# iscte

**INSTITUTO** UNIVERSITÁRIO **DE LISBOA** 

**Predicting stock price direction using machine learning models** 

Afonso Amado Santos Pinto

Master in Data Science

Supervisor: PhD José Joaquim Dias Curto, Associate Professor with Habilitation, Iscte-Iul

Co-Supervisor: PhD Diana Elisabeta Aldea Mendes, Associate Professor, Iscte-Iul

October, 2022



**BUSINESS SCHOOL** 



**TECNOLOGIAS E ARQUITETURA** 

**Predicting stock price direction using machine learning models** 

Afonso Amado Santos Pinto

Master in Data Science

Supervisor: PhD José Joaquim Dias Curto, Associate Professor with Habilitation, Iscte-Iul

Co-Supervisor: PhD Diana Elisabeta Aldea Mendes, Associate Professor, Iscte-Iul

October, 2022

# **Acknowledgements**

I would like to thank my supervisor Professor José Dias Curto and my co-supervisor Professor Diana Aldea Mendes, for their constant support and availability, as well as for all the knowledge they have passed on to me. During my master's degree, I had the opportunity to serve as a monitor for Professor José Dias Curto and this is a memory that I will cherish for the rest of my life.

Also, I would like to thank my parents, Duarte e Paula, and my siblings, André e Rita, who have always believed in me and served as a source of inspiration.

Finally, I would like to thank Patricia for her unconditional love.

#### **Resumo**

O presente trabalho visa avaliar a eficiência dos modelos de aprendizagem automática na previsão da direção do preço de ações no dia seguinte, utilizando indicadores técnicos e padrões de vela. Este estudo considerou 440 ações do índice Standard & Poor's 500 e, para cada uma delas, foram desenvolvidos três modelos binários diferentes com base nos seguintes algoritmos de aprendizagem automática: Deep Neural Network, Support Vector Machine, e Random Forest. Além disso, foi criado um naive predictor para ajudar a comparar os resultados dos modelos. Os modelos foram julgados com base na sua precisão e retorno financeiro.

Os resultados mostraram que os modelos de aprendizagem automática alcançaram resultados semelhantes aos do modelo naive e não conseguiram prever com precisão a direção do preço das ações no dia seguinte, utilizando as características selecionadas, indicando que não existe uma relação aparente entre os mesmos.

**Palavras-chave**: Inteligência Artificial, Aprendizagem Automática, Mercado de Ações, Previsão de Preço.

**Classificação JEL**:

G17 - Financial Forecasting and Simulation

C88 - Other Computer Software

## **Abstract**

The present work aims to evaluate the efficiency of machine learning models in predicting next-day stock price direction, using technical indicators and candlestick patterns. This study considered 440 stocks from the Standard & Poor's 500 index and for each one of them, three different binary models were developed based on the following machine learning algorithms: Deep Neural Network, Support Vector Machine, and Random Forest. Additionally, a naive predictor was created to help compare the models results. The models were judged based on their accuracy and financial returns.

The results showed that the machine learning models achieved similar results to the naive model and failed to accurately predict the next-day stock price direction using the selected features, indicating that there is no apparent relationship between them.

**Keywords**: Artificial Intelligence, Machine Learning, Stock Market, Price Prediction.

#### **JEL Classification**:

G17 - Financial Forecasting and Simulation

C88 - Other Computer Software

# **Index**



# **List of Tables**



# **List of Figures**



# **List of Abbreviations**

AI – Artificial Intelligence ANN – Artificial Neural Network DNN – Deep Neural Network DT – Decision Tree LSTM – Long Short-term Memory ML – Machine Learning NN – Neural Network OHLCV – Open-High-Low-Close-Volume RF – Random Forest S&P 500 – Standard & Poor's 500 SVM – Support Vector Machine

# **Chapter 1: Introduction**

The growth of processing power and data availability over recent years has enabled Machine Learning techniques to be successfully applied in a wide range of industries. Stock market price prediction has been an area of significant interest to the research community since an accurate forecast can create an investment opportunity or avoid potential losses.

Predicting stock prices has proven to be a challenging task as market data is frequently associated with hidden relationships and a high degree of uncertainty. Scholars, however, believe that Artificial Intelligence algorithms have made it possible to model some of the nonlinearity in the financial markets. While this topic has gained popularity and more studies have been published, many questions remain about those who have been successful. This is because most only study a small number of stocks and rarely show how the results would translate into actual returns in a financial market.

To understand if ML algorithms can successfully forecast the next day stock price direction using technical indicators and candlestick patterns, we selected all the Standard & Poor's 500 constituents that have a historical performance dating back to 2010. Three binary prediction models were developed using these algorithms: Deep Neural Network, Support Vector Machine with a polynomial and a radial basis kernel, Random Forest. Additionally, a naive predictor was created to help compare the models results. The models output predicted whether the stock price would increase or decrease the next day. Based on those predictions, financial returns were calculated and compared between the models and with a buy-and-hold strategy.

