Optimizing investment allocation: a combination of Logistic Regression and Markowitz model.

Authors: Fabiano Ramos Alves[a,b] ; Michael David de Souza Dutra[a,c,d]

[a] Affiliation: Federal University of Goiás. Av. Esperança, s/n - Chácaras de Recreio Samambaia, Goiânia - GO, 74690-900 – Brazil.

[b] Email: fabianoramos@discente.ufg.br

[c] Email: michaeldavid.dutra@polymtl.ca

[d] corresponding author.

ABSTRACT

One of the biggest challenges in quantitative finance is the efficient allocation of capital. Thus, in this study, a two-step methodology was proposed, in which a combination of logistic regression and Markowitz model was performed to determine optimized portfolios. In this context, in the first step, fundamentalist indicators were used as inputs to the logistic regression model in order to select the assets with the highest return potential. Thus, in the second step, the portfolio with these assets was balanced according to the covariances between the assets, so that the risk of the portfolio was as low as possible. The database consisted of the financial statements of 77 assets listed on the Ibovespa corresponding to the period from 2012 to 2020. To analyze the results, the model was implemented during the last 5 quarters and compared to three other approaches: 1 - only the implementation of logistic regression and the assignment of equal weights to the selected assets (LR + 1/N); 2 - traditional Markowitz model (mean-variance); 3 - Ibovespa index. The results showed that the combination of machine learning and optimization models can provide more efficient portfolios, and the cumulative return obtained by this investment strategy was 13.1% higher than the Ibovespa index, while the portfolio volatility was 3.1% lower in the same period.

Keywords: Portfolio Selection; Logistic Regression; Fundamental Analysis; Mean-Variance Model; Risk Management.

1 – INTRODUCTION

Tools to help decision makers about investments are crucial for the digital transformation (DUTRA, 2021). The construction and management of an investment

portfolio consist in allocating the capital in different financial assets, in certain proportions, in order to achieve the best results in terms of return and risk (WANG, LI, ZHANG and LIU, 2020). According to Freitas, Souza and Almeida (2009), the selection of investments is a central problem within the theory of finance or for economic analyses in general (DE SOUZA DUTRA et al., 2019). Corroborating this, In this context, portfolio optimization has become one of the main topics when it comes to research related to the investment decision-making process (CHEN, ZHANG, MEHLAWAT and JIA, 2020).This process, which aims to find the best portfolio in the return/risk ratio, can be divided into two steps: first, selecting assets with higher return potential; and second, determining the optimal composition of these assets in an investment portfolio.

Portfolio optimization consists in maximizing the return on investment given a certain level of risk, or reducing the risk, given a certain level of return. Thus, this optimization is related to the characteristics of the assets themselves, and particularly to the variances of these assets and their covariances among themselves. This fundamental idea was developed by Markowitz (1952), who developed a mathematical model involving a quadratic programming problem with linear constraints, called Mean-Variance model (FÁVERO and BELFIORE, 2013).

However, the Markowitz model depends on the accuracy of the statistical parameters used. In this sense, the potential error grows as the size of the set of assets analyzed increases. Thus, the preliminary selection step of assets with higher quality becomes a crucial factor in determining an optimal portfolio (CHEN, ZHANG, MEHLAWAT, & JIA, 2020; WANG, LI, ZHANG, & LIU, 2019). In this context, performing this selection of assets with higher return potentials proves to be a promising and at the same time challenging approach, since the stock market has a dynamic, complex and non-linear behavior (PAIVA, CARDOSO, HANAOKA and DUARTE, 2018).

Recent studies have addressed the use of different asset pre-selection techniques in conjunction with the classical model proposed by Markowitz. It stands out mainly the use of artificial intelligence associated with portfolio optimization (Paiva, Cardoso, Hanaoka, and Duarte, 2018; Wang, Li, Zhang, and Liu, 2019; Ma, Han, and Wang, 2020; Chen, Zhang, Mehlawat, and Jia, 2020). It is noticeable that these studies chose an approach based on technical analysis, i.e., the assets were selected by forecasting future returns considering the past behavior of asset prices. Thus, there is a gap regarding

research that addresses the use of fundamentalist indicators associated with artificial intelligence, in order to analyze the contribution that data from financial statements can bring in this optimization process.

Thus, the objective of this paper proposes an artificial intelligence technique for stock market forecasting associated with the Markowitz model. Specifically, this research seeks to analyze the performance of an approach consisting of two stages: asset selection by means of Logistic Regression with the use of fundamentalist indicators; and optimization of the selected assets by means of a model based on the Markowitz formulation. In this sense, the results of the proposed approach will be compared to strategies that implement only Logistic Regression and the Markowitz model as unique steps. To this end, the São Paulo Stock Exchange (Ibovespa) Index will be used as a reference.

The organization of this paper is described below. Section 2 reviews some empirical studies that addressed the use of artificial intelligence techniques in asset selection and portfolio optimization. In Section 3, the methodology of this work is described. In addition, the data used, the architecture of the proposed model, the Logistic Regression and the Markowitz model are presented. Section 4 presents and discusses the results obtained. Finally, in Section 5, the conclusion of the work is carried out, where the main findings and their implications are highlighted.

2 - LITERATURE REVIEW

Ma, Han, and Wang (2020) performed a comparison between different artificial intelligence models that perform return forecasting and portfolio optimization. The models reviewed were randon forest, support vector regression (SVR), LSTM neural network, deep multilayer perceptron, and convolutional neural network. The historical base of China Securites 100 index (CSI 100 Index) from 2007 to 2015 was used. For each asset, the return forecast for the next day was made based on the returns of the previous 60 days. The results obtained pointed out the best performance for the random forest model associated with the Markowitz model.

