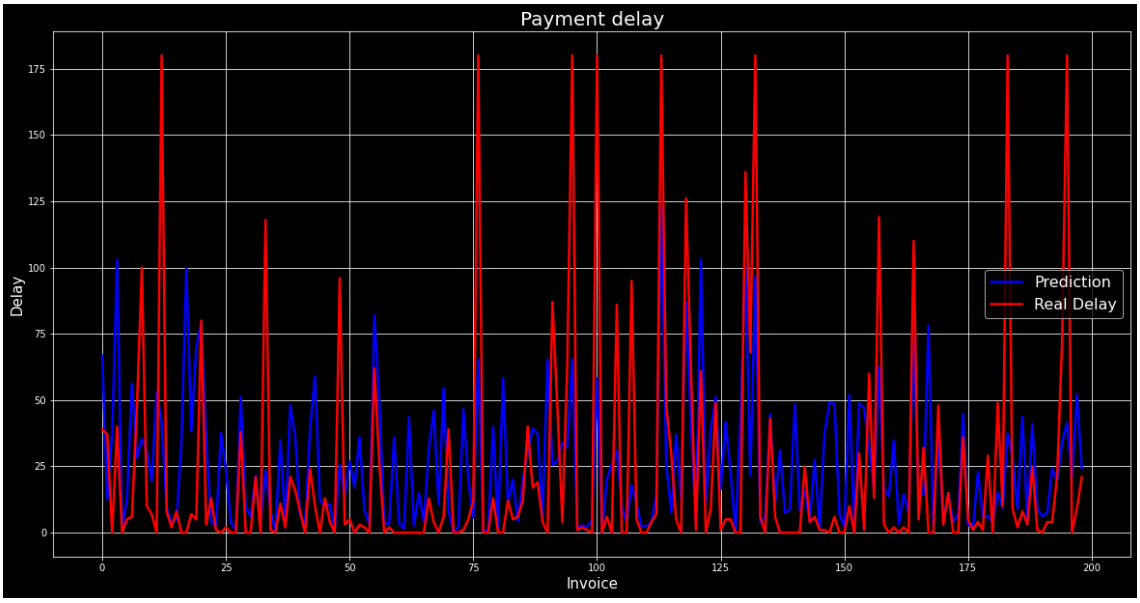

Figure 2.17: Payment Delay: Prediction vs Real - from 601 to 800 invoice

In metrics terms:

- MAE: 20
- MSE: 1203
- RMSE: 35
- Pseudo R^2 score: 18%

As it can be seen the metrics dramatically got worse, especially the R^2. However, the MAE is more than acceptable, and looking at the dataset this result seems more related

to some poverty in the dataset - to be intended as variance in the variables.
However, the regression model is still useful as a monitoring risk tool.

2.6 Future research.

The room for possible improvements and experiments is vast, and it spans several areas. First of all, it will be useful to compare our model's results with more baseline and/or traditional models. First of all, a comparison with the Altman Z-score and Altman Z-score plus models [2] would be interesting. These models are very popular among the practitioners, and over years and across markets have shown a very good prediction power. They are based on a bunch of financial ratios linearly combined so to provide the best possible discrimination among solvent and bankrupted companies.
Another noticeable contribution has been provided in [19], where has been proposed and tested a Distance from Default formula extracted through the Merton bond pricing model proposed in [86], proving its forecasting power in default prediction.
These attempts has not been already produced for two reasons, both related to the business needs: the only important comparison was with respect to the Mode Finance model and the solution we proposed needed to be Machine Learning based. As for improvements room, feature engineering based on domain expertise could be a first idea. We experienced the improvement produced by the creation of the *interactions* feature as well as the ratio among invoice *balance* and total costs faced by the given company. Also better search in this direction through ML techniques as PCA and Autoencoders deserves attention in the future. Furthermore, it could be interesting to provide as input in the Machine Learning model the output produced by classical models such those aforementioned, then using the Machine Learning algorithm as a standard model aggregator and optimizator.
Another important issue is that related to the dataset unbalance, both in the predictor variables as well as in the output one - especially in the last case when we are facing the prediction of *Strong Defaults*. For this problem, it could be useful to group the invoices paid on time in n clusters according to some features, and then to train n classifiers on n different datasets which are constitued of the i cluster of invoiced paid on time plus the invoices which exhibited *Strong Default*. Then a final model would be composed from a clustering part which assign the observation t to its pertaining model i, trained on the dataset produced by the invoices paid on time which are in the i cluster and the invoices labeled as *Strong Default*, which are fixed. This approach not only resizes the unbalance problem, but also provides a different and more informative point of view. The invoices paid on time are no more a monolithic group, but they are clustered according to some peculiarities, so that the classification models learn to specialize on different features leading to virtuous behaviors.
Talking about unbalance in the predictors part, we observed that we have for a single company more than 5000 invoices. This fact of course provides to the company combi-

nation of features and output a huge weight in the optimization scheme and so in the model learning. In the future, some research on how to capture the main information related to this company history hence reducing its presence - which is a delicate manner - in observations could be useful.
The last direction is strictly related to the computational power at use. For example we could not experiment with Support Vector Machines because it required too much power. Last, what would be interesting is a shift of paradigm, from a *naive* approach related to classification to a more business oriented purpose. The idea is then reformulate the problem by associating to the different invoices payment behaviors, the financial output faced by the company, summarized in a loss function, and then trying to optimize it with Machine Learning based approaches.

Part II

Computer Vision and Feature Selection

Chapter 3

Deep Learning approaches to detect damaged smartphone screens

This work aims to classify smartphone images based on whether the smartphone screen is damaged or not. The problem is divided into two complementary tasks: classify whether the image represents a smartphone or not and classify whether the smartphone's screen is damaged or not. To tackle the problem, several famous convolutional architectures are tested, and on the basis of these preliminary results an adapted architecture is obtaineden which proves to be a powerful and effective solution to the proposed problem.

3.1 Damaged smartphone screens recognition

On the basis of a real business problem, this paper proposes a novel approach to detect damaged smartphones screens. An insurance company is interested to provide for its customers a new insurance policy, which triggers when the smartphone screen get damaged. More precisely, the company asks to propose both in terms of methodology and computation an algorithm which automatically recognizes from a picture if the smartphone is damaged or not.
Very good results in this kind of problem can be obtained, as will be shown, by adapting only the final classification layer of pre-trained popular convolutional neural networks. The results thus obtained - in the case of our dataset around 90% accuracy on average - may however not be high enough for an implementation involving a real business initiative.
About 10% of errors, although on paper it is an excellent result, can in fact lead to problems and compromise the success of a product. In fact, such a result, translated into concrete terms, means that in 10% of cases either the customer who subscribes to the

insurance is dissatisfied or that the Insurance Company will have to guarantee further checks that will extend the time, however leading to a certain degree of dissatisfaction with the customer.
This percentage, depending on the customer base, can result in huge numbers, which in addition to direct losses can affect the long-term success of the insurance plan. Negative reviews or word of mouth from such a percentage of people can result in poor appeal to other potential insurance plan customers.
For these reasons, the approach proposed here focuses on obtaining the most excellent results possible in terms of accuracy as well as guaranteeing the robustness of the model with respect to a large and varied dataset.

Different approaches are available in the literature and a very natural solution was provided in [67], through Edge Detection Algorithm (i.s. Canny Detection Algorithm) [30]. Since the type of damage in a smartphone could be given by scratches, scrapes, cracks and so on, Edge Detection oriented solution is quite effective. That said, this approach was in general outmoded by the Convolutional Neural Networks (CNN) algorithms [73]. In fact a CNN, even if not specifically addressed for an edge detection task, spontaneously learns to capture the edges with a sufficient accuracy to pursue classification purposes, especially in the first layers. CNN have been widely used for similar problems as for example in [75] and [77], as well as in [42]. The paper is organized as follows: Section 2 describes how the datasets have been constructed and what kind of images have been used; Section 3 shows how the model has been constructed and improved and how the datasets evolved in order to better train the model; finally Section 4 reports conclusions.

3.2 The methodological approach

Two different classification models, based on Convolutional Neural Networks architectures have been trained. The first one recognizes if a given image depicts a smartphone or not and this is Model 1; then the second one classifies the image as *Broken* if the smartphone screen is damaged or *Unbroken* viceversa and this is Model 2. The Dataset is made out of 3122 images, collected within *a priori* defined categories: *Broken Back*, *Broken Screen*, *Unbroken Back*, *Unbroken Screen*, *Other* and *Other Tricky*. *Other Tricky* category collects images of that kind of objects which could typically decept the algorithm by inducing it to classify them as smartphones. Emblematic instances of these objects are MP4, Tablet, TVs, e-readers and Notebooks. This first taxonomy is relevant to understand how the categories are represented in the dataset.