The present study is organized into different chapters. Chapter 2 provides context to the findings by reviewing articles with similar themes. Chapter 3 presents the theoretical background that provided the framework for this study's methods. A detailed description of the methodologies used to gather the necessary data and to train the different prediction models is provided in Chapter 4. An analysis of the results obtained by each model is presented and compared in Chapter 5. Lastly, in Chapter 6, general conclusions are derived from the models results.

## **Chapter 2: Literature Review**

#### **2.1. Stock Market**

One of the most significant aspects of the current global economy is the stock market, as it can impact the economic progress of many nations and provides new opportunities for business expansion (Casson & Lee, 2011). In the financial market, people can trade publicly traded company shares, giving the buyer ownership of a company. The stock market can be essentially represented by its trends known as bull and bear markets (Edwards et al., 2018). A bull market is often represented by steadily rising prices and represents an investment opportunity, whereas a bear market is characterized by a drop in prices and generally results in possible losses for an investor. A third situation, which can also be considered a market condition, occurs when the prices fluctuate within a defined range. This situation is known as a sideways market.

For each of the market conditions, an investor can take different strategies to gain a financial hedge. A long position is advised in a bullish market, whereas a short position is recommended in a bearish market. When a position is long, the holder who owns the stock stands to gain if the price increases. In a short position, the investor loans the stock from a broker and instantly sells it at the current market price. He then waits for a price decrease to repurchase the stock and returns it to the broker, profiting from the price difference. In the sideways market, investors typically avoid making any kind of investment since there is no apparent trend and the market is perceived as unstable.

A variety of charts are used by traders to assist them in making decisions on the stock market (Folger, 2022). A chart is usually used as a simple visual representation of how a stock's price or volume has changed over time. Line charts display only one specific characteristic of a security asset, often the close price over time, while bar charts and candlestick charts show the open, high, low, and close prices.

In the stock market, indexes are widely used since they represent a segment of the financial market and can help investors track market movements. Nearly 500 of the top United States publicly traded corporations are represented by the S&P 500 index, which places a strong focus on market capitalization. The S&P 500 is regarded as one of the strongest indicators of significant United States stocks as well as the entire equities market due to its depth and diversity (Kenton, 2022). Due to these features, it is frequently used as a benchmark for equity funds (Liang, 1999).

#### **2.2. General Machine Learning**

A ML algorithm is a computer procedure that uses incoming data to complete a task without being explicitly programmed to do so. Ultimately, ML seeks to replicate the way humans process sensory information (El Naqa & Murphy, 2015).

ML algorithms are broadly classified into four categories: supervised, unsupervised, semisupervised and reinforcement. The most widely used methods belong to the supervised learning category. In supervised learning, the training data consists of a set of labeled examples, each with its own set of features and corresponding output (Brownlee, 2020). By using the features and the outputs from one dataset (training data), the algorithm predicts the outputs (labels) for another dataset (test data). A supervised learning problem can be further divided into two categories: regression and classification. Classification problems have discrete output variables, whereas regression problems have continuous output variables. Unlike supervised learning, unsupervised learning involves an unlabeled set of features and attempts to group the data into distinct groups. The dataset in semi-supervised learning contains both labeled and unlabeled samples. An initial model is trained on a few labeled samples before being applied iteratively to a larger set of unlabeled data. In reinforcement learning, after defining a set of rules, the algorithm tries to explore different options and possibilities in order to find an optimal solution (Kaelbling et al., 1996). This procedure is carried out through trial and error.

#### **2.3. Stock Market Prevision Using Machine Learning Models**

For a long time, predicting stock prices was considered unachievable. This assumption was supported by the Random Walk Theory (Cootner, 1964) and the Efficient Market Hypothesis (Fama, 1970). The Random Walk Theory states that the price of securities in the stock market evolves randomly. The Efficient Market Hypothesis posits that security prices reflect all available information and therefore, it should be impossible to outperform the overall market. Nonetheless, in recent years, numerous research studies have demonstrated some promise in predicting stock prices by achieving accuracy scores well over the 50% threshold (Ou & Wang, 2009; Chong et al., 2017).

ML approaches have become increasingly popular in estimating asset prices and trends. AI algorithms have made it possible to tackle computationally intensive models that can detect and predict some of the nonlinearity prevalent in financial markets. The research community has shifted its focus to ML models because traditional financial methods used to forecast an asset price, such as the autoregressive method (AR), moving average model (MA), and autoregressive moving average model (ARMA), have proven to have worst performance (Adebiyi et al., 2014) and rely solely on past returns to make a prediction rather than taking advantage of all data available through data mining processes. Not only have they outperformed the classical models on price prediction but also on portfolio management and credit evaluation (Bahrammirzaee, 2010).

The ability of Neural Networks to interpret noisy data has led to its widespread adoption by researchers. Massive amounts of data can be used to train these algorithms, and if they are applied to build DNNs with numerous hidden layers, they may be able to identify more nonlinear relationships in data. The results obtained by using Artificial Neural Networks surpassed those obtained by linear and logical regression models (Altay & Satman, 2005). This algorithm can be traced back to 1988, when White (1988) created a simple ANN to test its efficiency in predicting IBM daily stock returns. His results weren't promising as he couldn't constitute convincing statistical evidence against the Efficient Market Hypothesis.