Wang, Li, Zhang and Liu (2020) analyzed asset selection by applying the LSTM neural network model to a large volume of data from the UK Stock Exchange 100 index, between 1994 and 2019. In this case, the input variables chosen were 5 technical indicators and the return of the last 15 days. The experiments showed that the LSTM

neural network model showed higher predictive ability for financial time series than the support vector machine (SVM), random forest, deep neural network and autoregressive integrated moving average (ARIMA) models. Thus, the combination of the LSTM neural network model with the Average-Variance model showed the best results in terms of return and risk.

Paiva, Cardoso, Hanaoka, and Duarte (2018) implemented the SVM model to classify the assets of the São Paulo Stock Exchange Index (Ibovespa), associating the Markowitz formulation to optimize the portfolio. The database of the experiment was composed of a history of 3716 trading days, during 2001 and 2016. During this period, 135 assets were listed on Ibovespa. Twenty-two attributes were used as input data, including the following: opening and closing returns, maximum and minimum prices, as well as momentum, volatility, and volume indicators. The proposed model was then compared with 2 other models. The first is formed by SVM for classifying N assets, which integrates an equal distribution of investment among them (1/N) approach. The second model is based on a random choice of assets (Random) and optimization of the investment distribution using the Mean-Variance (MV) model. Simulations were performed with variations in the models' parameters aiming to find the configuration with the highest accumulated return. According to the empirical results of this study, the model proposed by the authors has a better efficiency when the defined return targets are high.

Heo and Yang (2016) evaluated the prediction of stock market volatility based on financial statements using artificial intelligence. Through the financial results released quarterly by companies, the SVM method was implemented to verify whether this data influenced asset prices. The database for this study was composed of financial information and prices of 200 companies listed on the KOSPI 200 between the first quarter of 2010 and the third quarter 2013. The experimental results showed that the proposed model had higher predictive ability than the predictions made by experts.

Chen, Zhang, Mehlawat and Jia (2020) also presented a model combining an artificial intelligence method to perform stock market forecasting, associated with the MV model to optimize the portfolio composed of the selected assets. For the first stage the authors chose the eXtreme Gradient Boosting (XGBoost) method, which was associated with the Improve Firefly algorithm (IFA) so that the hyper parameters of XGBoost were optimized. The data was collected from the Shanghai Stock Exchange (SSE) 50 index, and the analysis period was from November 2009 to November 2019. In

this regard, this database consisted of 19 indicators, including the last 15 returns and 4 technical indicators. The results pointed to the superiority of the MV-integrated IFAXGBoost over an traditional model. Another conclusion of the study was the confirmation of the importance of the Average-Variance model in the optimization process, since the results were better than the analyzed alternative of assigning equal weights to the chosen assets.

Kaczmarek and Perez (2021) used random forest in the asset selection step. This study was based on companies listed in the S&P 500 during the period from December 31, 1999 to December 31, 2019. In the optimization step, the expected return parameter in the mean-variance model was replaced by the random forest forecast vector. In addition, the maximum investment constraint of 10% in a single asset was added, to avoid a large concentration in only a few assets. The main conclusion of this research was the confirmation that the average-variance model fed by machine learning-based predictions is more efficient than applying these models separately.

Like the literature, this paper proposes the usage of a machine learning technique. Differently of the literature, this paper integrates this technique in a nonlinear optimization model based mathematical optimization. Thus, the contribution of this paper is to propose a simple technique that combines advantages of the Logistic Regression with mathematical optimization.

3 - METHODOLOGY

3.1 - Proposed Model

The model proposed for asset selection consists of predicting returns by means of logistic regression, where the input data for training purposes of the logistic regression are fundamentalist indicators. Initially, the financial statements of the assets listed on a stock exchange are collected. Then, with this historical data, the performance indicators of these assets are calculated. With this, the latter indicators are transformed into percentage variations over time. Then, logistic regression is used to capture the influence of these indicators on asset prices. The resulting regression equation is then used in the test data set for the purpose of selecting potential assets for purchase. Then, the proportion of buying these potential assets is determined in order to minimize the investment risk of the portfolio by optimizing a mathematical model based on the Markowitz formulation. All these steps can be implemented using the Phyton programming language, taking

advantage of several available libraries. Such a model is schematized in Figure 1 and detailed below.