	Model 1		**Model 2**	
	Phone	Other	Broken	Unbroken
Broken Screen	901	0	901	0
Unbroken Screen	768	0	0	768
Broken Back	12	0	0	12
Unbroken Back	82	0	0	82
Other	0	857	-	-
Other Tricky	0	492	-	-
Percentage	56%	44%	51%	49%

Table 3.1: A priori distribution across the two models and their respective classes.

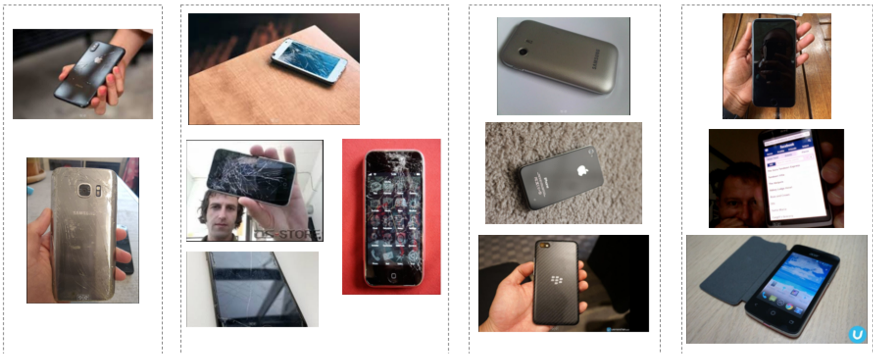

Figure 3.1: Picture istances respectively from *Broken Back*, *Broken Screen*, *Unbroken back* and *Unbroken screen a priori* categories.

Figure 3.2: Picture istances respectively from *Other* and *Other Tricky a priori* categories.

Table 1 depicts the image distribution across the *a priori* subclasses. At this point the images were distributed homogeneously with respect to their a priori subclasses in the Train, Validation and Test datasets, which collected respectively 75%, 15% and 10% of the observations.

Model 1 is characterized by a binary target variable with value equal to 1 if a smartphone

is in the picture and 0 if not. For *Model 1*, the a priori categories *Broken Back*, *Broken Screen*, *Unbroken Back* and *Unbroken Screen* were obviously collected in the class 1 or *Phone* class, while the *Other* and the *Other Tricky* subclasses were labeled as 0 or *Other* class.
Model 2 classifies as 0 or *Broken* the images belonging to the Broken Screen cathegory, while the subclasses *Broken Back*, *Unbroken Back* and *Unbroken Screen* were assigned to 1 or *Unbroken* label.

3.2.1 Model 1

For Model 1 some well known CNN architectures have been selected, to test their performance. We briefly describe these deep neural networks, to highlight their peculiarities and the relative importance in image recognition problem.
The **VGG16** CNN is a standard CNN firstly presented in [109] and recognized in Large Scale Visual Recognition Challenge 2014. VGG16 represents a very good starting point for image recognition. VGG16 is made up of 16 weights layers with ReLu as activation functions. It adopts only convolutional and max pooling layers and it uses (3×3) convolutional kernels with 1 stride and (2×2) max pooling layers, for a total of about 134 million of trainable parameters.
The **Residual Net 50** [62] is a direct answer to some well known issues related to the VGG16 architecture and more generally with the so called *very deep neural networks*. Since neural networks can be looked as universal functions approximations, it is clear that deeper architectures provide a better performance in these terms. The issue with such deep networks is the so called vanishing gradient problem, which is related to the optimization part. The fact is that, through the back-propagation algorithm [102] the gradient decays too much quickly, and then the information aquired by the first layers results to be poor. There have been several attempts to deal with this issue but none of them provided a real improvement. The Residual Net architecture provides a very natural solution to this problem through the *Residual Block* implementation, also known as *Identity Block*, because the mapping function is in fact the identity function. This approach consists in adding a shortcut (also called skip connection) to allow information to flow more easily from one layer to the next's layer, in this way preserving the information. Through this technique, the stacked layers try to fit a residual mapping instead of the desired underlying mapping. In this way, adding additional layers will not affect the model performance, because the related weigths will be setted to zero in the case in which the additional layers are not useful. At the same time, if the new layers learn useful information their training will be no affected by the new Residual block, given the linear relation which connects the Identity block with the feature map obtained through the layers. Residual Net is then another natural attempt for the problem we present here, and the Residual Block innovation has been preserved as a good solution to the highlighted problems even

in more recent architectures. Moreover, the kind of issue the Residual Net deal with could be important for our particular problem. In fact our algorithm needs to learn basic figures and patterns - at least for the first stage problem - given that smartphones are with an excellent approximation rectangular. These are exactly the kind of patterns which are learnt by the first architecture layers, i.e. the most affected ones by the gradient vanishing problem.

The **Residual Net 152** [62] represents an improvement in term of deepness of the already mentioned Residual Net 50. Its characteristics are almost the same as the Residual Net 50, and the main difference is that the Net exploit the power of Residual Blocks to achieve a greater deepness. Indeed when the architecture was published, it was the deepest net of its time. Then the reasons behind this choice are the same as Residual Net 50, but with far greater number of layers.

The **Residual Inception** architecture [**?**], combines in an optimal way the peculiarities of the Residual Net which we already discussed with the structure and flexibility of the Inception networks. The main contribution of the Inception modules, is that they drop some degrees of freedom in choosing the network structure and layers, by allowing in such a sense the net itself to *choose* the most fitting layers. Instead of choosing for each layer the kind of layer (e.g. convolutional layer, pooling layer, etc.)as well as the hyper-parameters such as convolutional kernel dimensions, kernels, etc., the Inception module allows to use all of them and then concatenate the outputs along the depth dimension. In order to put together these different outputs, usually you need them to share the same dimension. In order to pursue this aim, some tricks are adopted, as for instance using the *same* type convolutions.

Training and Error Analysis

For each network, only the last block of layers have been trained and the other blocks' weights have been kept freezed, pretrained on the ImageNet dataset. This is a common technique known as *transfer learning* [94], which allows both to make the training faster and to avoid overfitting. Indeed the transfer learning technique simulates in such a sense a much more large datasets, returning nets basically trained for very general tasks which can then be specialized on the addressed problem. Here the results obtained for the proposed neural networks are summarized. The optimization has been performed through The AdAm algorithm for all the proposed architectures [69].

	Train		**Validation**		**Test**		
	Loss	**Accuracy**	**Loss**	**Accuracy**	**Loss**	**Accuracy**	**Epochs**
ResNet50	0.0014	100%	0.027	96%	0.049	93.66%	14
ResNet152	0.0081	100%	0.77	78%	0.67	68%	31
VGG16	0.00049	100%	0.42	90.49%	0.36	91.22%	31
ResInception	0.55	75.31%	0.55	74.14%	0.59	72.20%	31

Table 3.2: Loss and Accuracy metrics for the CNNs

As it can be seen, the Residual Net 50 is the most performing architecture. The Residual Inception seems to struggle to get good results even in the train dataset.
The Residual Net 50 architecture is clearly the most promising one, even if the divergence among Train and Validation sets suggests that it may be some overfitting which we will have to deal with.
That said, the most important thing is that all the networks tested have shown great difficulties to correctly classify those kind of images which we called *tricky*, and this confirms that a peculiar work has to be done in order to have a production model.

Modified Residual Net 50 and final results

For the final version of Model 1 two parallel directions have been pursued. Well known augmentation techniques have been implemented. This kind of methodology belongs to a family of the more general preprocessing techniques which allow to *artificially increase* the size of the dataset by allowing to generate from each image a new one obtained through some transformations. These kind of transformations range from zoom intensity to shear range and horizontal flips. In this way the algorithm learns that the same object - let's say - rotated is in fact the same object, belonging to the same class, so to mitigate the overfitting problems.
At the same time some more layers have been added to the Residual Net 50 architecture, so to better adapt it to our task. In particular, after a *Flatten Layer*, we added a sequence of three blocks composed by respectively a Batch Normalization Layer, a *Dense Layer* with a *ReLu* activation function and a *Dropout Layer* [111] with 0.5 probability. The Dense Layers took respectively 256, 128 and 64 neurons. Then we applied a final Batch Normalization Layer before the prediction one, for which we chose a logistic function.

