

## **Modelling of S&P 500 Index Price Based on U.S. Economic Indicators: Machine Learning Approach**

**Ligita Gaspareniene<sup>1</sup>, Rita Remeikiene<sup>2</sup>, Aleksejus Sosidko<sup>3</sup>, Vigita Vebraite<sup>4</sup>**

<sup>1,2,4</sup>*Vilnius University, Law Faculty*

*Sauletekio av. 9, Vilnius, Lithuania*

*E-mail. <sup>1</sup>ligita.gaspareniene@tf.vu.lt; <sup>2</sup>rita.remeikiene@tf.vu.lt (Corresponding author); <sup>4</sup>vigita.vebraite@tf.vu.lt*

<sup>3</sup>*Mykolas Romeris University, Faculty of Public Governance and Business*

*Ateities str. 20, Vilnius, Lithuania*

*E-mail. <sup>3</sup>alsosidko@mruni.eu*

**crossref** <http://dx.doi.org/10.5755/j01.ee.32.4.27985>

*In order to forecast stock prices based on economic indicators, many studies have been conducted using well-known statistical methods. Meanwhile, since ~2010 as the power of computers improved, new methods of machine learning began to be used. It would be interesting to know how those algorithms using a variety of mathematical and statistical methods, are able to predict the stock market. The purpose of this article is to model the monthly price of the S&P 500 index based on U.S. economic indicators using statistical, machine learning, deep learning approaches and finally compare metrics of those models. After the selection of indicators according to the data visualization, multicollinearity tests, statistical significance tests, 3 out of 27 indicators remained. The main finding of the research is that the authors improved the baseline statistical linear regression model by 19 percent using a ML Random Forest algorithm. In this way, model achieved accuracy 97.68 % of prediction S&P 500 index.*

**Keywords:** *S&P 500 Index; Economic Indicators; Machine Learning; Deep Learning; Fundamental Analysis; Stock.*

### **Introduction**

As it was noted by Prazak (2018), “the basic notion of the stock market is the market value of the stocks” (p. 1613). Shares are widely recognized as a very risky investment since fluctuations in their price are determined by a number of factors both quantitative (macro and microeconomic) and qualitative (social, political, psychological, etc.). According to the theory of arbitrage, particular fundamental macroeconomic and financial indicators represent the components of investment risk, which means that particular indicators sufficiently fully reflect the risk stock prices are exposed to.

The main purpose of stock market indices is to serve as benchmarks with reference to which investors can assess their current or potential investment performance. Composed of 500 U.S. largest companies operating in key industries, the S&P 500 is considered to be one of the most influential indicators reflecting and even determining the movements in the U.S. stock market. The stocks included in the index account for about 80 percent of the total U.S. stock market capitalization, and since a large proportion of U.S. corporations operate globally, fluctuations in the index value can as well cause the changes in equity markets worldwide, especially the ones most tied to the U.S. economy. Therefore, a deeper notion of the S&P 500 value determinants can help more accurately forecast prospective stock market returns (losses).

The S&P 500 stock index modelling is not a new issue. Previous models are based on various methodologies (statistical learning – Liu *et al.*, 2016; classification – Tang, 2020; genetic programming – Sheta *et al.*, 2013;

GARCH – Awartani, Corradi, 2005; SV mixture – Durham, 2007; time series analysis and simulation – Chan, 2009, etc.). *The novelty of this research* is that we use machine learning algorithms to help predict stock prices. With these algorithms, more accurate results are usually obtained than with basic statistical models.

The main purpose of the article is to model the monthly price of the S&P 500 index based on U.S. economic indicators using statistical, machine learning, deep learning approaches and finally compare metrics of those models. The defined purpose was detailed into the following objectives: 1) to categorize most influential indicators affecting the stock market performance; 2) to select and substantiate methods of the research; 3) to model the S&P 500 index price based on U.S. economic indicators using statistical, machine learning, deep learning methods. Research methods include comparative and systematic literature analysis, statistical, machine learning and deep learning algorithms.

### **Categorization of the Most Influential Indicators Affecting the Stock Market Performance**

Although in particular cases stock price fluctuations seem to have been caused by some divorced factors, macroeconomic environment certainly has a significant impact on the general movement of the stock market, which is confirmed by the results of previous studies. Nevertheless, as it was noted by Pilinkus (2010), the direction of the causality running between macroeconomic environment and the stock market has not still been comprehensively researched, just as the number of the

most influential factors that could be employed to explain stock market volatility. Literature analysis helped to categorise the factors that, with reference to previous

studies, can be considered as most influential in terms of the stock market performance (see Table 1).

Table 1

**Categorisation of the Most Influential Indicators Affecting Stock Markets**

Category	Indicators	Author(s), year	Category	Indicators	Author(s), year
<b>General macroeconomic indicators</b>	Consumer price index	Pilinkus, 2010; Ligocka <i>et al.</i> , 2016; Prazak, 2018; Akhtar <i>et al.</i> , 2020	<b>Monetary supply indicators</b>	M1	Kumari, 2012
	Industrial production index	Shanken, Weinstein, 2006; Liu, Shrestha, 2008; Prazak, 2018		M2	Baharumshah, 2004; Baharumshah, <i>et al.</i> , 2009; Raymond, 2009
<b>Labour market indicators</b>	Unemployment rate	Boyd <i>et al.</i> , 2005; Pilinkus, 2010; Prazak, 2018		Year treasury 10	Kvietkauskienė, Plakys, 2017
	Unemployed persons	Boyd <i>et al.</i> , 2005		Funds rate	Patelis, 1997; Bernanke, Kuttner, 2005
	Employed persons	Kvietkauskienė, Plakys, 2017		Currency in circulation	Brzenk, 2018
	All employees	Wertheim, Robinson, 2004		<b>Consumer (household) financial behaviour indicators</b>	Personal savings rate
	Labour force participation rate	Abudy <i>et al.</i> , 2019	Personal consumption expenditures		Poterba <i>et al.</i> , 1995; LeCompte, 2015; Sosa Cueto, 2017
	Population	Quayes, Jamal, 2016	Real disposable personal income		Santos, Veronesi, 2006; Tarver, 2020
<b>Real estate market indicators</b>	Housing starts	Conrad, Loch, 2015; Brzenk, 2018	Personal saving		LeCompte, 2012
	New private housing building permits	Cresap, 2017; Cortes, Weidenmier, 2019	Demand deposits	Lin, 2017	
<b>Credit market indicators</b>	Interest rate	Adam, Tweneboah, 2008; Alam, Uddin, 2009; Pilinkus, 2010; Addo, Sunzuoye, 2013	<b>Commodity market indicators</b>	WTI crude oil price	Wang, Xie, 2012; Shaeri, Katircioglu, 2018
	Treasure bill	Adam, Tweneboah, 2008; Addo, Sunzuoye, 2013		Motor vehicle retail sales	Khouri, 2015; Aliu <i>et al.</i> , 2017
	Commercial and industrial loans	Lummer, McConnell, 1989			
	Bank prime loan rate	Ippolito <i>et al.</i> , 2017			
	Bank credit	Ippolito <i>et al.</i> , 2017			

Source: compiled by the authors.

As it can be seen in Table 1, the indicators that are recognised to have the most significant impact on stock market return in general and on S&P 500 in particular can be attributed to the following categories: general macroeconomic indicators, labour market indicators (unemployment rate and jobs reports), real estate indicators, credit market indicators, monetary supply indicators, consumer (household) financial behaviour indicators and commodity market indicators.