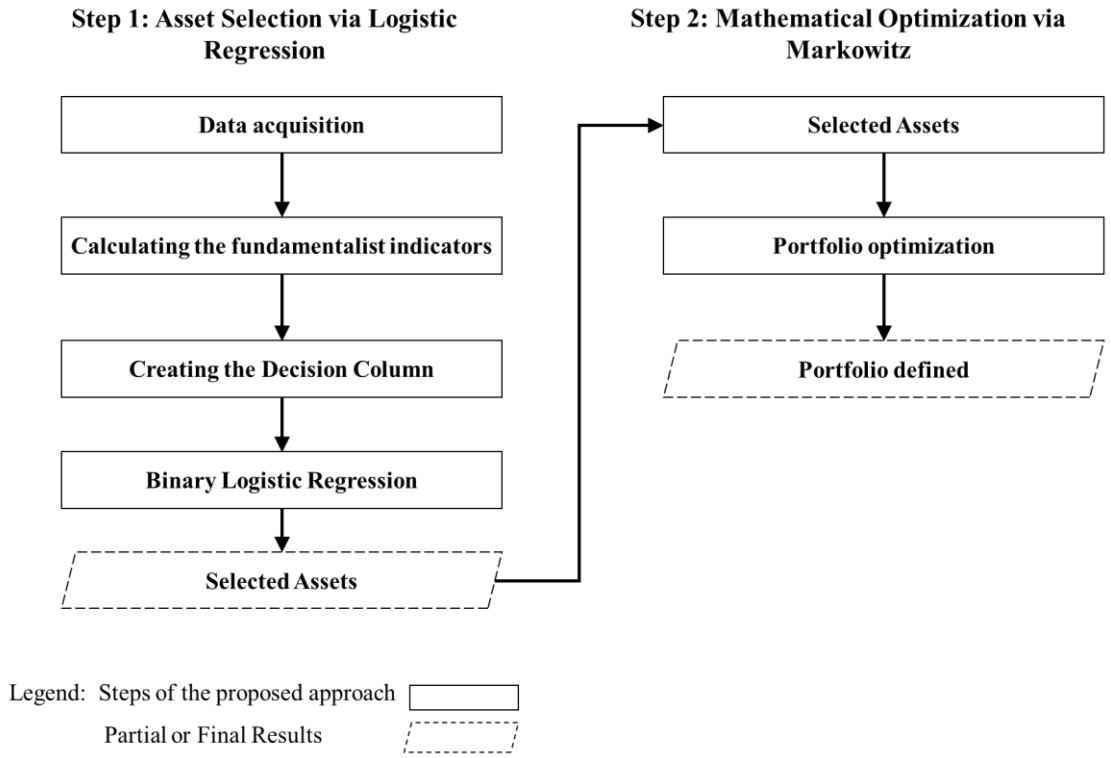

Figure 1: Scheme with the steps of the proposed model.

3.1.1 - Data acquisition

This study used the public data of companies listed on Ibovespa during the period December 2012 through June 2020, obtained from the Fundamentus (2021). These data are the financial values present in the balance sheets and income statements. In total, data was collected for 77 assets. The quotations of these assets and the Ibovespa index in the same period were also used, according to the Yahoo Finance (2021).

3.1.2 - Calculating the fundamentalist indicators

For this study 6 fundamentalist indicators were chosen: Gross Margin, Net Margin, Return on Equity (ROE), EBITDA Margin, Return on Assets (ROA) and Net Debt over EBITDA. The calculation of these indicators is shown in Table 1.

Table 1: Indicators and their formulae

Indicator	Formulae
Gross Margin	Gross Profit / Net Revenue
Net Margin	Net Profit / Net Revenue

Return on Equity (ROE)	Net Profit / Shareholders' Equity
EBTIDA Margin	EBITDA / Net Revenue
Return on Assets (ROA)	Net Profit / Total Assets
Debt over *EBITDA*	Net Debt / *EBTIDA*

These indicators were then transformed into percentage changes from the previous quarter to avoid distortions in the data due to the different sizes of companies listed on the stock exchange.

3.1.3 - Creating the Decision Rule and Column

From the history of fundamentalist indicators, their percentage variation (Δ%) was calculated in relation to the previous quarter. Meanwhile, the variation of asset prices and the stock market index were calculated in relation to the following quarter. In this sense, the machine learning will identify, from the historical data, the influence that the variation of the indicators has on the future return of the assets. It will also identify if this return exceeds the stock market index by a certain threshold (φ), this being the criterion for asset selection. Figure 2 presents the construction and operation of the decision rules.

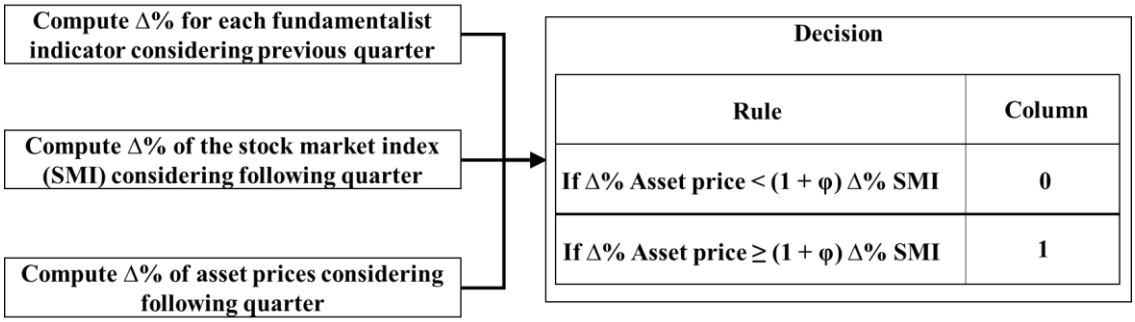

Figure 2: Schema for Decision Rules.

3.1.4 - Binary logistic regression

Logistic regression is one of the fundamental techniques when dealing with a classification problem. For Buckinx and Poel (2005), the main advantages of logistic regression are that its assumptions are satisfied by many families of distributions and that it provides fast and robust results.

The main objective of this technique consists in studying the probability of an event Y occurring (value 1) or not occurring (value 0) (FÁVERO E BELFIORE, 2019). For this, a set J = {1,2,...,k} of explanatory variables is used, as shown in equation (1).

$$Z_i = \alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki} \quad (1)$$

In this equation, Z_i is called the logit function, X_{ji} represents the explanatory variable $j \in J$ of observation i, α is a constant, β_j are the parameter estimates for each variable X_{ji} and the index i indicates each observation of the sample (i =1, 2, ..., n).

Since the response of the binary logistic regression must be limited to 0 or 1, a linear equation might not provide a good fit to the model data. To solve this, the logit function is defined as (FÁVERO AND BELFIORE, 2019):

$$Z_i = \ln\left(\frac{p_i}{1-p_i}\right) \quad (2)$$

Equation (2) associates the probability of occurrence of event Y of observation i with its linear predictors, by means of a nonlinear function. The ratio of the probability of success p_i to the probability of failure 1-p_i is called the odds. Isolating p_i as a function of Z_i in equation (2), equation (3) can be written as:

$$p_i = \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki}}}{1+e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki}}} \quad (3)$$

Then the probability of non-occurrence of the event is given by equation (4) as:

$$1 - p_i = \frac{1}{1+e^{Z_i}} = \frac{1}{1+e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki}}} \quad (4)$$

In equations (3) and (4), p_i is a function with image between 0 and 1, in which Z_i is the input value, which in turn is a linear function with domain =]-∞,+ ∞[.