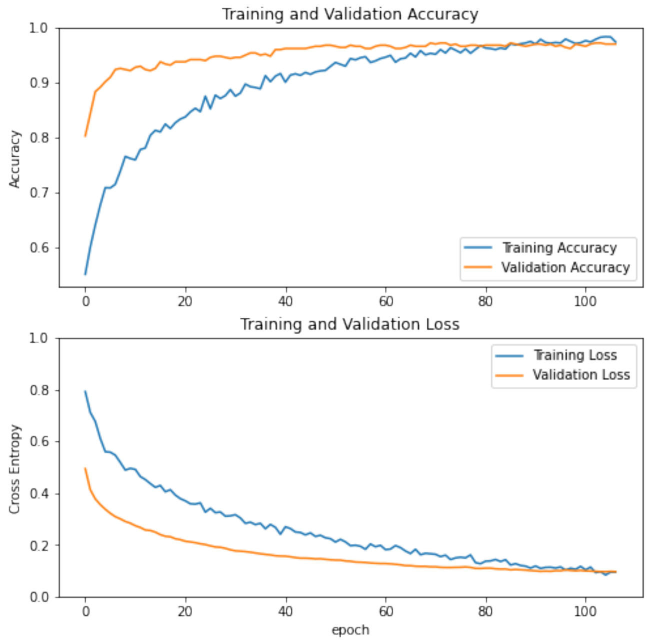

Figure 3.3: Model 1: training

With these last improvements, Model 1 reached very satisfying results in term of metrics: Besides the metrics, there are different perspectives from which the goodness of a model can be evaluated, especially for effective real life business implementations. A particular example in this sense will be discussed in the next section.

Figure 3.4: Misclassified images from the test set.

Confusion Matrix						
	Train		Validation		Test	
No Phone (T)	99.61%	0.39%	94.30%	5.70%	95.07%	4.93%
Phone (T)	0.28%	99.72%	1.32%	98.68%	3.51%	96.49%
	No Phone (P)	Phone (P)	No Phone (P)	Phone (P)	No Phone (P)	Phone (P)

Table 3.3: Confusion Matrices for the Train, Validation and Test datasets. Where T and P respectively stay for *True* and *Predicted*. The overall accuracies are 99.57%, 96.98% and 95.95% for Train, Validation and Test.

Probability distributions and possible thresholds

In order to further investigate the quality of our model, i.e. how much the model is robust with respect to its predictions, the distribution of the probabilities that our architecture assigns to the images are analyzed.
Across the train, validation and test datasets the probabilities occurrences are quite concentrated to the extreme probabilities for the two classes, and at the same time the majority of error probabilities are sparse on more central values.

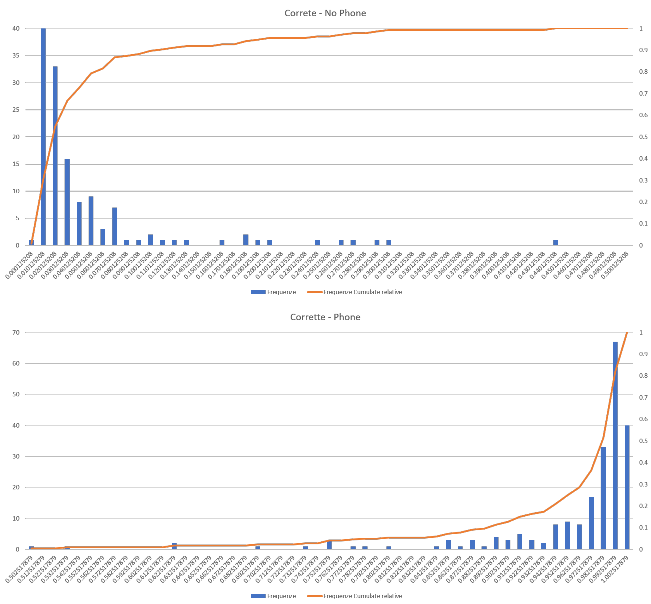

Figure 3.5: Model 1: correctly predicted probabilities distribution for Test set

The probability occurrences distribution is particularly good for a company which aims to face this kind of problem, because it allows to effectively choose two probability thresholds so that within this range the ambiguous images can be collected. Indeed the probability we obtained are particularly well suited to construct an a priori ambiguous range for practical implementation, to further increase the model precision. To highlight this applied perspective three different thresholds have been chosen heuristically for each of the two classes, i.e. three different maximum probability values to be classified as No Phone and three different minimum probability values to be classified as Phone, and analyzed how recall and precision vary according to the ambiguity range.

	Treshold No Phone: 10%	**No Phone: 20%**	**No Phone: 25%**
Treshold Phone: 70%	Precision: 99.81% Recall: 99.53% Ambiguous: 2.25%	Precision: 99.81% Recall: 99.53% Ambiguous: 0.81%	Precision: 99.81% Recall: 99.53% Ambiguous: 0.56%
Treshold Phone: 80%	Precision: 99.91% Recall: 99.25% Ambiguous: 2.44%	Precision: 99.91% Recall: 99.25% Ambiguous: 1%	Precision: 99.91% Recall: 99.25% Ambiguous: 0.72%
Treshold Phone: 90%	Precision: 99.90% Recall: 97% Ambiguous: 3.59%	Precision: 99.90% Recall: 97% Ambiguous: 2.15%	Precision: 99.90% Recall: 97% Ambiguous: 1.87%

Table 3.4: Training Set tresholds analysis

	Treshold No Phone: 10%	**No Phone: 20%**	**No Phone: 25%**
Treshold Phone: 70%	Precision: 97.07% Recall: 98.03% Ambiguous: 8.16%	Precision: 97.07% Recall: 98.03% Ambiguous: 2.04%	Precision: 97.07% Recall: 98.03% Ambiguous: 0.82%
Treshold Phone: 80%	Precision: 97.36% Recall: 97.04% Ambiguous: 8.98%	Precision: 97.36% Recall: 97.04% Ambiguous: 2.86%	Precision: 97.36% Recall: 97.04% Ambiguous: 1.63%
Treshold Phone: 90%	Precision: 97.13% Recall: 89.14% Ambiguous: 13.88%	Precision: 97.13% Recall: 89.14% Ambiguous: 7.76%	Precision: 97.13% Recall: 89.14% Ambiguous: 6.53%

Table 3.5: Validation Set tresholds analysis

	Treshold No Phone: 10%	**No Phone: 20%**	**No Phone: 25%**
Treshold Phone: 70%	Precision: 97.73% Recall: 94.30% Ambiguous: 8.38%	Precision: 97.73% Recall: 94.30% Ambiguous: 5.68%	Precision: 97.73% Recall: 94.30% Ambiguous: 5.14%
Treshold Phone: 80%	Precision: 98.12% Recall: 91.67% Ambiguous: 10.27%	Precision: 98.12% Recall: 91.67% Ambiguous: 7.57%	Precision: 98.12% Recall: 91.67% Ambiguous: 7.03%
Treshold Phone: 90%	Precision: 98.98% Recall: 85.53% Ambiguous: 14.59%	Precision: 98.98% Recall: 85.53% Ambiguous: 11.89%	Precision: 98.98% Recall: 85.53% Ambiguous: 11.53%

Table 3.6: Test Set tresholds analysis

If a balance among precision and recall is considered and at the same time the company wants to keep the ambiguous observations as low as possible, the combination $(70\%, 25\%)$ is probably the best. But given the business context for which this kind of analysis has been performed, there could be more adapted criteria. Without going into a business analysis, we can for example consider that, being this one a new policy and in a certain sense a new area of business for the company, in the first times it will be not yet a large customers base. Then it could be better for the company to prefer a larger precision. In this way the product may guarantee some degree of customer satisfaction, and even with relatively large percentages of ambiguous classified images will still remain manageable given the customer base. With the business increasing, then new images could be collected and use to refine the algorithm, so to gain with time higher robustness. Then this company may prefer a very conservative couple of thresholds as $(90\%, 10\%)$ based on this kind of cryterium.

3.2.2 Model 2

As already said, Model 2 aims at classifying smartphone images in order to distinguish broken or damaged smartphones from the unbroken ones. The detailed last version of the dataset used for Model 2 follows:

	Train		**Validation**		**Test**	
	Broken	**Ubroken**	**Broken**	**Ubroken**	**Broken**	**Ubroken**
Percentage	41.63%	58.47%	56.60%	43.40%	81.73%	18.27%
Number	438	614	150	115	170	38

Table 3.7: Final Datasets

This last training has been done through the modified Residual Net 50 and in this final version also augmentation techniques have been used, choosing the same parameters already proposed for Model 1.