Evaluating the impact of the main general macroeconomic indicators, most researchers invoke OECD composite leading indicator (CLI), inflation rate (commonly represented by CPI) and industrial production index. The OECD composite leading indicator provides early signals of economic turning points, and thus allows to timely evaluate the current and short-term economic situation (OECD, 2012). Significance of the impact of CLI on the effectiveness of investors' decisions is empirically confirmed in some studies (e.g., Topcu & Unlu, 2013). Inflation rate measures fluctuations in the prices of goods/services. Higher prices diminish consumer purchase power and can force Federal Reserve to raise interest rates,

which, in its turn, causes an economy to cool and squelches stock rallies, while falling inflation rates result in the opposite effects. Industrial production index measures the overall industrial output that serves as an indicator of the health of national production factories. The co-integrating interrelationship between the industrial production index and stock prices was identified by Shanken and Weinstein (2006), Liu and Shrestha (2008), and others. According to Prazak's (2018) findings, general macroeconomic indicators provide a statistically significant impact on stock prices in the long run, though strength of the impact may substantially vary among economic sectors.

Labour market indicators affect stocks since they reflect the general state (i.e. strength or weakness) of the economy, contribute to the national GDP and income structure, and ensure smooth functioning of the target economic and social instruments (Otoi, 2015). In essence, a higher number of the employed is linked to increasing economic output, which, in its turn, determines higher volumes of retail sales and leads to generation of corporate profits, and vice versa.

On the other hand, as it was found by Boyd et al. (2005), growing unemployment rate can cause stock prices to rise during the periods of economic development, but cause them to drop during the periods of economic decline. Although vast majority of literature sources (Boyd *et al.*, 2005; Pilinkus, 2010; Prazak, 2018; Akhtar *et al.*, 2020, etc.) addressing the relationship between macroeconomic indicators and stock market return tend to focus on generalising unemployment or employment rates (unemployment rate is considered to affect three types of corporate information: information on future interest rates, equity risk premiums and corporate earnings and dividends (Boyd *et al.*, 2005)), for more accurate results it is no less important to consider particular labour market components such as the number of persons employed and unemployed, all employees and labour force participation rate reflecting the (un)employment situation in households which, through mediation of goods, service and capital markets, actually forms business-household economic relations (Otoi, 2015) and determines the real level of household income subsequently used for consumption or saving.

The real estate market, and in particular housing market indicators serve as reflectors of consumer confidence: feeling save and secure, consumers are more inclined to apply for loans and start building new housing, which positively affects both the future performance of the construction industry and a national economy in general. The robust evidence of the causality running from the housing market to stock market was provided by Li et al. (2020), although the relationship between both markets was found to be either positive or negative depending on the wealth effect. Housing starts and building permits are considered to be most market-moving indicators sending stock prices higher or lower. Among other macroeconomic variables, housing starts were found to have the highest predictive ability for stock market volatility by Conrad and Loch (2015), Brzenk (2018) and others, while the predictive ability of building permits was confirmed by Cresap (2017), Cortes and Weidenmier (2019), etc.

The impact of credit market indicators, first of all, manifests through interest rate which reflects the cost of capital. As it was noted by Alam and Uddin (2009), the relationship between interest rate and share prices is inverse because increasing deposit interest rates prompt agents to switch their capital from stock markets to the banking sector, and vice versa. Treasury bill rate is commonly employed as a measure of interest rate, and the relationship between Treasury bill rate and stock prices is often reported negative (Adam & Tweneboah, 2008; Addo & Sunzuoye, 2013, etc.). Commercial and industrial loans (C&I) provide businesses with working or financial capital. Thus, the demand from business companies for these loans serves as a sign of economic growth and therefore positively affects share prices (Lummer & McConnell, 1989). Similar effects are observed with growing bank prime loan rate (which measures the rate at which banks lend to their clients; the rate is correlated with the federal funds rate and tends to move along with it) (Ippolito *et al.*, 2017) and bank credit rate (Ippolito *et al.*, 2017). Hence, bank loan and credit rates play an important role as information conveyors to capital markets.

Money supply that measures the quantity (abundance or scarcity) of money in circulation reflects consumer purchasing power and thus is considered to be a predictor of potential demand. Abundance of money in the economy is commonly linked to growing stock prices as large money supply means more money available to investing, and vice versa. The relationship between various components of money supply and stock market fluctuations was confirmed by numerous studies (M1 and stock market – Kumari, 2012; M2 and stock market – Baharumshah, 2004, Baharumshah et al., 2009, Raymond – 2009, etc.; currency in circulation – Brzenk, 2018, etc.). According to Kvietkauskienė and Plakys (2017), a positive impact on stock market returns is also exerted by the 10-year government benchmark yields, while Patelis (1997) found that a significant negative impact on stock returns (i.e., anticipated excess returns rather than dividend or expected returns) in the short term and positive in the long term is exerted by the federal funds rate. Bernanke and Kuttner's (2005) estimations showed that a surprise 25-basis-point cut in the federal funds rate caused stock prices to rise by nearly one percent.

In the category of consumer (household) financial behaviour indicators, the results of previous studies suggest that the minor changes in personal consumption expenditure indicate a relatively low volatile stock market (LeCompte, 2015). The statistically significant positive relationship between consumption, wealth and S&P500 was also found by Poterba et al. (1995) and Sosa Cueto (2017). An increase in real disposable personal income causes higher stock valuations, and thus determines an increase in the overall value of the stock market (Tarver, 2020). What concerns personal saving and personal savings rate, an inverse relationship between these indicators and the stock market performance is commonly declared (LeCompte, 2012), although some studies (e.g., Owiredo et al.'s (2016) study) suggest that domestic savings do not have any significant impact on stock market fluctuations as their regression coefficients are not statistically significant at the 5 % level. Lin's (2020) study proposes that there exists a negative relationship between the deposit and stock market: “deposit growth tends to be smaller or even negative when the stock market is booming” (p. 3), and vice versa.

Among the commodity market indicators, WTI crude oil prices and motor vehicle retail sales can be categorised as most influential factors on the stock market. Rising crude oil prices are linked to the declining potential economic growth and higher inflation in a short term. Vague economic growth prospects, in their turn, undermine corporate earnings and thus dampen stock prices. The cross-correlation between WTI and stock markets was confirmed by Wang and Xie (2012), Shaeri and Katircioglu (2018), etc. Motor vehicle retail sales indicator represents the automotive industry as a major global economic and industrial power that produces 60 million cars and trucks a year, consumes nearly half the global consumption of petroleum and employs 4 million people (Papatheodorou & Harris, 2007), thus significantly affecting the overall stock market. The U.S. auto market has been the second largest in the world with three multinational corporations – ‘General Motors’, ‘Ford Motor’ and ‘Chrysler’, whose monthly sales

announcements substantially contribute to (un)attractiveness of not only auto, but also auto industry-related (e.g., satellite radio, car battery industries, etc.) stocks to investors.

Summarizing, the literature analysis helped identify 27 indicators that can have the most significant impact on stock markets and S&P 500 index in particular. Further in the research, the S&P 500 price based on U.S. most influential economic indicators will be modelled.

### Research Methodology

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system.

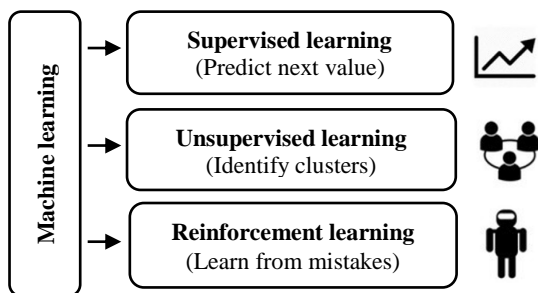


Figure 1. Machine Learning Categories

Source: created by the authors based on Vitola J., Pozo F., Tibaduiza D., & Anaya M. (2017).

**Supervised learning:** The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

**Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (independent variables learning).

**Reinforcement learning:** An algorithm interacts with a dynamic environment in which it has a specific purpose, e.g., drive a vehicle, or play a game against an opponent. Every action performed has an impact on the environment, and the environment returns feedback, according to which the algorithm continues to learn and make decisions.

This article uses the supervised learning approach to analyse the current problem, which in turn is further subdivided into supervised learning classification and supervised learning regression. Below are the most used algorithms for modelling. The category of deep learning, which is a branch of machine learning, is singled out mostly due to its unique methodology.