3.1.4.1 - Parameter Estimation

The general method used to find the parameters $\beta_j : j \in J$ in logistic regression is called maximum likelihood. This method consists in finding the parameters that maximize the probability of finding the observed values in the data set (HOSMER, LEMESHOW and STURDIVANT, 2013). Therefore, this method provides the values of the parameters so that the model fits the pattern of the observed data.

Each decision variable Y_i follows the Bernoulli distribution and can assume the values 0 or 1. Thus, the probability function, $f(y_i)$, is given by:

$$f(y_i) = p_i^{y_i} \times (1-p_i)^{1-y_i} \quad (5)$$

Since the observations are independent, the likelihood function, $l(x)$, is defined as the product of the terms in equation (5) for the n observations (HOSMER, LEMESHOW, and STURDIVANT, 2013). This results in the equation (6):

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{Z_i}}{1+e^{Z_i}}^{y_i} \times \frac{1}{1+e^{Z_i}}^{1-y_i} \qquad (6)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots \beta_k)$.

For practicality, the natural logarithm is applied to equation (6) to obtain (7):

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \left[(y_i) \times \ln\left(\frac{e^{Z_i}}{1+e^{Z_i}}\right) + (1-y_i) \times \ln\left(\frac{1}{1+e^{Z_i}}\right) \right] \qquad (7)$$

This is the case of a nonlinear programming problem with multiple variables, without constraints. To find the optimal solution of $L(\boldsymbol{\beta})$, it is necessary that its first-order partial derivatives in relation to the parameters β_j are equal to 0 (FÁVERO and BELFIORE, 2013, p. 434). Performing these operations, the resulting equations are the following expressions (HOSMER, LEMESHOW and STURDIVANT, 2013, p. 37):

$$\sum_{i=1}^{n} \left(y_i - \frac{e^{Z_i}}{1+e^{Z_i}} \right) = 0 \qquad (8)$$

$$\sum_{i=1}^{n} x_{ij} \left(y_i - \frac{e^{Z_i}}{1+e^{Z_i}} \right) = 0 \quad \forall\, j = 1,2, \dots, k \qquad (9)$$

Finally, replacing Z_i in equations (8) and (9) with their linear predictors, one can obtain equations (10) and (11):

$$\sum_{i=1}^{n} \left(y_i - \frac{e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki}}}{1+e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki}}} \right) = 0 \qquad (10)$$

$$\sum_{i=1}^{n} x_{ij} \left(y_i - \frac{e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki}}}{1+e^{\alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \cdots + \beta_k \times X_{ki_i}}} \right) = 0 \quad \forall\, j = 1,2, \dots, k \qquad (11)$$

Since these are nonlinear equations, it is necessary to use some numerical method to find the vector $\boldsymbol{\beta}$, which forms a critical point. To define if this critical point is a maximum, minimum or cell point, it is necessary to calculate the determinant of the hessian matrix and its principal minors (FÁVERO and BELFIORE, 2013, p. 431).

However, Silva (2012) cites that Amemiya (1985) has already shown that the log of the likelihood function is globally concave, so one has a single resulting critical point that will be a maximum. Silva (2012) points out that one of the optimization algorithms that can be used in this case is the Newton-Raphson method, which will converge to the single maximum due to the global concavity of the likelihood function. Finally, we recall that in the proposed model y_i represents a decision to buy or not to buy asset *i* and X_{ij} refers to indicator *j* of asset *i*.

3.1.5 - Portfolio optimization

The mean-variance model proposed by Markowitz (1952) can be represented by the following multi-objective model:

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_J \sigma_{ij} \ , \max \sum_{i=1}^{n} x_i \mu_i \qquad (12)$$

Subject to:

$$\sum_{i=1}^{n} x_i = 1 \qquad (13)$$

$$0 \leq x_i \leq 1, \forall\ i = 1, 2, \ldots, n \qquad (14)$$

where:

σ_{ij} = covariance between assets i and j, (i = 1, 2,..., n) e j = (1, 2,..., n)

μ_i = expected return on asset i, (i = 1, ..., n)

x_i = percentage of asset i in the portfolio, (i = 1, ..., n)

Following de Souza Dutra et al. (2020) and Chen et al. (2020), this formulation can be transformed into a problem with only one objective by introducing the risk aversion parameter, $\lambda \in [0, 1]$, proposed by Chang et al (2009):

$$\min\ \lambda \left[\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_J \sigma_{ij} \right] - (1-\lambda) \left[\sum_{i=1}^{n} \sum_{j=1}^{n} x_i \mu_i \right] \qquad (15)$$

Subject to:

$$\sum_{i=1}^{n} x_i = 1 \qquad (16)$$

$$0 \leq x_i \leq 1, \forall\, i = 1, \ \ldots, n \qquad (17)$$

The solution of this model is composed of a set of points, which form the so-called efficient frontier. However, the Mean-Variance model does not indicate which of these optimal solutions should be chosen, leaving it up to the investor to make this choice according to his level of risk aversion and return expectation (PAIVA, CARDOSO, HANAOKA, and DUARTE, 2018).

Through this model, Markowitz showed the importance of diversification of investments to reduce the total risk of a portfolio (REILLY and BROWN, 2012). In this sense, the need to analyze the relationships between the different assets in a portfolio was highlighted, avoiding only an individual look at each component of the portfolio.

In the model proposed in this paper, the parameter λ in equation (15) is set equal to 1. That is, only the risk minimization objective is considered, and the optimization of returns is performed in a previous step through the Logistic Regression.