Confusion Matrix						
	Train		Validation		Test	
Broken (T)	99.77%	0.23%	99.33%	0.67%	98.82%	1.18%
Unbroken (T)	0.00%	100%	0.00%	100%	0.00%	100%
	Broken (P)	Unbroken (P)	Broken (P)	Unbroken (P)	Broken (P)	Unbroken (P)

Table 3.8: Confusion Matrices for the Train, Validation and Test datasets. Where T and P respectively stay for *True* and *Predicted*. The overall accuracies are 99.90%, 99.62% and 99.04% for Train, Validation and Test.

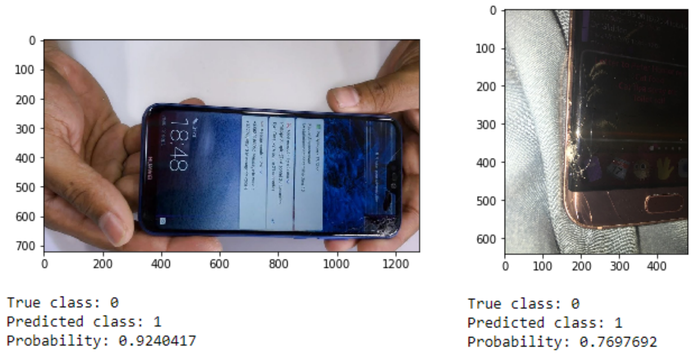

Figure 3.6: Mislcassified images from the test set.

As it can be seen, the results are very good. In order to assess the robustness of a so apparently good model we ran the same kind of probability occurrences analysis as done for Model 1. In the Model 2 case, even more concentrated distributions for the predicted probabilities have been obtained.

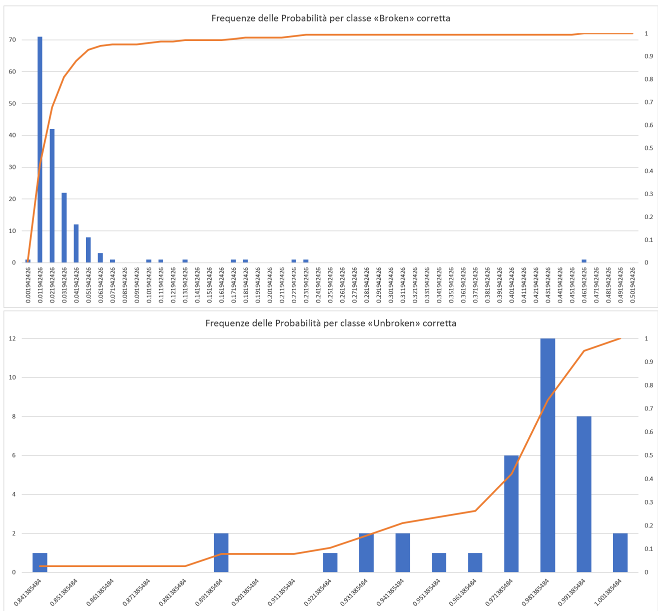

Figure 3.7: Model 2: correctly predicted probabilities distribution for Test set

As for Model 1, this kind of analysis reveals how strong and safe the model predictions are, and allows to possibly construct a range of probabilities within the predictions are considered ambiguous, which could be useful for business applications as already discussed for Model 1. The probability occurrences are strongly concentrated in the extreme probability values, and the few errors are quite sparsed in more soft values. With few errors, the kind of threshold analysis proposed for Model 1 does not really make sense. The thresholds to construct the ambiguous probability range can in this case be chosen freely by the company uniquely based on its needs and business ideas.

Grad-Cam analysis

Grad-Cam is an algorithm that allows to evaluate with a good approximation what the machine has learned [106]. In particular, for each image it highlights through a heat-map the pixels which have most stressed the neural network and influenced it in making the prediction. This type of control allows to check whether some unwanted correlation occurs by unfortunate coincidence in the selection of datasets and in particular within the classes. A classic example in that case could be that the machine has learned that there is a high probability for a smartphone to be broken if fingers are holding it. This kind of problems, which can occur even in much more unpredictable and sneaky patterns than the example provided here, are really dangerous and often discriminate between what model works in laboratory and what in real business applications. Here some instances from the Grad-Cam algorithm based analysis are shown:

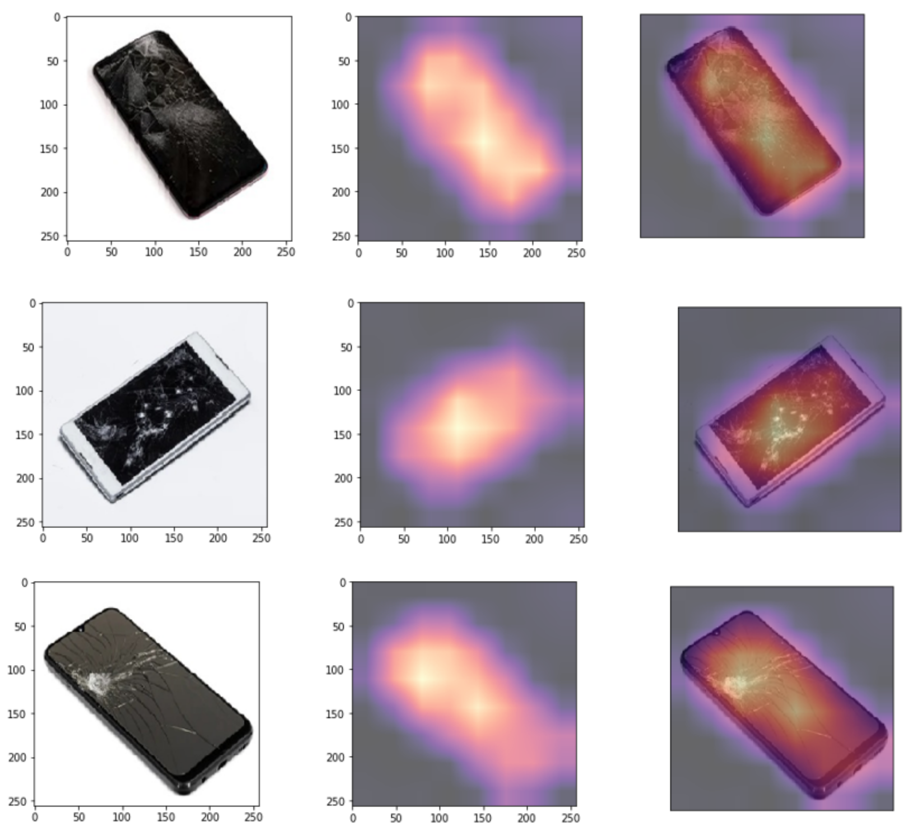

Figure 3.8: Instances from the GRAD-CAM algorithm.

As it can be noted, Model 2 worked well, and the Net *looked* almost always where it has to look. We analized all the Grad-Cam images, and no spurious correlations have been spotted, and Model 2 always concentrate where it has to.

3.3 Conclusion

In this paper we provided an effective solution to a concrete business problem which required a Deep Learning approach. We split the task in two different problems, Model 1 and Model 2. For Model 1, first we selected and tested some well known architecture on an initial dataset. Then we chose the most performing architecture and through some error analysis we improved the initial dataset, and added some further layers to te architecture to better fit our problem. The results were extremely good, especially for Model 2.

Chapter 4

Feature selection variable based on multicollinearity.

This paper introduces a novel approach to perform variables selection in predictive models. Starting from a generalization of the conditioning number, a new index is derived to detect multicollinearity and the select the most relevant variables in predictive modeling. The contribution of this paper is two folds: in terms of methodological development, the paper introduces a new approach for future selection; in terms of computational innovation a new algorithm is provided to implement our proposal. Empirical evidence achieved on simulated and real data set.

4.1 Introduction

Variables and feature selection have become the focus of much research in terms of methodological and computational development. The objective of variables selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data (see e.g. Guyon et al. 2003) [58].

Collinearity, or excessive correlation among explanatory variables, can affect the identification of an optimal set of explanatory variables for a statistical model, producing inconsistent results, first of all providing a poor precision in the estimates due to high variances in the parameters. More precisely, parameters estimates may be unstable, standard errors on estimates inflated and consequently inference statistics biased.

In variables selection is crucial to look at the intercorrelation or multi-collinearity: the existence of predictor variables that are (highly) correlated among themselves.