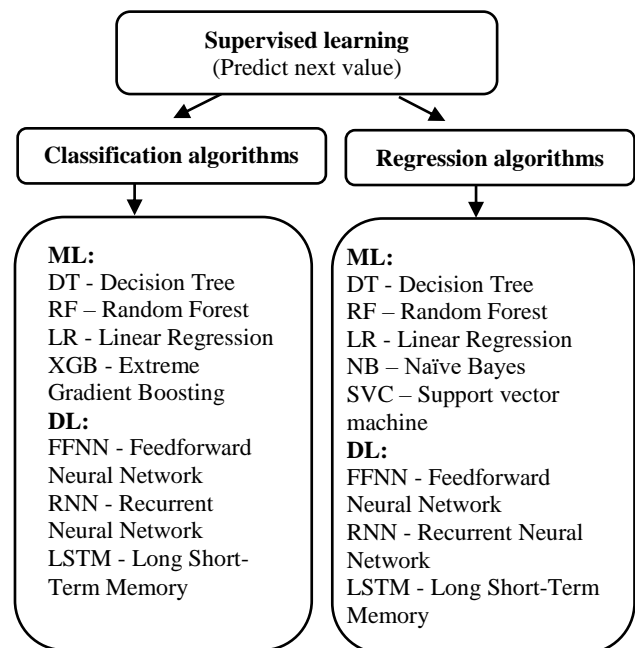


Figure 2. Supervised Learning Categories and Most Used Algorithms

Source: created by the authors

Thus, Figure 2 provides a summary of the most used algorithms. It should be noted that in order to model stock markets, researchers use regression and classification-type algorithms, and some of them are universal.

With the help of regression algorithms, we aim to predict real value, such as stock returns or price. Variables can be discrete or continuous. The aim of classification algorithms is to predict a categorical variable. For example, whether a stock needs to be sold at a specific point in time or bought, or maybe even held. Classification algorithms are not considered in this work.

In this work, we will use the Linear regression algorithm from statistics as baseline model. From machine learning algorithms we are going to use Decision tree, Random forest, Linear regression, and Extreme Gradient Boosting. From deep neural network algorithms, we will use FFNN.

Linear Regression algorithms are used to investigate the dependence of one variable on one or more variables and to predict subsequent mean values of the variables.

$$Y = b_0 + b_1X + b_2Z + b_3W + e \tag{1}$$

From here:

- Y – dependent variable
- b<sub>0</sub>, b<sub>1</sub>, b<sub>2</sub>, b<sub>3</sub> - model coefficients
- X, Z, W – independent variables
- e – error.

This equation is used for both quantitative and qualitative research. The coefficients indicate whether the dependent variable Y increases or decreases as the regressors increase.

A decision tree algorithm is used to determine a course of action or show a statistical probability. It forms the outline of the namesake woody plant, usually upright but sometimes lying on its side. Each branch of the decision tree

represents a possible decision, outcome, or reaction. The farthest branches on the tree represent the end results. The advantages of this method are that it is understood intuitively, does not require special data preparation, such as data normalization. The model allows you to work with large amounts of data without special preparatory procedures.

Meanwhile Random Forest algorithm make up of many simple Decision trees. Random forest consists of many trees grown on randomly selected traits. With such a methodology and the use of many cultivated trees, the average accuracy is improved compared to a single solution tree.

Using a random forest, the initial data set is randomly divided into several smaller sets of the same size. The decision tree is then learned for each split smaller set. After creating multiple decision trees and in order to classify or predict new data, the random forest algorithm receives a “voice” for each decision tree, which should be the class or value of the dependent variable.

XGB is another machine learning method that is based on the logic of the decision tree method with the addition of gradients. This model iterates and improves itself many times over. It starts with initiating an ensemble with a single model whose predictions can be quite naive and even if his predictions are very inaccurate, subsequent additions to the ensemble will help eliminate these errors.

The last algorithm we are going to use is Feedforward Neural Network. Artificial neural networks have been developed based on human biological neural networks. The brain is made up of billions of neurons that connect to form neuronal connections. A neuron describes a cell that can receive, process, and transmit an electrochemical signal. The simplest neural network is called a single-layer perceptron, which is made up of two layers, the input, and the output. More complex networks, meanwhile, have one or more hidden layers with multiple neurons. Such networks are called multilayer perceptron’s.

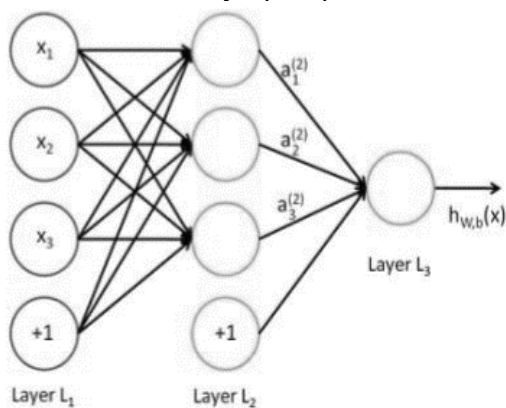


Figure 3. Example of Multilayer Perceptron

Source: <https://www.altoros.com/blog/introduction-to-neural-networks-and-metaframeworks-with-tensorflow/>

The input layer  $L_1$  contains a training data set, each neuron of which contains different variables (independent variables). When moving to hidden layer  $L_2$ , a weighting factor is assigned. Each neuron in the hidden layers has summation and activation functions. The summation

function is identical to the linear regression function. The activation function, meanwhile, plays the role of a director, which by its purpose passes neurons with its weights further to other neurons or transforms their values. The output layer  $L_3$  also applies a function that presents the results in the desired form. The stochastic gradient descent (SGD) strategy, implemented by the backpropagation algorithm, is the most common method for learning multilayer perceptron’s. SGD is a way to minimize the error function, one of the main components of which is the learning rate.

The S&P 500 index close price was modelling in this paper. The study used a python programming language with many libraries in Google Colab environment. The analysis period between 1970-01-01 / 2020-12-01, and both S&P 500 index close price and U.S economic indicators data frequency are month.

Below we are presenting a logical diagram of the research, which consists of 4 stages, which has at least few points inside.

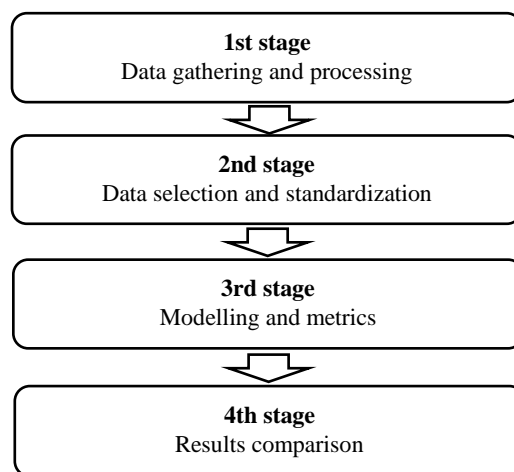


Figure 4. Logical Diagram of the Research

Source: created by the authors

The first step is data collection process from the finance.yahoo.com and fred.stlouisfed.org websites using panda’s data reader library. Our dependent variable is the closing price of the S&P 500 index, and the independent variables are 27 US economic indicators: consumer price index, industrial production index, unemployment rate, unemployed persons, employed persons, all employees, labour force participation rate, population, housing starts, new private housing building permits, interest rate, 3-month treasure bill, commercial and industrial loans, bank prime loan rate, bank credit. The obtained data were combined into one data set, which was checked for missing values.

The second phase of the research began with the analysis of the dependent variable. We performed Augmented Dickey-Fuller, Dickey-Fuller GLS, Phillips-Perron, KPSS, Zivot-Andrews tests to determine if S&P 500 close prices is stationary. Later, we made visualization of the data between the closing price and each independent variable to determine if there is a linear relationship between the variables and thus to select only those variables that it has. After, we checked whether the independent variables correlated with each other using the

VIF test. Finally, because the indicators are measured in different scales, the data were normalized using z score criteria to produce accurate and reliable models.