3.2 Approaches for comparison.

To evaluate the performance of the proposed model, comparisons of the results with three other approaches described below were performed.

3.2.1 Approach 1: LR + 1/N

This first approach consists only of running the Logistic Regression (LR), the first step in Figure 1, followed by assigning equal proportions of the investment value to the selected assets. In this sense, the proportion of each asset will be equal to 1/N, where N is the number of selected assets.

3.2.2 Approach 2: Markowitz Model

In this other approach, only the optimization model defined by equations (12)-(14) was used, in which the solution was defined as being the one with the lowest risk. In this case, this model was implemented on a portfolio composed of all assets in the database, without taking into account a previous selection.

3.2.3 Approach 3: Ibovespa Index

According to the B3 stock exchange (2021), the Ibovespa index is the main indicator of performance of the stocks traded in Brazil and consists of a theoretical portfolio of assets that correspond to 80% of the capital market transactions carried out in the country.

4 - RESULTS AND DISCUSSIONS

For the proposed model, the exogeneous parameter φ is set to 0.01. Applying the proposed model results in the composition of the optimized portfolios for the analyzed period shown in Table 2. The selection of these assets was determined by the Logistic Regression while the investment percentage in each asset was defined in the mathematical optimization step.

Table 2 - Results of the proposed model.

30/06/2020		30/09/2020		31/12/2020		31/03/2021		30/06/2021	
B3SA3	0%	ABEV3	14%	ABEV3	8%	B3SA3	0%	ABEV3	10%
BEEF3	6%	AMER3	0%	AMER3	0%	BEEF3	5%	B3SA3	0%
BRKM5	0%	COGN3	0%	BEEF3	6%	BRKM5	0%	BEEF3	7%
COGN3	0%	CPFE3	9%	BRFS3	3%	BRML3	0%	BRKM5	1%
CPLE6	0%	CSNA3	0%	BRML3	0%	CCRO3	0%	CMIG4	0%
CSNA3	0%	CYRE3	0%	CCRO3	0%	CMIG4	0%	COGN3	0%
CYRE3	0%	ECOR3	0%	CIEL3	1%	CPFE3	6%	CPFE3	7%
EGIE3	10%	EQTL3	12%	COGN3	0%	CPLE6	0%	CPLE6	0%
ELET6	0%	EZTC3	0%	CSAN3	0%	CSAN3	0%	CSAN3	1%
ENBR3	5%	FLRY3	11%	CSNA3	0%	CSNA3	0%	CYRE3	0%
ENEV3	3%	GGBR4	0%	CYRE3	0%	CYRE3	0%	ENGI11	9%
EQTL3	9%	GOAU4	0%	EGIE3	8%	ELET6	0%	EQTL3	8%
EZTC3	0%	ITSA4	4%	ELET6	0%	EMBR3	4%	EZTC3	0%
HGTX3	2%	MULT3	3%	EMBR3	4%	ENBR3	3%	GGBR4	0%
HYPE3	6%	RADL3	13%	ENBR3	3%	ENGI11	8%	GOAU4	0%
IGTA3	2%	RENT3	0%	ENEV3	2%	EZTC3	0%	GOLL4	0%
ITSA4	3%	SULA11	8%	EQTL3	7%	FLRY3	7%	HYPE3	6%
JBSS3	0%	TAEE11	18%	EZTC3	0%	GGBR4	0%	IGTA3	1%
JHSF3	0%	UGPA3	2%	FLRY3	7%	GOAU4	0%	ITSA4	2%
LREN3	0%	VALE3	5%	GGBR4	0%	HYPE3	4%	JBSS3	1%
MRFG3	0%	VIIA3	1%	GOAU4	0%	ITSA4	2%	JHSF3	0%
MRVE3	0%			GOLL4	0%	JBSS3	0%	LCAM3	3%
PRIO3	0%			HYPE3	5%	JHSF3	0%	MRFG3	0%
QUAL3	2%			IGTA3	0%	LCAM3	2%	MRVE3	0%
RADL3	10%			ITSA4	0%	MGLU3	0%	MULT3	0%
SULA11	4%			JBSS3	0%	MULT3	0%	PETR4	0%
SUZB3	18%			JHSF3	0%	PRIO3	0%	RADL3	9%
TIMS3	6%			LAME4	0%	QUAL3	1%	RENT3	0%
TOTS3	5%			LCAM3	2%	RADL3	8%	SBSP3	0%
VALE3	2%			MGLU3	0%	RENT3	0%	TAEE11	14%
VIIA3	0%			MRFG3	0%	SBSP3	0%	TIMS3	5%
WEGE3	6%			MRVE3	0%	SULA11	3%	USIM5	0%
				MULT3	0%	SUZB3	15%	VALE3	2%

PRIO3	0%	TAEE11	12%	VIVT3	9%
QUAL3	0%	TIMS3	4%	WEGE3	7%
RADL3	8%	TOTS3	4%	YDUQ3	0%
RENT3	0%	UGPA3	0%		
SBSP3	0%	USIM5	0%		
TAEE11	13%	VALE3	1%		
TIMS3	3%	VIIA3	0%		
TOTS3	3%	VIVT3	7%		
UGPA3	0%	WEGE3	5%		
USIM5	0%				
VALE3	1%				
VIIA3	0%				
VIVT3	8%				
WEGE3	5%				

Thus, based on the composition of these portfolios in each quarter, it was verified what the return would have been if this model had been followed to support the investment strategy. It is important to point out that the portfolio is updated every quarter according to the balance sheet data of the previous quarter. Figure 3 shows the returns accumulated during the analyzed period in the different approaches.