The problem of multicollinearity is well known in the literature as described in Farrar et al (1967) [47]. Multicollinearity can be a good criterion for selecting the best set of $n-k$ variables out of n possible regressive variables in a linear model, taking into account stability and forecasting.

The methods proposed in the literature vary according to the perspective from which the problem is approached, whether statistical, algebraic or numerical. There are well-known tests for multicollinearity, as well as approaches based on indices of algebraic and numerical sources, applied relying on heuristically accepted threshold values.
The VIF (Variance Inflation Factor) quantifies the severity of multicollinearity in an ordinary least squares regression analysis, providing an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.
VIF calculations are straightforward and easily comprehensible: the higher the value, the higher the collinearity. A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable against all other explanatory variables:

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{4.1}$$

where the VIF for variable j is the reciprocal of the inverse of "R^2 from the regression. A VIF is calculated for each explanatory variable and those with high values are removed. The definition of high is somewhat arbitrary but values in the range of $5 - 10$ are commonly used. Removing individual variables with high VIF values is insufficient in the initial comparison using the full set of explanatory variables. The VIF values will change after each variable is removed. Accordingly, a more thorough implementation of the VIF function is to use a stepwise approach until all VIF values are below a desired threshold. For example, using the full set of explanatory variables, calculate a VIF for each variable, remove the variable with the single highest value, recalculate all VIF values with the new set of variables, remove the variable with the next highest value, and so on, until all values are below the threshold. The statistical literature offers several quantifications of collinearity, with the most common being the pairwise correlation coefficient, the condition index (the square root of the ratio of each eigenvalue to the square root of the smallest eigenvalue of X), the variance inflation factor (VIF) and its generalized version [51], and the variance decomposition proportions (VD, which gives more specific information on the eigenvectors' contribution to collinearity: Belsley et al. 1980 [17], Brauner and Shacham 1998) [24]. There are also approaches that estimate a single value to describe the degree of collinearity in the full dataset (variable set indices). Most commonly used are the determinant of the correlation matrix and the condition number.
The condition number, which is the ratio among the maximum and the minimum square root eigenvalue (i.e. the singular value) of a generic correlation matrix quantifies multicollinearity. An eigenvalue close to zero indicates that in the original data some variable is approximately a linear combination of some of the other ones. Indeed a high value in the conditioning number highlights multicollinearity. A solid comparison between the condition number and the VIF approach can be found in Salmerón et al. (2017) [103].
The conditioning number generally fails to treat some cases that may arise in a variable selection problem. Indeed a proper example can be shown in which the conditioning number struggles to capture the presence of more than one multicollinear vector.

Suppose for example to have two N rank matrices which are equally "close" to the space of $N-1$ rank matrices: the condition number does not discriminate which of these two matrices is "closer" to the space of $N-k$ rank matrices (with $k>1$).
The conditioning number does not take into account situations where more than one variable is approximately a linear combination of some other variables in the data. A proper example will be constructed in order to highlight this structural weakness.
The most useful class of indices depends on the complexity of the dataset. Variable-set indices are preferable when quickly checking for collinearity in datasets with large numbers of explanatory variables. Per-variable-indices give a more detailed picture of the number of variables involved and the degree of collinearity. Sometimes the per-variable-indices may indicate collinearity although the variable-set indices miss it. This paper proposes a new contribution in collinearity detection and it is organized as follows: Section 2 describes our methodological proposal; Section 3 focus on the choice of the parameter k and then Section 4 shows an application both on simulated data and real data at hand. Conclusions and further idea of research are reported in Section 5.

4.2 Our proposal: A multicollinearity detection index

Consider the following $(T \times 1)$ random column vectors:

$$x_1 \sim N(0,1) \times 10^{(-2)}, x_2 \sim N(0,1) \times 10^{(-2)},$$

$$x_3 = x_1 + \epsilon_1 \times 10^{(-5)}, x_4 = x_1 + \epsilon_2 \times 10^{(-4)}$$

with ϵ_1 and ϵ_2 standard normal random vectors.
Then derive the following sub-datasets:

$$X = [x_1 \ x_2 \ x_3]$$

$$X' = [x_1 \ x_2 \ x_4]$$

$$Y = [x_1 \ x_3 \ x_4]$$

Detecting multicollinearity, the best dataset is X', followed by X and Y. The condition number works well to capture the relationship between X' and X; X' and Y, but it fails to detect X and Y. In order to overcome this weakness, a generalization of the condition number is proposed, which accounts for the last k singular values corresponding to k possibly multicollinear variables [48].
Let X be a $T \times N$ matrix, with $T > N$ normalized column vectors, such that $X_{i,j} \sim D(0,1)$, $\forall i \in [1,T] \cap \mathbf{N}$ and $\forall j \in [1,N] \cap \mathbf{N}$. Let C be the correlation matrix:

$$C = X'X.$$

Let $L = \{\lambda_n\}_n$ with $1 \leq n \leq N$ where λ_n are the eigenvalues of C, i.e. $\lambda_n \geq \lambda_{n+1}$; the singular values of X are $\sigma_n = \sqrt{\lambda_n}$, $\forall n = 1, ..., N$. Define $L_k = \{\lambda_k\}$ with $1 \leq k < N$, such that $\lambda_k \geq \lambda_{k+1}$.
Our index is:

$$s_k(X) = \frac{\sigma_1}{\sigma_N}, \ \textit{with } L_k \textit{ empty set}$$

and:

$$\frac{\sigma_1}{\left(\prod_{i=k}^{N} \sigma_i\right)^{\frac{1}{k}}}, \ \textit{otherwise}$$

Our proposal satisfies the following properties:

1. The index is invariant with respect to the scale transformations, i.e. $\forall X \subset \mathbf{R}^{T \times N}$ with $T > N$ and $\forall \alpha \in \mathbf{R}$,

   $$s_k(\alpha X) = s_k(X)$$

   Proof: A linear scale transformation in the original matrix is reflected on the singular values:

   $$s_k(\alpha X) = \frac{\alpha \sigma_1}{\left(\prod_{i=k}^{N} \alpha \sigma_i\right)^{\frac{1}{k}}}$$

   then:

   $$s_k(\alpha X) = \frac{\alpha \sigma_1}{\left(\alpha^{k\frac{1}{k}} \prod_{i=k}^{N} \sigma_i\right)^{\frac{1}{k}}} = \frac{\sigma_1}{\left(\prod_{i=k}^{N} \sigma_i\right)^{\frac{1}{k}}}$$

2. $s_k(X)$ is non-increasing with respect to k, i.e. $1 \leq s_{k+1}(X) \leq s_k(X)$.

   Proof The index depends on k through the denominator, which is constructed as a geometric mean of the last k terms. Since the singular values are ranked in decreasing order, then the property is naturally satisfied. Notice that $1 \leq s_k(X) \leq k(X)$, where $k(X)$ is the conditioning number.

4.3 The choice of k

The index depends on a parameter k which reflects the dimensions considered at risk of loss (i.e. how many variables are over a threshold of redundancy in explaining the variability of the data).
From a different perspective, the choice of k means assessing a range of proximity to zero regarding the eigenvalues, i.e. establishing how close to zero an eigenvalue must be to consider the matrix potentially singular. This perspective converts the solution of the problem by resorting to a numerical approach.
Several methods can be proposed whose effectiveness may differ depending on the scope of the index and the practical meaning of the variables involved. For example, considering redundant all eigenvalues exceeding a treshold or choosing a threshold for the singular

values. Depending on the context, the choice of the threshold value can be linked to exogenous variables.

In this contribution, a statistical asymptotic approach was chosen, which relies on well-known results on Random Matrix Theory [81]. We underline how k can be linked to Random Matrix Theory [72], [71], [121], [29], [96].

Let X be a $T \times N$ matrix, with T number of observations and N number of standardized variables; let E be the empirical correlation Matrix:

$$E = X'X$$

Using RMT, E is unbiased with respect to the true correlation matrix C.

λ_i for $i = 1, ..., N$ are the eigenvalues of a correlation matrix E, with Λ defined as:

$$\Lambda_\lambda = \{\lambda_i : \lambda_i < \lambda\}$$

and:

$$\mu(\lambda) = \frac{1}{N} \#(\Lambda_\lambda)$$

with # the counting measure; the eigenvalue density function for the correlation Matrix E is given by:

$$\rho_E(\lambda) = \frac{d\mu(\lambda)}{d\lambda}.$$

It can be shown that for $\lim_{T,N \to \infty} T/N = Q$ with $Q > 1$, if the correlation matrix E was generated by a $T \times N$ matrix whose entries were i.i.d. and standardized, then [?]:

$$\rho_E(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}$$

where $\lambda_{min}/\lambda_{max}$ is:

$$\lambda_{min}^{max} = \left(1 \pm \sqrt{\frac{1}{Q}}\right)^2$$

In this paper, λ_{min} represents a threshold beyond that the eigenvalues of the correlation matrix will be considered closed to zero. It will be shown in the empirical analysis that this approach allowed to capture a good estimate of the correct number of correlated variables considering all the possible cases.