In the third stage of the study, S&P 500 price modelling is performed, which starts with a baseline statistical linear regression model. Following this model, it was finally determined which independent variables should remain and which should be eliminated. To decide that T - (Student's) test for each regressor were used. It helps to understand if a regressor is statistically significant or not. If the value of the regressor p is <0.05, then we can say that the regressor is statistically significant and we do not delete it. If the value of the regressor p is ≥ 0.05, then the regressor is statistically insignificant and we remove it from this and further models. After baseline model we made other models which we mentioned earlier and to understand how well they performed we calculated following indicators:

1. Coefficient of determination (R<sup>2</sup>). This is the most important indicator of model confidence in the data, which is mandatory in all descriptions of regression models. The interpretation of R<sup>2</sup> is as follows - what percentage of Y's behaviour is explained by the behaviour of the variables X, Z, W. The coefficient of determination values acquires between 0 and 1. The higher value indicates more reliable model.

2. Mean square error (MSE). This value measures the root mean square error between actual and predicted values. The smaller it is, the more accurate is model.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

From here:

- $Y_i$  is the actual value of the dependent variable.
- $\hat{Y}_i$  is the predicted value of the dependent variable.
- $n$  is the number of predicted values.

In the last stage of the study, we compared model's accuracy and error metrics with each other. The obtained results are also compared with the research discussed in theory part.

### Results of Empirical Research

As everywhere in the studies, data is needed, so the S&P 500 index and 27 US indicators were downloaded. The price of the S&P 500 index and its history is shown below (see figure 5). It is clear from the picture that the biggest longer recessions were two: dotcom and the 2008 financial crisis. At COVID19 we also see larger declines, but also rapid recoveries.

After downloading all the data (S&P 500 and 27 U.S. Indicators), they were cleared of missing values and merged into one dataset of the period 1970-01-01 / 2020-11-01. Data frequency – month.

In the second stage of the study, we examine whether our dependent variable is non-stationary or stationary (see table 2). Examples of non-stationary processes are random walk with or without a drift (a slow steady change) and deterministic trends (trends that are constant, positive, or negative, independent of time for the whole life of the series). When forecasting or predicting the future, most time series models assume that each point is independent

of one another. The best indication of this is when the dataset of past instances is stationary. For data to be stationary, the statistical properties of a system do not change over time.

Table 2

Tests for S&P 500 Price Stationarity

Test name	Test value	Test crit. Value (5%)
Augmented Dickey-Fuller	0.210	-3.42
Dickey-Fuller GLS	0.983	-2.88
Phillips-Perron	0.741	-3.42
KPSS	0.484	0.15
Zivot-Andrew	0.667	-5.07

Source: created by the authors

The rule to understand if variable is stationary or not is described follow:

1. If test value > Test crit → accept null hypothesis, i.e., data non-stationary. Unit root exists.
2. If test value < Test crit → reject null hypothesis, i.e., data stationary. Unit root does not exist.
3. Rules above need to understand vice versa on KPSS test.

From the table we can see that the tests results confirms that data **is not** stationary, which means that past prices have an impact on current prices. Non – stationarity we can also confirmed and by Figure 5.

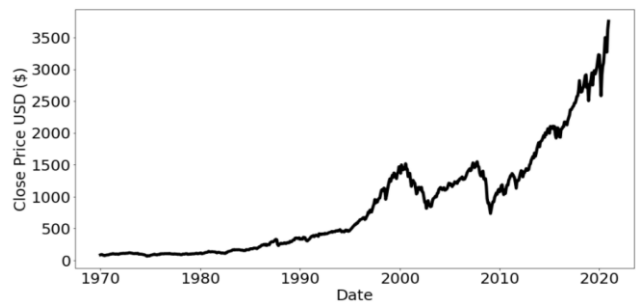
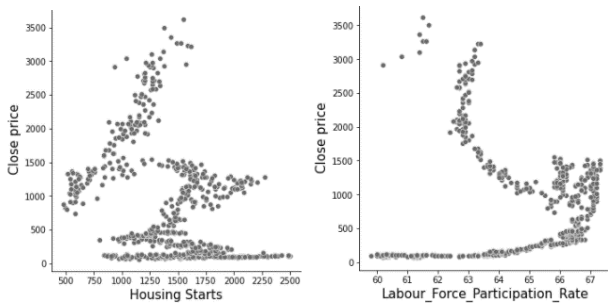


Figure 5. S&P 500 Index Close Price History

Source: created by the authors

To make the data stationary, it was necessary to create a new dependent variable, which was calculated from the current monthly price minus the previous month's price. In this case, the tests mentioned earlier were recalculated and the data were found to be stationary. The problem was that in further analysis of the linear relationship between the new dependent and independent variables, no relationship was found, so the research with the modified dependent variable was not worthwhile. In this case we came back to the usual dependent variable, which is understood as the monthly closing price of the S&P 500 index. So, the limitation of this study and the assumption now appeared that the index is only affected by U.S. economic indicators and the impact of past prices on current prices are not examined.

Next step of second stage we visualized the relationship between close price and economic indicators. We found that 11 of them did not meet the linear regression rule that the data should be linear. See examples below.



**Figure 6.** Examples of Non-Linearity from the Research  
 Source: created by the authors

Thus, after such visualization, 16 indicators were retained for further analysis.

The next step in the second stage was to if independent variables did not correlate with each other. The VIF test was chosen for this purpose. A VIF values between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity. Thus, our aim was to leave those indicators below which VIF values do not exceed 5.

Table 3

**VIF Test Results**

No	Variables	VIF value
1	Employment Persons	41112
2	3-month Treasury bill	48
3	CPI	3575
4	M2	3588
5	M1	1439
6	All employees	48520
7	Industrial production index	2922
8	Personal consumption expenditures	20958
9	Commercial and industrial loans	462
10	Read disposable personal income	24669
11	Currency in circulation	4809
12	WTI	42
13	Population	4356
14	Bank credit	2340
15	Personal savings	170
16	Demand deposits	633

Source: created by the authors

From the table it can be seen that the VIF test values for each variable are well above 5. In this case, the indicator with the highest VIF value is removed and the whole set is recalculated. This is done as long as all variables with values less than 5 remain.

Table 4

**VIF Test Results after Variables Dropping**

No	Variables	VIF value
1	3-month Treasury bill	1.482
2	WTI	2.638
3	Personal savings	2.256

Source: created by the authors

We can see, that from 16 indicators remained just 3: 3-month Treasury bill, WTI and personal savings.

The last thing in our second stage was to take those 3 indicators together with S&P 500 index price data and normalized them by z score criteria.

$$Z = \frac{x - \mu}{\delta} \tag{3}$$

This is due to the fact that the scales for our variables are different, so we would like to normalize them to produce accurate and reliable models. A Z-score itself is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

We started the third stage of the study by constructing a statistical linear regression model, which is our baseline. At the time of compilation, it was also finally verified if the existing remaining three variables are reliable for modelling part. This was verified by the T - (Student's) test, the results of which are given below.

Table 5

**Selected Indicators by p-Value**

No	Variables	P-value
1	3-month Treasury bill	0.000
2	WTI	0.000
3	Personal savings	0.000

Source: created by the authors

If its p-value is greater than 0.05, we must remove the indicator as statistically insignificant. But in our case, all the p-values are zero, so all models with these regressor would be statistically significant.