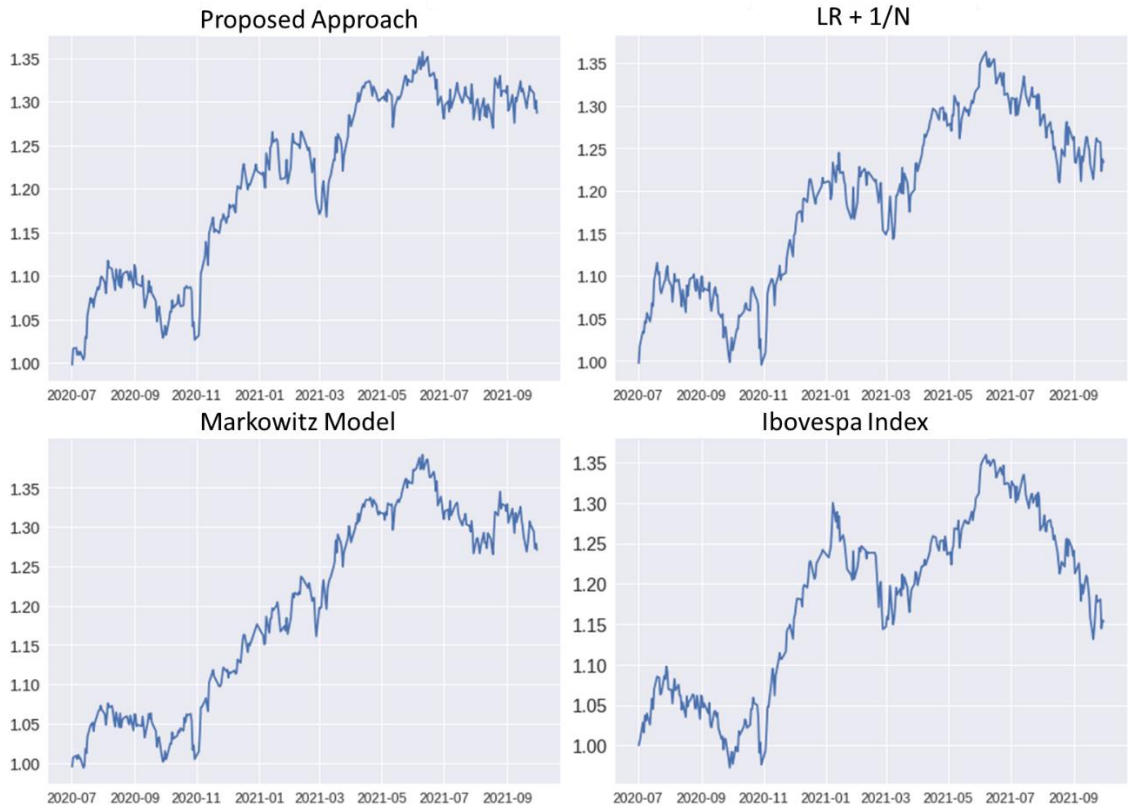

Figure 3: Cumulative returns.

Then, the approaches were compared via the Sharpe and Maximum Drawdown indexes, which are detailed in Appendix 1. The Sharpe ratio measures the portfolio's performance in relation to its risk, compared to a risk-free asset. The higher this indicator, the greater the efficiency of the portfolio. Similarly, the Maximum Drawdown index is also an indicator that measures risk, in which it indicates what would be the greatest possible loss when investing in a given portfolio. In this sense, the lower the Drawdown, the lower the risk of investment losses.

Table 3 summarizes the performance measures related to the aforementioned indexes, return and volatility. For this, an average rate of 4% was used as the return for a risk-free investment.

Table 3 - Performance Comparison.

	Proposed Approach	**LR + 1/N**	**Markowitz Model**	**Ibovespa Index**
Return	28,7%	23,3%	27,1%	15,4%
Volatility	20,4%	23,5%	19,2%	23,5%
Sharpe Index	1,05	0,95	0,98	0,59
Maximum Drawdown	- 8,1%	- 11,2%	- 9,1%	-16,8%

As can be seen in Table 3, the scenarios in which the logistic regression was used to select the assets that should compose the investment portfolio obtained significantly higher returns than the Ibovespa Index in the same period. The proposed model reached an accumulated return 13.3% higher than the Ibovespa Index, while its volatility was 3.1% lower. Thus, it is verified that machine learning associated with fundamentalist indicators has the capacity to help decision making in the investment allocation process. In this context, between the two models in which the logistic regression was used, the one that had the additional step of portfolio optimization reached the highest return, 28.7%, compared to the scenario in which equal weights were attributed to the selected assets, 23.3%. This demonstrates the efficiency of the portfolio when there is diversification based on the covariances between the assets. Still, the proposed model also obtained the highest Sharpe ratio 1.05, and the lowest Drawdown, -8.1%. This indicates greater portfolio efficiency, in which a greater return was achieved with lower risks.

The traditional Markowitz model also achieved a relatively high return, 27.1%, and also achieved a Sharpe Index very close to 1, representing an efficient portfolio. What

explains this positive result of the Average-Variance model, despite the use of average returns as parameters of the expected returns, is the normal distribution of the historical data, as can be seen in Chart 1.