4.4 Empirical analysis - Simulated datasets

The simulated dataset is made up of $i = 100$ observations and $j = 10$ variables.

X is a data matrix created as follows:

$$\begin{cases} X_{i,j} \sim N(0,1), j = 1, ..., 6 \\ X_{i,j} = X_{i,1} + X_{i,2} + 10^{-j+6}\epsilon_{i,j}, j = 7, ..., 10 \end{cases}$$

with $\epsilon_{i,j} \sim N(0,1)$ and $i = 1, ..., 100$. 100 data sets with this structure has been simulated. The simulation data at hand has 4 multicollinear variables. In order to translate our methodological proposal described in Section 3, an algorithm has been implemented and it is composed of the following steps:

- Detect the number $\bar{k}$ of potentially multicollinear variables through the RMT approach.

- Single out all the possible combinations of $M \equiv N - \bar{k}$ random variables. Obviously, since they are potentially $\bar{k}$ redundant variables out of the initial N it is meaningful to deal just with $N - \bar{k}$ of them.

- Derive k as $k = min\{M-1, \bar{k}\}$, i.e. the maximum number of multicollinear variables which can occur in a combination of M variables. In particular, $k = \bar{k}$ if $N \geq 2\bar{k}+1$.

- Compute the index for all the possible $\frac{N!}{(N-\bar{k})!(\bar{k})!}$ combinations.

On the basis of the data at hand, $\lambda_{min} = 0.4675$.
For all the simulated matrices, the random matrix approach provides $\bar{k} = 4$. Fixing $k = 4$, 210 possible combinations of 6 variables are analyzed. In order to assess the validity of the RMT approach to detect multicollinear variables, the number of predicted redundant terms were computed for each possible combination. Figure 1 depicts the number of predicted multicollinear variables against the real number of multicollinear variables. As we can see, the RMT approach provides the correct number for each combination.

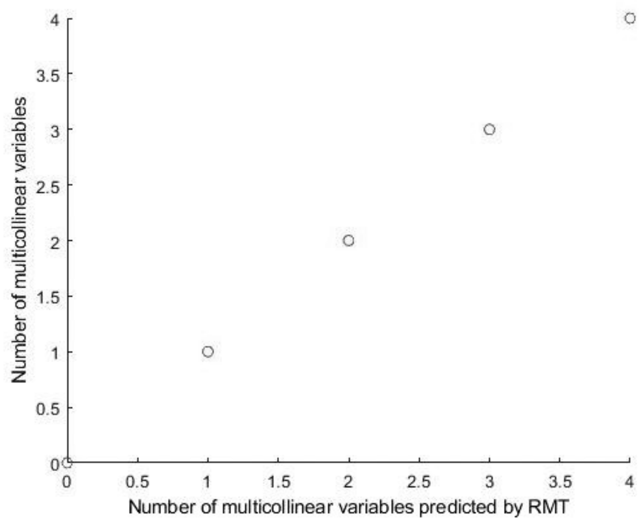

Figure 4.1: Index conditional distribution with respect to the correlated vectors number, across the 100 simulated datasets

The RMT method seems to work well on simulated data. It is interesting to show how the index is distributed across the 100simulated datasets changing the number of multicollinear variables. For each case the *less* multicollinear vectors have been chosen.

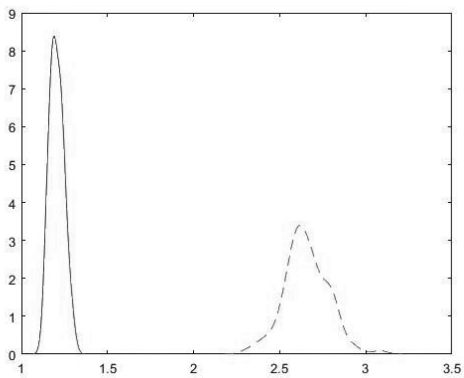

Figure 4.2: Conditional distributions with respect to the zero and one multicollinear variables case

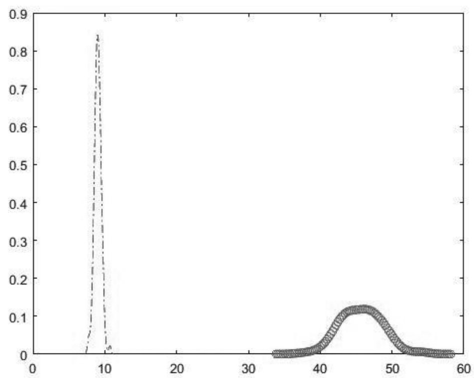

Figure 4.3: Conditional distributions with respect to the two and three multicollinear variables case

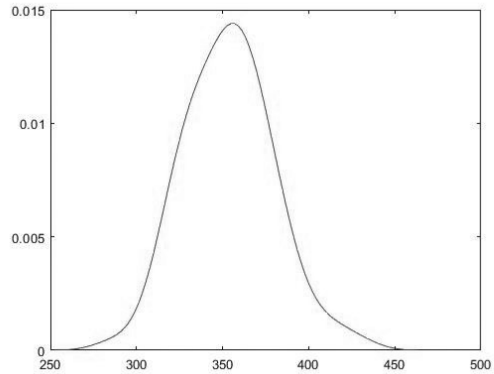

Figure 4.4: Conditional distributions with respect to the four multicollinear variables case

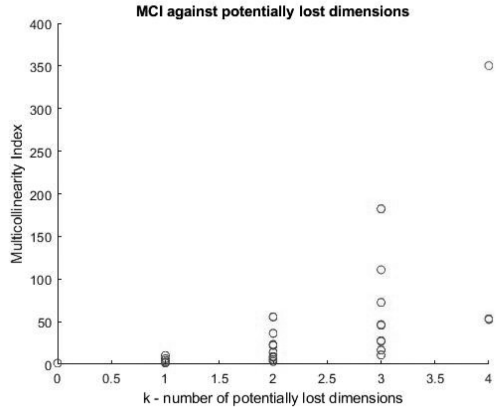

Figure 4.5: The index with fixed k is increasing with respect to the number of potentially lost dimensions predicted by RMT in each six variables combination

In the simulated data the index seems to well separate the different cases obtained according to the number of multicollinear variables, as it depicted from the conditional distributions.

	Mean	St. Dev.	Skewness	Kurtosis
X_0	1.21	0.04	0.18	2.83
X_1	2.66	0.11	0.15	2.88
X_2	9.04	0.45	0.24	2.76
X_3	45.74	2.76	0.24	2.47
X_4	353.92	25.44	0.17	2.33

Table 4.1: Conditional moments.

	μ	Σ	s
$X_{0,1}$	1.44	0.32	1.12
$X_{1,2}$	6.38	1.14	5.24
$X_{2,3}$	36.70	6.42	30.27
$X_{3,4}$	308.18	56.40	251.77

Table 4.2: Conditional distributions comparison.

Let X_i be one of the datasets, with $1 = 1, ..., 4$ the number of multicollinear vectors, let $m(X_i)$ be the mean of the index distribution conditioned on the dataset with i multicollinear vectors, and σ_i be its standard deviation. Then $\mu_{i,i+1} = |m(X_{i+1}) - m(X_i)|$ is the distance among the i and $i + 1$ conditional distributions first moments. Finally, $\Sigma_{i,i+1} = 2(\sigma_i + \sigma_{i+1})$ and $s_{i,i+1} = \mu_{i,i+1} - \Sigma_{i,i+1}$, provides a measure of separation between the different spaces. Figure 4 plots the index against the number of multicollinear variables for the six variables combination with respect to the first of the 100 simulated datasets. The index increases with respect to the number of multicollinear variables.

4.5 Empirical analysis - Inclusive Internet dataset

In this section the Inclusive Internet dataset is considered. The variables considered are: Electricity access, Literacy, Urban population percentage, Peace index, Democracy Index, Corruption Index, Relevant Content, Internet Quality, Infrastructure, Policy, Mobile Gender Gap, Trust and safety. The dataset is composed of 100 observations which account for the 96% of global GDP and for the 91% of global population. Running our algorithm, we obtain $\lambda_{min} = 0.5408$ and $k = \bar{k} = 5$. The index has been computed across all the combinations of 7 variables out of 12. For sake of comparison, we have derived: the predicted number of multicollinear variables using RMT for each combination, the conditioning number, VIFs euclidean norms and the proposed index, MCI. Figure 3 depicts the results.