So, the first statistical linear regression model was created with all remained data - 3 independent variables and one depended variable. The most important model characteristics and metrics mentioned before has been shown below:

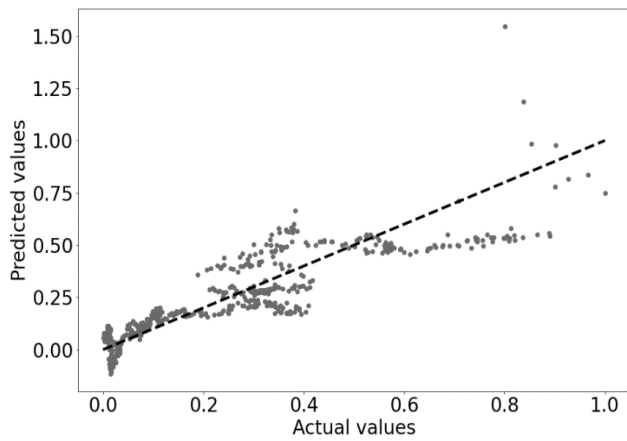
Table 6

**Baseline Regression Model Characteristics**

Name of characteristics	Characteristics
Model name	Baseline regression model
Model type	OLS Linear regression
Datapoints	2444
R <sup>2</sup>	78,50 %
MSE	0.0113

Source: created by the authors

And the model and its actual points are shown in the following 7 figure.



**Figure 7.** Baseline Regression Model

*Source: created by the authors*

It should be noted that the data is normalized, so the scale is small. Our line is our model, and the points are the true values of the dependent variable. As we can see, some of them are quite far from the line, whereas others – not. That is why we have an accuracy of 78.50% and an error of 0.0113.

The model itself can be written by the following equation:

$$S \ \& \ P500indexcloseprice = 0.2769 - 0.4839treasurebill + 0.2044WTI + 1.2499Personalsavings$$

Personal savings has the highest coefficient, which means the highest impact on the model. The 3-month treasury bill has a negative impact on the index price, which means a decrease in the index price.

The second Linear regression model was created using a machine learning approach. The most important model characteristics and metrics mentioned before have been shown below:

Table 7

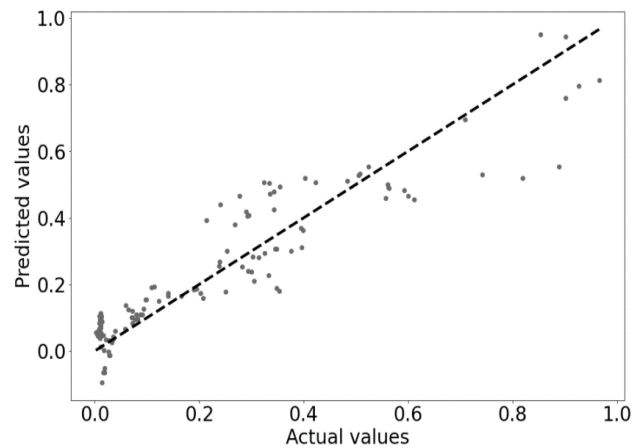
**ML Regression Model Characteristics**

Name of characteristics	Characteristics
Model name	ML Regression model
Model type	OLS Linear regression
Train - test split*	20 %*
Datapoints in train dataset	1952
Datapoints in test dataset	492
R <sup>2</sup>	86.39 %
MSE	0.0083

\*Most cases percent is between 20-30%. Also see annex.

*Source: created by the authors*

And the model and its actual points are shown in the following Figure 8.



**Figure 8.** ML Regression Model

*Source: created by the authors*

Compared to Figure 7, a slightly lower distance between points and model is seen, and the line itself is steeper. The accuracy of the model is 86.39% and is 7.89% higher than our baseline model. Meanwhile, MSE is 0.0083 and is 0.003 points lower.

The model itself can be written by the following equation:

$$S \ \& \ P500indexcloseprice = 0.2856 - 0.4935treasurebill + 0.2159WTI + 1.1616Personalsavings$$

Personal savings has the highest coefficient, which means the highest impact on the model. But compared to the baseline regression model, this coefficient is lower and its effect on the S&P 500 price is smaller.

The third model was created using a deep learning approach. The most important model characteristics and metrics mentioned before have been shown below:

Table 8

**DL FFNN model characteristics**

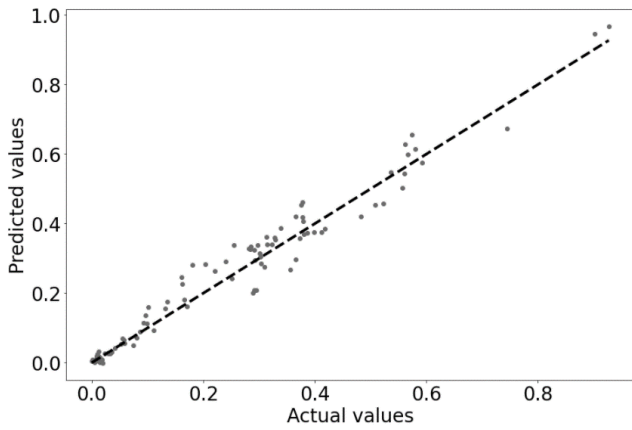
Name of characteristics	Characteristics
Model name	DL FFNN model
Model type	Sequential NN
Train - test split	20 %
Datapoints in train dataset	1952
Datapoints in test dataset	492
Number of hidden layers	3 (512, 256, 128 neurons)
Activation function	Relu*
Optimizer*	Adam*
Batch size*	20
Epoch*	100
R <sup>2</sup>	96.94 %
MSE	0.0013

\* see annex.

*Source: created by the authors*

And the model and its actual points are shown in the following Figure 9.





**Figure 9.** DL FFNN Model

Source: created by the authors

Compared to Figure 8., lower distance between points and model is seen. The accuracy of the model is 96.94 % and is on the 10.55 % higher than our ML Linear Regression model. Meanwhile, MSE is 0.0013 and is on the 0.007 points lower.

The model itself cannot be written by the simple form as Linear regression models, thus we do not put in equation.

The fourth model was created using ML approach. The most important model characteristics and metrics mentioned before has been shown below:

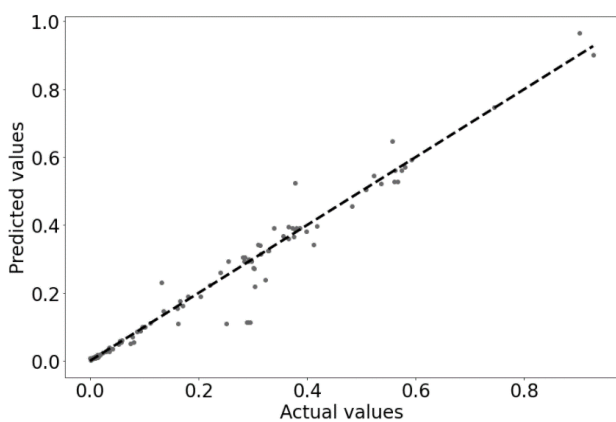
Table 9

**ML Decision Tree Model Characteristics**

Name of characteristics	Characteristics
Model name	ML Decision Tree model
Model type	Decision Tree Regressor
Train - test split	20 %
Datapoints in train dataset	1952
Datapoints in test dataset	492
R <sup>2</sup>	96.24 %
MSE	0.0016

Source: created by the authors

And the model and its actual points are shown in the following 10 figure.



**Figure 10.** ML Decision Tree Model

Source: created by the authors

Compared to Figure 9., no significant difference was observed. The accuracy of the model is 96.24 % and is on the 0.7 % lower than our DL FFNN model. Meanwhile, MSE is 0.0016 and is on the 0.0003 points higher.

The model itself cannot be written by the simple form as Linear regression models, thus we do not put in equation.