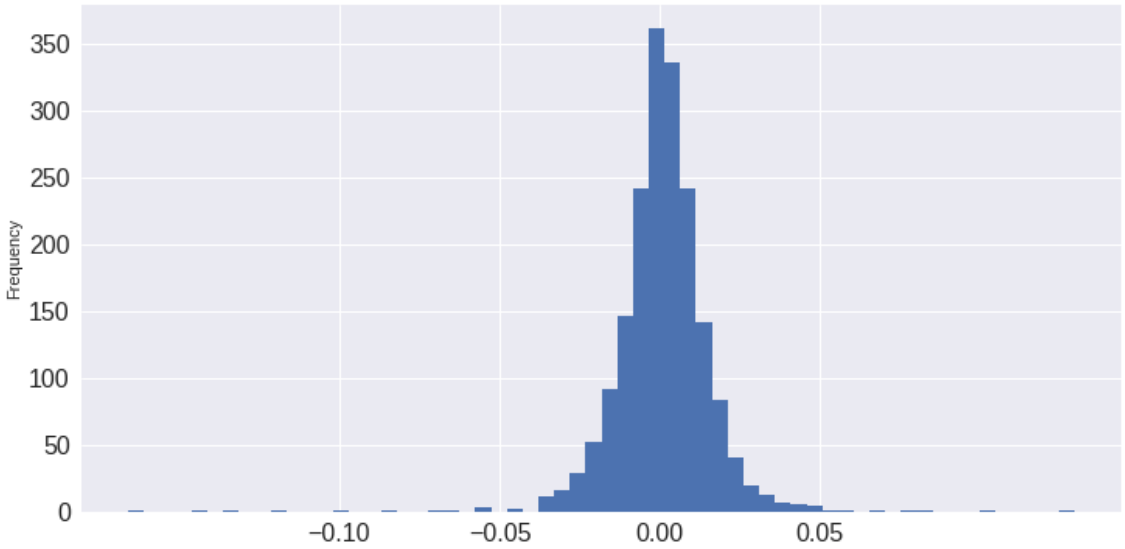

Chart 1: Histogram of the daily returns of the database assets.

Table 4 shows the descriptive measures of the daily returns of the models analyzed. Figure 4 graphically presents the dispersion of these returns.

Table 4 - Descriptive Measures of Daily Returns.

	Proposed Approach	**LR + 1/N**	**Markowitz Model**	**Ibovespa Index**
Average Daily Return	0,088%	0,077%	0,084%	0,055%
Standard Deviation	1,2%	1,3%	1,1%	1,3%
Minimum Return	- 3,7%	- 4,6%	- 3,5%	- 5,1%
Maximum Return	4,3%	4,1%	3,4%	3,1%

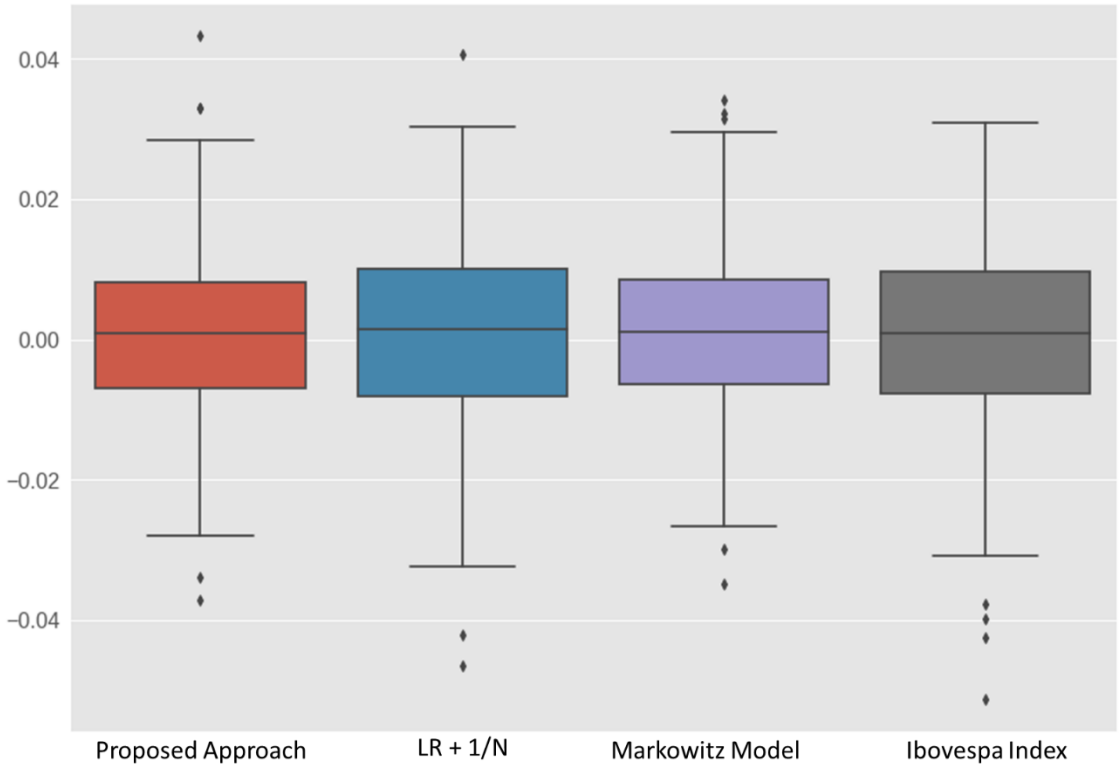

Figure 4: Daily returns.

From these results one can see the best performance of the Logistic Regression associated with risk optimization, in which it obtained the highest average daily return, 0.088%, reaching daily returns above 4%. On the other hand, the Ibovespa Index in the same period obtained an average daily return equal to 0.055%, in which it was observed a higher number of negative daily returns. The proposed model also obtained an average daily return 0.004% higher than the approach in which only Markowitz's model was used. These small incremental gains in daily returns represent a large gain on investments, as the cumulative return grows exponentially over time. These results reinforce the value that artificial intelligence methods can add to the investment allocation decision process.

5 – CONCLUSIONS

This study aimed to analyze the efficiency of asset selection through logistic regression and fundamentalist indicators, as well as the balancing of these portfolios by minimizing the covariance between the assets. In this sense, the experimental results pointed out the superior performance of this model in relation to the Ibovespa index, in which a 13.3% higher accumulated return was verified during the 5 quarters. At the same time, portfolio volatility was 3.1% lower over the same period. These results indicate that the theory of efficient markets is valid in the medium and long term, and that the price of

assets tends to converge to the equity value of companies. Thus, the analysis of the percentage variations of the fundamentalist indicators makes it possible to predict the assets with the highest earning potentials. In this context, machine learning is a tool that allows capturing this relationship between balance sheet data and the future performance of assets.