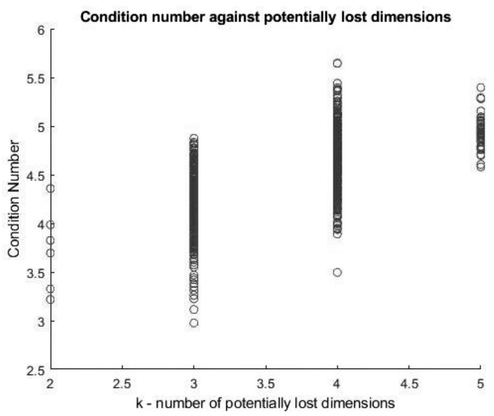

Figure 4.6: Conditioning number against predicted number of multicollinear variables

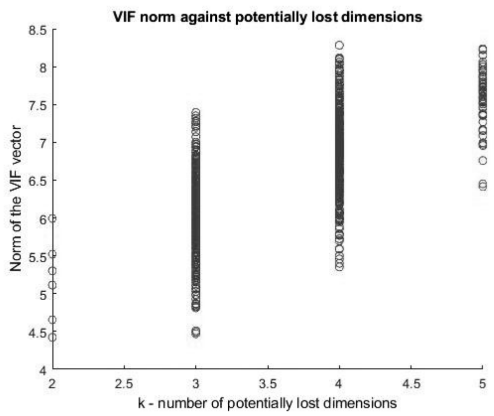

Figure 4.7: VIFs norms against predicted number of multicollinear variables

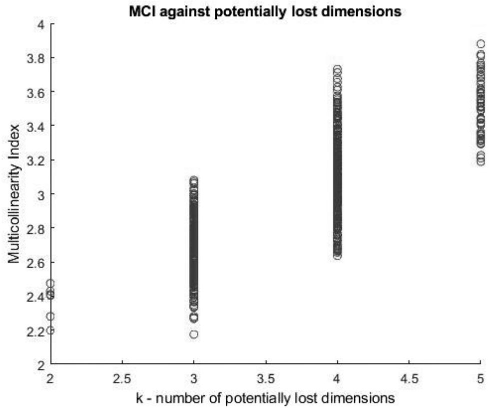

Figure 4.8: MCI against predicted number of multicollinear variables

It seems that RMT is a parsimonious criteria to identify the number of redundant vectors. Furthermore, the number of predicted multicollinear vectors show a direct relationship with respect to the VIFs euclidean norm and the conditioning number. More precisely, the proposed index shows as expected a stronger increasing relationship and the results obtained are concordant with VIF and the condition number.

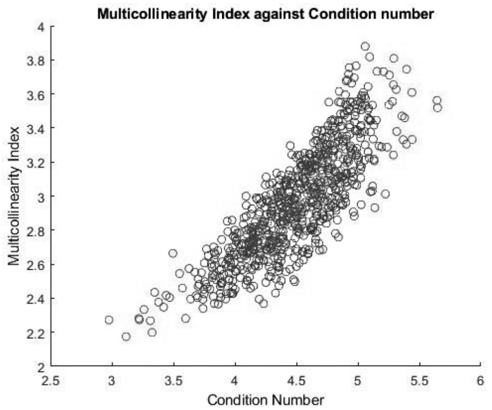

Figure 4.9: MCI against condition number

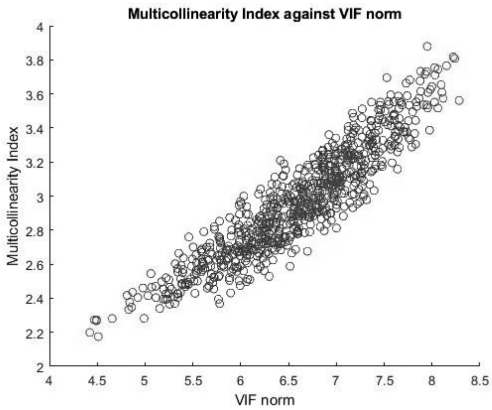

Figure 4.10: MCI against VIFs norms

4.6 Conclusions

In this paper a new approach is proposed to deal with multicollinearity.
The results obtained on simulated and real data underline that the new indicator seems a robust approach to detect correlation among variables, reducing collinearity and redundancy. The validity of the approach does not depend on the explained variable and it is self-consistent with respect to the application context.

List of Figures

List of Tables

Bibliography

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[2] Edward I Altman. Predicting financial distress of companies: revisiting the z-score and zeta® models. In *Handbook of research methods and applications in empirical finance*. Edward Elgar Publishing, 2013.

[3] Antonio Amendola, Dennis M Montagna, and Mario Maggi. Analysis of equity β components: New results and prospectives in a low β framework. *Journal of Economics and Financial Analysis*, 3(1):1–26, 2019.

[4] Andrew Ang, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang. High idiosyncratic volatility and low returns: International and further us evidence. *Journal of Financial Economics*, 91(1):1–23, 2009.

[5] Clifford S Asness, Andrea Frazzini, and Lasse H Pedersen. Leverage aversion and risk parity. *Financial Analysts Journal*, 68(1):47–59, 2012.

[6] Clifford S Asness, Andrea Frazzini, and Lasse H Pedersen. Low-risk investing without industry bets. *Financial Analysts Journal*, 70(4):24–41, 2014.

[7] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(6), 2008.

[8] Louis Bachelier. Théorie de la spéculation. In *Annales scientifiques de l'École normale supérieure*, volume 17, pages 21–86, 1900.

[9] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[10] David H Bailey and Marcos Lopez de Prado. The sharpe ratio efficient frontier. *Journal of Risk*, 15(2):13, 2012.

[11] Malcolm Baker, Brendan Bradley, and Ryan Taliaferro. The low-risk anomaly: A decomposition into micro and macro effects. *Financial Analysts Journal*, 70(2):43–58, 2014.

[12] Malcolm Baker, Mathias F Hoeyer, and Jeffrey Wurgler. The risk anomaly tradeoff of leverage. Technical report, National Bureau of Economic Research, 2016.

[13] Nardin L Baker and Robert A Haugen. Low risk stocks outperform within all observable markets of the world. *Available at SSRN 2055431*, 2012.

[14] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[15] Turan G Bali, Stephen J Brown, Scott Murray, and Yi Tang. A lottery-demand-based explanation of the beta anomaly. *Journal of Financial and Quantitative Analysis*, 52(6):2369–2397, 2017.

[16] Flavio Barboza, Herbert Kimura, and Edward Altman. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417, 2017.

[17] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

[18] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[19] Sreedhar T Bharath and Tyler Shumway. Forecasting default with the merton distance to default model. *The Review of Financial Studies*, 21(3):1339–1369, 2008.

[20] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

[21] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[22] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. In *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 1: Foundations of CCA and Equity Valuation*, pages 3–21. World Scientific, 2019.

[23] David Blitz. Agency-based asset pricing and the beta anomaly. *European Financial Management*, 20(4):770–801, 2014.

[24] Shacham M. Brauner, N. *An introduction to variable and feature selection.* 1998.

[25] L Breiman, J Friedman, R Olshen, and C Stone. Cart. *Classification and Regression Trees*, 1984.

[26] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[28] Axel Buchner and Niklas Wagner. The betting against beta anomaly: Fact or fiction? *Finance Research Letters*, 16:283–289, 2016.

[29] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

[30] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[31] Scott Cederburg and MICHAEL S O'DOHERTY. Does it pay to bet against beta? on the conditional performance of the beta anomaly. *The Journal of finance*, 71(2):737–774, 2016.

[32] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[33] Yves Choueifaty and Yves Coignard. Toward maximum diversification. *The Journal of Portfolio Management*, 35(1):40–51, 2008.

[34] George M Constantinides and Anastasios G Malliaris. Portfolio theory. *Handbooks in operations research and management science*, 9:1–30, 1995.

[35] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[36] David Cowan and Sam Wilderman. Re-thinking risk: what the beta puzzle tells us about investing. *GMO White Paper*, pages 1–18, 2011.

[37] Matej Črepinšek, Shih-Hsi Liu, and Marjan Mernik. Exploration and exploitation in evolutionary algorithms: A survey. *ACM computing surveys (CSUR)*, 45(3):1–33, 2013.