The fifth model was created using ML approach. The most important model characteristics and metrics mentioned before has been shown below:

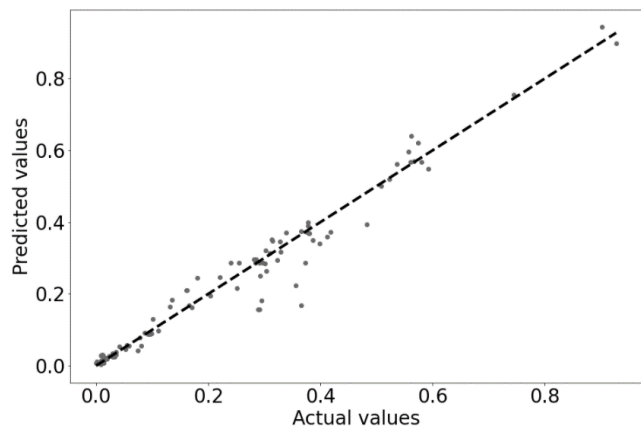
Table 10

**ML XGB Model Characteristics**

Name of characteristics	Characteristics
Model name	ML XGB model
Model type	Gradient Boosting Regressor
Train - test split	20 %
Datapoints in train dataset	1952
Datapoints in test dataset	492
R <sup>2</sup>	96.65 %
MSE	0.0015

Source: created by the authors

And the model and its actual points are shown in the following 11 figure.



**Figure 11.** ML XGB model

Source: created by the authors

Compared to Figures 9,10., no significant difference was observed. The accuracy and MSE of FFNN, Decision Tree and XGB models are similar.

The model itself cannot be written by the simple form as Linear regression models, thus we do not put in equation.

The last model was created using ML approach. The most important model characteristics and metrics mentioned before has been shown below:

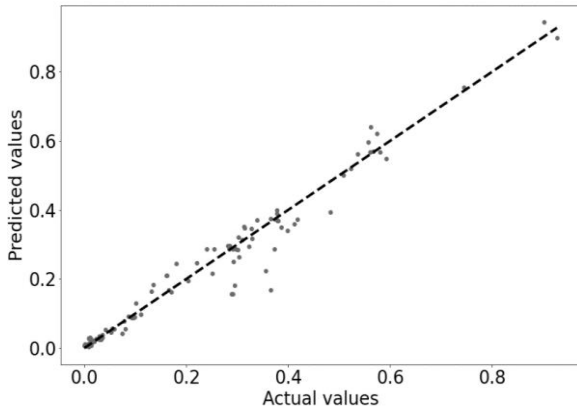
Table 11

**ML Random Forest Model Characteristics**

Name of characteristics	Characteristics
Model name	ML Random Forest model
Model type	Random Forest Regressor
Train - test split	20 %
Datapoints in train dataset	1952
Datapoints in test dataset	492
R <sup>2</sup>	97.68 %
MSE	0.0010

Source: created by the authors

And the model and its actual points are shown in the following 12 figure.



Source: created by the authors

**Figure 12.** ML Random Forest model

Compared to Figures 9, 10, 11, no significant difference was observed. The accuracy and MSE of FFNN, Decision Tree and XGB models are similar.

The model itself cannot be written by the simple form as Linear regression models, thus we do not put in equation.

In order to provide as structured information as possible, a Table 12 was created with the name of each model and its metrics.

Table 12

**Models Summary**

Model name	R <sup>2</sup> (%)	MSE
Baseline LR	78.50	0.0113
ML LR	86.39	0.0083
DL FFNN	96.94	0.0013
ML Decision Tree	96.24	0.0016
ML XGB	96.65	0.0015
ML Random Forest	97.68	0.0010

Source: created by the authors

From the table we can see that the best model for us was developed by the Random Forest method. The statistical linear regression model was improved by 19 % and the error was reduced 11 times. Since some algorithms can be interpreted more easily than others, the Random Forest algorithm cannot be presented as an equation, so we should compare the research results with other authors using the ML Linear regression model.

**Discussion**

After elimination of the indicators with lowest statistical reliability, it was found that S&P 500 index is most significantly affected by 3 remaining ones: 3-month Treasury bill, WTI and Personal savings. Concerning the negative relationship between 3-month Treasury bill and S&P 500 index, it can be stated that the results of this research are in line with the findings provided by many previous studies (Adam and Tweneboah, 2008; Addo and Sunzuoye, 2013, etc.). The rationale for the negative correlation between these two variables is interpretation of Treasury bill rate as a measure of interest rate. In this regard, a high interest rate is considered to be an obstacle

to higher economic growth and vice versa, i.e. high interest rates raise corporate borrowing costs, while lower interest rates mean cheaper crediting, which, in its turn, can accelerate corporate development and thus lead to higher stock prices. Nevertheless, contrary to Addo and Sunzuoye’s (2013) findings, proposing that Treasury bill rate has weak predictive power on stock market return, the results of this research suggest that the dynamics of 3-month Treasury bill is a significant predictor of the movements in the stock market.

The results of this research also disclose a positive correlation between crude oil prices, represented by WTI, and S&P 500 index, which contradicts the findings provided, for instance, by Chen (2010), Filis (2010), Kelikume and Muritala (2019) and others who argue that cross-correlation between crude oil and stock prices is negative. The rationale for the negative correlation between these two variables is the declining potential of economic growth reflected in rising crude oil prices. On the other hand, Wang and Xie (2012) admit that cross-correlation between crude oil market and U.S. stock market can be multifractal. The discrepancies between the results of this research and some previous findings are likely to be determined by the research duration as well as inputs and methodologies that build subtle biases into the outcomes.

Finally, the relationship between Personal savings and S&P 500 index was found to be positive, which contradicts, for instance, LeCompte’s (2012), Lin’s (2020), etc. findings based on the rationale that when accumulating savings, economic agents do not commonly make investment in stock i.e. they do not spend their money. Nevertheless, growing savings may indicate an increasing level of surplus money that is to be spend not only on consumption, but also on investment, including the investment in stock.

**Conclusions**

The literature analysis helped identity 27 indicators that can have the most significant impact on stock markets and S&P 500 index in particular. Statistical reliability of the 27 theoretical indicators was verified by employing data visualization, multicollinearity and statistical significance tests. Based on the results of these tests, 3 out of 27 indicators – 3-month Treasury bill, WTI and Personal savings – were found to have most significant impact on the S&P 500 index price because these indicators passed the regression analysis assumptions and the models developed with them were statistically reliable. Personal savings had the highest estimated coefficient which means its strongest impact on the models (in particular, Baseline LR and ML LR) and thus on the S&P 500 index price.

The analysis was based on six models: 4 - using machine learning approach, 1 – using deep learning approach, and 1 – using statistical approach. When comparing statistical, machine learning and deep learning approaches, it can be stated that the statistical model (Baseline LR) provides the lowest accuracy, deep learning model (FFNN) – slightly higher accuracy, and the accuracy provided by the other machine learning models is highest. Baseline LR and ML LRT models are easy to interpret and

can be compared to each other, but in the context of all the models developed, only  $R^2$  and MSE characteristics can be compared. In general, machine learning approach is superior to other approaches because it allows both training and testing a model, i.e. initially a model is trained, and later tested. Training is based on less data, but the data is still sufficient to ensure high accuracy and less

errors. The ML Random Forest model was found to be the best machine learning model because it provided highest accuracy and lowest error rate. The statistical linear regression model was improved by 19 % and the error rate was reduced 11 times. In this case, the accuracy of the ML Random Forest was 97.86 %, and MSE was equal to 0.001.