The results also showed the gain that is obtained when defining the proportion of an asset in the portfolio taking into account the covariance matrix. For the same set of assets selected by the logistic regression, the one with this type of risk optimization achieved a cumulative return 5.4% higher than the same portfolio with equal weights. In addition, the volatility of the portfolio with the optimized weights was 3.1% lower. Thus, one can see that the correlation between the assets is a relevant factor in the return and risk optimization process of an investment portfolio.

There are several opportunities that can still be explored in this theme. Future work could, for example, verify the performance of other artificial intelligence models, as well as it is also possible to add external input data to organizations, such as macroeconomic indicators, such as the exchange rate and inflation. It would also be possible to analyze the model's behavior by adding new restrictions, such as limitations of the maximum and minimum percentage that can be invested in a single asset.

REFERENCES

B3. **Ibovespa Index**. Available at https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm. Accessed in: October/2021.

BROWN, R. **Analysis of Investment & Management of Portfolios**. 10. ed. Cengage Learning: 2012.

BUCKINX, Wouter; POEL, Dirk Van den. **Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting**. European Journal of Operational Research. V. 164(1), p. 252-268, July. 2005. https://doi.org/10.1016/j.ejor.2003.12.010

CHEN, W.; ZHANG, H.; MEHLAWAT, M. K.; JIA, L. **Mean–variance portfolio optimization using machine learning-based stock price prediction**. Applied Soft Computing, v. 100, p. 106843, March. 2021. https://doi.org/10.1016/j.asoc.2020.106943

DUTRA, M. D. **Avaliação financeira de um projeto de casa inteligente para uma residência no Ceará**. Exacta. v. 20(1), p.176{197, Jun. 2021. https://doi:10.5585/exactaep.2021.17173.

DE SOUZA DUTRA, M. D.; JUNIOR, G. C.; DE PAULA FERREIRA, W.; CAMPOS CHAVES, M. R. A **customized transition towards smart homes: A fast framework for economic analyses**. Applied Energy. vol. 262, p.114559, Mar. 2020. https://doi.org/10.1016/j.apenergy.2020.114549.

DE SOUZA DUTRA, M.D.; ANJOS, M.F.; LE DIGABEL, S. **A general framework for customized transition to smart homes**. Energy. vol. 189, p.116138, Dec. 2019. https://doi:10.1016/j.energy.2019.116138.

FÁVERO, L. P. BELFIORE, P. **Data Science for Business and Decision Making**. Elsiever, 2019.

FÁVERO, L. P.; BELFIORE, P. **Pesquisa operacional para cursos de engenharia**. Rio de Janeiro: Elsevier, 2013.

FREITAS, F. D.; SOUZA, A. F.; ALMEIDA, A. R. **Prediction-based portfolio optimization model using neural networks**. Neurocomputing. v. 72 (10-12), p. 2155-2170, June. 2009. https://doi.org/10.1016/j.neucom.2008.08.019

FUNDAMENTUS. Available at https://fundamentus.com.br/index.php. Accessed in: August/2021.

HEO, J.; YANG, J. Y. **Stock Price Prediction Based on Financial Statements Using SVM.** International Journal of Hybrid Information Technology. v. 9(2), p. 57-66, 2016. http://dx.doi.org/10.14257/ijhit.2016.9.2.05

HOMER, D. W.; LEMESHOW, S.; STURDIVANT, R. **Applied logistic regression**. 3rd ed. New Jersey: John Wiley & Sons, 2013.

KACZMAREC, T. PEREZ, K. **Building portfolios based on machine learning predictions.** Economic Research- Ekonomska Istraživanja, v. 1(0), p. 1-19, Fev. 2021. https://doi.org/10.1080/1331677X.2021.1875865

MA, Y.; HAN, R.; WANG, W. **Portfolio optimization with return prediction using deep learning and machine learning**. Expert Systems with Applications, v. 165(1), p. 113973, March. 2021. https://doi.org/10.1016/j.eswa.2020.113973

MARKOWITZ, H.M. **Portfolio selection**. The Journal of Finance 7 (1), p. 77–91, 1952. http://dx.doi.org/10.1111/j.1540-6261.1952.tb01525.x

PAIVA, F. D.; CARDOSO, R. T. N.; HANAOKA, G. P.; DUARTE, W. M. **Decision-making for financial trading: A fusion approach of machine learning and portfolio selection**. Expert Systems with Applications. v. 115, p. 635-655, Jan. 2019. https://doi.org/10.1016/j.eswa.2018.08.003

SILVA, A. M. **Técnicas de Data Mining na aquisição de clientes para financiamento de Crédito Direto ao Consumidor – CDC**. Master's Dissertation. Escola Superior de Agricultura "Luiz de Queiroz". Piracicaba, 2012.

WANG, W.; LI, W.; ZHANG, N.; LIU, K. **Portfolio formation with preselection using deep learning from long-term financial data**. Expert Systems with Applications, v. 143 (1), p.113042, April. 2020. https://doi.org/10.1016/j.eswa.2019.113042

YAHOO FINANCE. Available at https://finance.yahoo.com. Accessed in August/2021.

APPENDIX 1

Sharpe Index Formula

$$S = \frac{R_i - R}{\sigma_i}$$

Where:

S – Sharpe Index

R_i – Return on investment i

R – Return on a risk-free investment

σ_i – Volatility of investment i

Maximum Drawdown Formula (MDD)

$$MDD = \frac{Maximum\ Value}{Minimum\ Value} - 1$$