[38] Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[39] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953, 2009.

[40] Victor DeMiguel, Alberto Martin-Utrera, and Francisco J Nogales. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. *Journal of Banking & Finance*, 37(8):3018–3034, 2013.

[41] David J Disatnik and Simon Benninga. Shrinking the covariance matrix. *The Journal of Portfolio Management*, 33(4):55–63, 2007.

[42] Cao Vu Dung, Hidehiko Sekiya, Suichi Hirano, Takayuki Okatani, and Chitoshi Miki. A vision-based method for crack detection in gusset plate welded joints of steel bridges using deep convolutional neural networks. *Automation in Construction*, 102:217–229, 2019.

[43] Eugene F Fama and Kenneth R French. The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465, 1992.

[44] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

[45] Eugene F Fama and Kenneth R French. *Common risk factors in the returns on stocks and bonds*. University of Chicago Press Chicago, 2021.

[46] Eugene F Fama and James D MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636, 1973.

[47] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

[48] Silvia Figini, Mario Maggi, and Pierpaolo Uberti. The market rank indicator to detect financial distress. *Econometrics and Statistics*, 14:63–73, 2020.

[49] Irving Fisher. *The nature of capital and income*. Macmillan and Cie, 1906.

[50] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[51] John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

[52] Wenjiang Fu and Keith Knight. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

[53] Andrea Gamba and Francesco A Rossi. A three-moment based portfolio selection model. *Rivista di matematica per le scienze economiche e sociali*, 21(1):25–48, 1998.

[54] Stefano Giglio and Dacheng Xiu. Asset pricing with omitted factors. *Journal of Political Economy*, 129(7):1947–1990, 2021.

[55] Jeremiah Green, John RM Hand, and X Frank Zhang. The supraview of return predictive signals. *Review of Accounting Studies*, 18(3):692–730, 2013.

[56] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[57] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021.

[58] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[59] Peter E Hart, David G Stork, and Richard O Duda. *Pattern classification*. Wiley Hoboken, 2000.

[60] Campbell R Harvey and Yan Liu. Luck versus skill and factor selection. *The Fama Portfolio: Selected Papers of Eugene F. Fama*, 2017.

[61] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[63] WEm Henley and David J Hand. Ak-nearest-neighbour classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(1):77–95, 1996.

[64] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[65] Thomas Idzorek. A step-by-step guide to the black-litterman model: Incorporating user-specified confidence levels. In *Forecasting expected returns in the financial markets*, pages 17–38. Elsevier, 2007.

[66] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[67] James Max Kanter. Color crack: Identifying cracks in glass. *dated Dec*, 9, 2014.

[68] Bryan Kelly and Seth Pruitt. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316, 2015.

[69] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[70] Vaclav Kozeny. Genetic algorithms for credit scoring: Alternative fitness function performance comparison. *Expert Systems with applications*, 42(6):2998–3004, 2015.

[71] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Physical review letters*, 83(7):1467, 1999.

[72] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03):391–397, 2000.

[73] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[74] Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[75] Shengyuan Li and Xuefeng Zhao. Convolutional neural networks-based crack detection for real concrete surface. In *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018*, volume 10598, page 105983V. International Society for Optics and Photonics, 2018.

[76] Shih-Hsi Liu, Marjan Mernik, and Barrett R Bryant. To explore or to exploit: An entropy-driven approach for evolutionary algorithms. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 13(3-4):185–206, 2009.

[77] Zhenqing Liu, Yiwen Cao, Yize Wang, and Wei Wang. Computer vision-based concrete crack detection using u-net fully convolutional networks. *Automation in Construction*, 104:129–139, 2019.

[78] Andrew W Lo. The statistics of sharpe ratios. *Financial analysts journal*, 58(4):36–52, 2002.

[79] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

[80] Semyon Malamud, Andreas Schrimpf, Teng Andrea Xu, Giuseppe Matera, and Antoine Didisheim. Benign autoencoders. *arXiv preprint arXiv:2210.00637*, 2022.

[81] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[82] Harry Markowitz. The utility of wealth. *Journal of political Economy*, 60(2):151–158, 1952.

[83] Jacob Marschak. Money and the theory of assets. *Econometrica, Journal of the Econometric Society*, pages 311–325, 1938.

[84] Thomas Mazzoni. *A First Course in Quantitative Finance.* Cambridge University Press, 2018.

[85] Elmar Mertens. Comments on variance of the iid estimator in lo (2002). Technical report, Technical report, Working Paper University of Basel ..., 2002.

[86] Robert C Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, 29(2):449–470, 1974.

[87] Richard O Michaud and Robert Michaud. Estimation error and portfolio optimization: a resampling solution. *Available at SSRN 2658657*, 2007.

[88] Melanie Mitchell. *An introduction to genetic algorithms.* MIT press, 1998.

[89] R Mitchell, J Michalski, and T Carbonell. An artificial intelligence approach. *Machine learning. Berlin, Heidelberg: Springer*, 2013.

[90] John M Mulvey and Han Liu. Identifying economic regimes: Reducing downside risks for university endowments and foundations. *The Journal of Portfolio Management*, 43(1):100–108, 2016.

[91] Marcus D Odom and Ramesh Sharda. A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks*, pages 163–168. IEEE, 1990.

[92] John Douglas JD Opdyke. Comparing sharpe ratios: so where are the p-values? *Journal of Asset Management*, 8(5):308–336, 2007.

[93] Szilárd Pafka, Marc Potters, and Imre Kondor. Exponential weighting and random-matrix-theory-based filtering of financial covariance matrices for portfolio optimization. *arXiv preprint cond-mat/0402573*, 2004.

[94] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[95] Zdzisław Pawlak. Rough set approach to knowledge-based decision support. *European journal of operational research*, 99(1):48–57, 1997.

[96] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A Nunes Amaral, Thomas Guhr, and H Eugene Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.

[97] C Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[98] Brian D Ripley. *Pattern recognition and neural networks.* Cambridge university press, 2007.

[99] Richard Roll. A critique of the asset pricing theory's tests part i: On past and potential testability of the theory. *Journal of financial economics*, 4(2):129–176, 1977.

[100] Richard Roll and Stephen A Ross. An empirical investigation of the arbitrage pricing theory. *The journal of finance*, 35(5):1073–1103, 1980.

[101] Thierry Roncalli. *Introduction to risk parity and budgeting.* CRC Press, 2013.

[102] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[103] Román Salmerón, CB García, and J García. Variance inflation factor and condition number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12):2365–2384, 2018.

[104] Fritz W Scholz and Michael A Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

[105] Shruti Sehgal, Harpreet Singh, Mohit Agarwal, V Bhasker, et al. Data analysis using principal component analysis. In *2014 international conference on medical imaging, m-health and emerging communication systems (MedCom)*, pages 45–48. IEEE, 2014.

[106] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[107] William F Sharpe. Efficient capital markets: a review of theory and empirical work: discussion. *The Journal of Finance*, 25(2):418–420, 1970.

[108] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

[109] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[110] QingJun Song, HaiYan Jiang, and Jing Liu. Feature selection based on fda and f-score for multi-class classification. *Expert Systems with Applications*, 81:22–27, 2017.

[111] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[112] Lyn C Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172, 2000.

[113] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[114] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[115] Maria-Laura Torrente and Pierpaolo Uberti. Connectedness versus diversification: two sides of the same coin. *Mathematics and Financial Economics*, 15(3):639–655, 2021.

[116] Chih-Fong Tsai, Yu-Feng Hsu, and David C Yen. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 24:977–984, 2014.

[117] Chih-Fong Tsai and Jhen-Wei Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34(4):2639–2649, 2008.

[118] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[119] Jue Wang, Kun Guo, and Shouyang Wang. Rough set and tabu search based feature selection for credit scoring. *Procedia Computer Science*, 1(1):2425–2432, 2010.

[120] Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2008.

[121] Pavel Yaskov. A short proof of the marchenko–pastur theorem. *Comptes Rendus Mathematique*, 354(3):319–322, 2016.

[122] Giovanni Maria Zambruno and Pierpaolo Uberti. Higher moments asset allocation.

[123] Qinghua Zhang, Qin Xie, and Guoyin Wang. A survey on rough set theory and its applications. *CAAI Transactions on Intelligence Technology*, 1(4):323–333, 2016.

[124] Long Zhao. Reappraise the maximum-sharpe portfolio. *Available at SSRN 3497169*, 2019.

[125] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Acknowledgements

A Mamma e Papà. Mi basta sentirvi parlare nella stanza accanto per essere selvaggiamente felice.