## References

- Abudy, M., Mugeran, Y., & Wiener, Z. (2019). Stock Markets and Female Participation in the Labor Force. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3509228>
- Adam, M., & Tweneboah, G. (2008). Macroeconomic Factors and Stock Market Movement: Evidence from Ghana. Social Science Research Network. Available from Internet: [https://www.researchgate.net/publication/23543215\\_Macroeconomic\\_Factors\\_and\\_Stock\\_Market\\_Movement\\_Evidence\\_from\\_Ghana](https://www.researchgate.net/publication/23543215_Macroeconomic_Factors_and_Stock_Market_Movement_Evidence_from_Ghana). <https://doi.org/10.2139/ssrn.1289842>
- Addo, A., & Sunzuoye, F. (2013). The Impact of Treasury Bill Rate and Interest Rate on the Stock Market Returns: Case of Ghana Stock Exchange. *European Journal of Business and Economics*, 8(2), 15–24. <https://doi.org/10.12955/ejbe.v8i2.378>
- Akhtar, H., Jabeen, R., & Shah, S. Z. A. (2020). Inflation Disagreement and Its Impact on Stock Market Volatility, Unemployment Rate and Financial Soundness. [https://doi.org/10.1007/978-981-15-5660-9\\_37](https://doi.org/10.1007/978-981-15-5660-9_37)
- Alam, M., & Uddin, G. S. (2009). Relationship between Interest Rate and Stock Price: Empirical Evidence from Developed and Developing Countries. *International Journal of Business and Management*, 4(3), 43–51. <https://doi.org/10.5539/ijbm.v4n3p43>
- Aliu, F., Pavelkova, D., & Dehning, B. (2017). Portfolio risk-return analysis: The case of the automotive industry in the Czech Republic. *Journal of International Studies*, 10(4), 72–83. <https://doi.org/10.14254/2071-8330.2017/10-4/5>
- Awartani, B. M. A., & Corradi, V. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting*, 21(1), 167–183. <https://doi.org/10.1016/j.ijforecast.2004.08.003>
- Baharumshah, A. Z. (2004). Stock Prices and Long Run Demand for Money Evidence from Malaysia. *International Economics Journal*, 18(3), 389–407. <https://doi.org/10.1080/1016873042000270027>
- Baharumshah, A. Z., Mohd, S., H., & Yol, M. A. (2009). Stock Prices and demand for Money in China: New Evidences. *Journal of International Financial Markets Institutions and Money*, 19, 171–187. <https://doi.org/10.1016/j.intfin.2007.11.001>
- Bernanke, B. S., & Kuttner, K. N. (2005). What explains the stock market's reaction to Federal reserve policy? *The Journal of Finance*, 60(3), 1221–1257. <https://doi.org/10.1111/j.1540-6261.2005.00760.x>
- Boyd, J. H., Hu, J., & Jagannathan, J. (2005). The Stock Market's Reaction to Unemployment News: 'Why Bad News is Usually Good for Stocks'. *The Journal of Finance*, 60(2), 649–672. <https://doi.org/10.1111/j.1540-6261.2005.00742.x>
- Brzenk, P. (2018). The Impact of the Global Economy on the S&P 500. Available from Internet: <https://www.spglobal.com/en/research-insights/articles/the-impact-of-the-global-economy-on-the-sp-500>
- Chan, E. G. (2009). Forecasting the S&P 500 index using time series analysis and simulation methods. Available from Internet: <https://dspace.mit.edu/handle/1721.1/55206>
- Chen, S. S. (2010). Do higher oil prices push the stock market into bear territory? *Energy Economics*, 32(2), 490–495. <https://doi.org/10.1016/j.eneco.2009.08.018>
- Conrad, C., & Loch, K. (2015). Anticipating Long-Term Stock Market Volatility. *Journal of Applied Econometrics*, 30(7), 1090–1114. <https://doi.org/10.2139/ssrn.2154882>
- Cortes, G. S., & Weidenmier, M. D. (2019). Stock Volatility during the Great Depression. *The Review of Financial Studies*, 32(9), 3544–3570. <https://doi.org/10.1093/rfs/hhy134>
- Cresap, W. (2017). The Real Estate and Stock Market during the Great Depression: Construction Permit Growth as a Leading Economic Indicator for Stock Returns. Available from Internet: [https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=2728&context=cmc\\_theses](https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=2728&context=cmc_theses)
- Durham, G. B. (2007). SV mixture models with application to S&P 500 index returns. Available from Internet: <https://core.ac.uk/download/pdf/32426012.pdf>; <https://doi.org/10.1016/j.jfineco.2006.06.005>
- Filis, G. (2010). Macro economy, stock market and oil prices: do meaningful relationships exist among their cyclical fluctuations? *Energy Economics*, 32(4), 877–886. <https://doi.org/10.1016/j.eneco.2010.03.010>

- Ippolito, F., Ozdagli, A. K., & Perez-Orive, A. (2017). The Transmission of Monetary Policy through Bank Lending: The Floating Rate Channel. *Finance and Economics Discussion Series 2017–026*. Washington: Board of Governors of the Federal Reserve System. <https://doi.org/10.17016/FEDS.2017.026>. <https://doi.org/10.17016/FEDS.2017.026>
- Kelikume, I., & Muritala, O. (2019). The impact of changes in oil price on stock market: evidence from Africa. *International Journal of Management, Economics and Social Sciences*, 8(3), 169–194. Available from Internet: <https://www.econstor.eu/bitstream/10419/205142/1/full-3.pdf>; <https://doi.org/10.32327/IJMESS/8.3.2019.11>
- Khouri, R. M. (2015). Do macroeconomic factors matter for stock returns? Evidence from the European automotive industry. *International Journal of Monetary Economics and Finance*, 8(1), 71–84. <https://doi.org/10.1504/IJMEF.2015.069170>
- Kumari, J. (2012). Relationship between stock prices, exchange rate and the demand for money in India. *Economics, Management, and Financial Markets*, 7(3), 31–52.
- Kvietkauskienė, A., & Plakys, M. (2017). Impact indicators for stock markets return. *Poslovna Izvrnost*, 11(2), 59–84. <https://doi.org/10.22598/pi-be/2017.11.2.59>
- LeCompte, S. (2012). Personal savings rate and the stock market. Available from Internet: <https://www.cxoadvisory.com/economic-indicators/personal-savings-rate-and-the-stock-market/>
- LeCompte, S. (2015). Personal consumption expenditures and the stock market. Available from Internet: <https://www.cxoadvisory.com/economic-indicators/personal-consumption-expenditures-and-the-stock-market/>
- Li, F. L., Wang, M. C., & Kung, H. H. (2020). Housing and Stock Market Nexus in the US. *European Research Studies Journal*, XXIII(3), 114–130. <https://doi.org/10.35808/ersj/1628>
- Ligočka, M., Prazak, T., & Stavarek, D. (2016). The Effect of Macroeconomic Factors on Stock Prices of Swiss Real Estate Companies. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 64(6), 2015–2024. <https://doi.org/10.11118/actaun201664062015>
- Lin, L. (2020). Bank Deposits and the Stock Market. *Review of Financial Studies*, 33(6), 2622–2658. <https://doi.org/10.1093/rfs/hhz078>
- Liu, M. H., & Shrestha, K. (2008). Analysis of the Long-Term Relationship between Macroeconomic Variables and the Chinese Stock Market Using Heteroskedasticity Cointegration. *Managerial Finance*, 34, 744–755. <https://doi.org/10.1108/03074350810900479>
- Liu, C., Wang, J., Xiao, D., & Liang, Q. (2016). Forecasting S&P 500 stock index using statistical learning models. *Open Journal of Statistics*, 06(06), 1067–1075. <https://doi.org/10.4236/ojs.2016.66086>
- Lummer, S. L., & McConnell, J. J. (1989). Further evidence on the bank lending process and the capital market response to bank loan agreements. *Journal of Financial Economics*, 25, 99–122. [https://doi.org/10.1016/0304-405X\(89\)90098-6](https://doi.org/10.1016/0304-405X(89)90098-6)
- OECD. (2012). OECD system of composite leading indicators. Available from Internet: <http://www.oecd.org/sdd/leading-indicators/41629509.pdf>
- Otoiu, A. (2015). Key labour market indicators. Main drivers and impact on other socio-economic phenomena: PhD thesis summary. Bucharest Academy of Economic Studies. Available from Internet: [https://www.researchgate.net/publication/278965977\\_KEY\\_LABOUR\\_MARKET\\_INDICATORS\\_MAIN\\_DRIVERS\\_AND\\_IMPACT\\_ON\\_OTHER\\_SOCIO-ECONOMIC\\_PHENOMENA\\_PhD\\_Thesis\\_Summary](https://www.researchgate.net/publication/278965977_KEY_LABOUR_MARKET_INDICATORS_MAIN_DRIVERS_AND_IMPACT_ON_OTHER_SOCIO-ECONOMIC_PHENOMENA_PhD_Thesis_Summary)
- Owiredu, A., Opong, M., & Asomaning, S. (2016). Macroeconomic Determinants of Stock Market Development in Ghana. *International Finance and Banking*, 3(2), 33. <https://doi.org/10.5296/ifb.v3i2.9555>
- Papatheodorou, Y., & Harris, M. (2007). The automotive industry: economic impact and location issues. Available from Internet: <https://www.industryweek.com/the-economy/article/21958422/the-automotive-industry-economic-impact-and-location-issues>
- Patelis, A. D. (1997). Stock Return Predictability and the Role of Monetary Policy. *Journal of Finance*, 52(5), 1951–1972. <https://doi.org/10.1111/j.1540-6261.1997.tb02747.x>
- Pilinkus, D. (2010). Macroeconomic Indicators and Their Impact on Stock Market Performance in the Short and Long Run: The Case of the Baltic States. *Technological and Economic Development of Economy*, 16(2), 291–304. <https://doi.org/10.3846/tede.2010.19>
- Poterba, J. M., Samwick, A. A., Shleifer, A., & Shiller, R. J. (1995). Stock Ownership Patterns, Stock Market Fluctuations, and Consumption. *Brookings Papers on Economic Activity* 1, 2, 295–372. <https://doi.org/10.2307/2534614>
- Prazak, T. (2018). The Effect of Economic Factors on Performance of the Stock Market in the Czech Republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 66(6), 1613–1626. <https://doi.org/10.11118/actaun201866061613>

- Quayes, S., & Jamal, A. M. M. (2016). Impact of demographic change on stock prices. *The Quarterly Review of Economics and Finance*, Elsevier, 60(C), 172–179. <https://doi.org/10.1016/j.qref.2015.08.005>
- Raymond, K. (2009). Is There a Long Run Relationship Between Stock Prices and Monetary Variables? Evidence from Jamaica. Available from Internet: [http://www.boj.org.jm/uploads/pdf/papers\\_pamphlets/papers\\_pamphlets\\_Is\\_There\\_a\\_Long\\_Run\\_Relationship\\_Between\\_Stock\\_Prices\\_and\\_Monetary\\_Variables\\_Evidence\\_from\\_Jamaica.pdf](http://www.boj.org.jm/uploads/pdf/papers_pamphlets/papers_pamphlets_Is_There_a_Long_Run_Relationship_Between_Stock_Prices_and_Monetary_Variables_Evidence_from_Jamaica.pdf)
- Santos, T., & Veronesi, P. (2006). Labor Income and Predictable Stock Returns. *The Review of Financial Studies*, 19(1), 1–4. <https://doi.org/10.1093/rfs/hhj006>
- Shaeri, K., & Katircioglu, S. (2018). The nexus between oil prices and stock prices of oil, technology and transportation companies under multiple regime shifts. *Economic Research*, 31(1), p. 681–702. <https://doi.org/10.1080/1331677X.2018.1426472>
- Shanken, J., & Weinstein, M. I. (2006). Economic forces and the stock market revisited. *Journal of Empirical Finance*, 13(2), 129–144. <https://doi.org/10.1016/j.jempfin.2005.09.001>
- Sheta, A., Faris, H., & Al-Kasassbeh, M. (2013). A genetic programming model for S&P 500 stock market prediction. *International Journal of Control and Automation*, 6(6), 303–314. <https://doi.org/10.14257/ijca.2013.6.6.29>
- Sosa Cueto, R. (2017). Stock market wealth impact on consumption expenditure. Available from Internet: [https://digitalcommons.buffalostate.edu/cgi/viewcontent.cgi?article=1024&context=economics\\_theses](https://digitalcommons.buffalostate.edu/cgi/viewcontent.cgi?article=1024&context=economics_theses)
- Tang, E. (2020) Forecasting S&P 500 stock index using classification models. Available from Internet: <https://towardsdatascience.com/forecasting-s-p-500-stock-index-using-classification-models-eb41510a896d>
- Tarver, E. (2020). The impact of disposable income on the stock market. Available from Internet: <https://www.investopedia.com/ask/answers/042315/what-impact-does-disposable-income-have-stock-market.asp>
- Topcu, M., Unlu, U. (2013). Do investors consider composite leading indicators? Time series evidence from emerging countries. Available from Internet: <http://store.ectap.ro/articole/901.pdf>
- Vitola J., Pozo F., Tibaduiza D., & Anaya M. (2017). Distributed Piezoelectric Sensor System for Damage Identification in Structures Subjected to Temperature Changes Sensors, 17, 1252. <https://doi.org/10.3390/s17061252>
- Wang, G. J., & Xie, C. (2012). Cross-correlations between WTI crude oil market and U.S. stock market: a perspective from econophysics. *Acta Physica Polonica Series B*, 43(10), 2021–2036. <https://doi.org/10.5506/APhys PolB.43.2021>
- Wertheim, P., & Robinson, M. A. (2004). Market reactions to company layoffs: evidence on the financial distress versus potential benefit hypothesis and the effect of predisclosure information. *Journal of Applied Business Research*, 20(10), 51–62. <https://doi.org/10.19030/jabr.v20i1.2195>
- FFNN example. Available from Internet: <https://www.altoros.com/blog/introduction-to-neural-networks-and-metaframeworks-with-tensorflow/>
- S&P 500 index data source. Available from Internet: [finance.yahoo.com](https://finance.yahoo.com)
- Economic indicators data source. Available from Internet: [fred.stlouisfed.org](https://fred.stlouisfed.org)

## **Annex 1. Keywords of machine and deep learning**

The train-test split is a technique for evaluating the performance of a machine learning algorithm. Train-test split function is using for splitting data arrays into two subsets: for training data and for testing data. With this function, you do not need to divide the dataset manually. By default, Sklearn library train-test split will make random partitions (with random state parameter) for the two subsets. Most used split is 20–30 %, where test dataset is 20–30 % and train dataset 70–80 %.

Relu - The rectified linear activation function or Relu for short is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance.

Optimizer - Optimizers are used to update weights and biases i.e., the internal parameters of a model to reduce the error.

Adam Optimizer - Adam is a replacement optimization algorithm for stochastic gradient descent for training deep learning models. Adam combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm that can handle sparse gradients on noisy problems.

Batch size - Batch size is a term used in machine learning and refers to the number of training examples utilized in one iteration.

Epoch - An epoch is a term used in machine learning and indicates the number of passes of the entire training dataset the machine learning algorithm has completed.

### Authors' biographies

**Ligita Gasparėnienė** is the Head of the Research and Innovation Department, Vilnius University; the leader of the national project "Welfare society." Research interests: shadow economy, digital shadow economy, corruption, money laundering, investments in gold, monetary policy. ORCID ID: <https://orcid.org/0000-002-5535-6552>

**Rita Remeikienė** is Dr. in Social Sciences, a project leader at the Research and Innovation Department, Vilnius University, the senior researcher of the national project "Welfare society." Research interests: shadow economy, digital shadow economy, corruption, money laundering, self-employment. ORCID ID: <https://orcid.org/0000-0002-3369-485X>

**Aleksejus Sosidko** is a PhD student of the Faculty of Business and Economics, Mykolas Romeris University. Research interests: Fundamental analysis, stock market analysis, machine learning. ORCID ID: <https://orcid.org/0000-0002-0438-668X>

**Vigita Vėbraitė** is Dr., an associate professor at Vilnius University Faculty of Law (Department of Private Law). Research interests: private law, civil procedure, electronic justice, alternative dispute resolution. ORCID ID: <https://orcid.org/0000-0003-4351-061X>

The article has been reviewed.

Received in November 2020; accepted in October 2021.



This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (<http://creativecommons.org/licenses/by/4.0/>).