 Fernandez-Rodrıguez et al. (2000) created a simple ANN model solely based on the time series of the Madrid Stock Exchange General Index. The preceding nine days of the index were utilized as inputs, and the output was a buy/sell signal. The study concluded that the NN trading model outperformed a simple buy-and-hold strategy during a bear market and a stable market, but it underperformed the buy-and-hold strategy during a bull market.

Long Short-term Memory neural networks, which are a type of Recurrent Neural Networks, have been increasingly popular in stock price prediction. These networks were developed by Hochreiter and Schmidhuber (1997) in order to improve performance by addressing the vanishing gradient problem that recurrent networks have when dealing with large data sequences. Shad et al. (2018) compared a DNN and an LSTM on stock price prediction, and the findings revealed that these two NNs topologies performed similarly.

The SVM is a different ML technique that has grown in popularity for predicting stock prices and trends. Vapnik (1999) presented this approach with the ability to divide classes by looking for a hyperplane in higher dimensions. Since then, numerous academics assert that they have successfully used it to forecast stock prices. Kim (2003) forecasted the direction of daily stock price change in the Korea Composite Stock Price Index in 2003 using the SVM model and thirteen technical indicators. Many academics cite his work and consider it successful due to its up to 57% accuracy in predicting the test set, which is much higher than the 50% accuracy threshold (Madge & Bhatt, 2015). However, there is no indication in his study of how his forecasts would translate into financial gains. The one thing we can conclude about Kim's work is that the SVM model outperformed a back-propagation neural network (BPN) and the casebased reasoning (CBR) since he compared the results obtained by all these models.

More recently Kara et al. (2011) compared the performance between an ANN and a SVM in predicting the direction of movement in the daily Istanbul Stock Exchange (ISE) National 100 Index. In his work, the ANN attained an average accuracy of 75.65%, significantly better than the 71.52% achieved by the SVM.

Another type of ML algorithm that has been employed in stock price prediction is the Random Forest. Breiman (2001) proposed it and, like the SVM, it has found some success since it avoids the issue of modeling the underlying distribution and instead focuses on producing accurate predictions for some variables given other factors.

Ballings et al. (2015) benchmarked ensemble methods consisting of RF, AdaBoost, and Kernel Factory against single classifier models such as NNs, Logistic Regression, SVM, and K-Nearest Neighbor using data from 5767 publicly listed European companies. The authors used Area Under the Curve (AUC) as a performance measure for predicting long term stock price direction and reported Random Forest as the top algorithm.

The features used in ML methods are another important topic of discussion since they directly impact the model's predictive power. Most of the input features used by researchers in asset price forecast can be categorized into four categories: Open-High-Low-Close-Volume data, technical indicators, fundamental data, and word embeddings. The first category is an aggregated form of market data. The second category is calculated mathematically from the OHLCV data and, examples include the moving average indicator, relative strength index indicator, and percentage price oscillator indicator (Agrawal et al., 2019). The third category is related to the quantitative variables that are assumed to influence the underlying company's intrinsic worth (Huang et al., 2021). The fourth and last category is used for text analysis to represent words, often in the form of a real valued vector (Mohan et al., 2019).

Although less common, candlestick pattern identification has also been used by researchers to predict stock market movements. For DJIA components from 1992 to 2002, Marshall et al. (2006) found that the candlestick charting strategy did not profit under a holding period of 10 days. On the other hand, Goo et al. (2007) found some profitable candlestick chart patterns in 25 Taiwanese companies. The inconsistent results across the literature may be due to differing stock samples.

## **Chapter 3: Theoretical Framework**

#### **3.1. Artificial Neural Network**

An ANN is a mathematical model that attempts to replicate the structure and capabilities of biological neural networks. A neuron is the fundamental building block of every NN. Similarities in design between the artificial and the biological neuron can be seen in Figure 3.1. In a biological neuron, information enters the neuron via dendrite, which is processed by the body and transmitted via axon. In an artificial neuron, information enters the body via weighted inputs, then all the weighted inputs are summed together with a bias. At the exit, the previously calculated sum is passed through an activation function. The mathematical formulation of the artificial neuron model below demonstrates its simplicity:

$$
y = f\left(\sum_{i=0}^{n} (w_i * x_i) + b\right)
$$
 (3.1)

Where x is the input value, w is the respective weight, **b** is the bias, f is the activation function and  $y$  is the final output value. The characteristics of an artificial neuron are defined by its activation function. The criteria for selecting an activation function are determined by the problem at hand. Sigmoid  $\sigma$  and rectified linear (ReLU) are examples of functions that allow to solve non-linear problems.



Figure 3.1 - Relation between biological and artificial neurons

The perceptron is considered one of the simplest NN architectures. Containing only a neuron inside a hidden layer, it is only capable of solving linear problems however, a multilayer perceptron has the capability of solving non-linear complex problems. There are many ways of interconnecting the neurons inside an ANN but the two most known are the feedforward topology and the recurrent topology. In feed-forward neural networks, information only flows in one direction, from input to output without back loops. On the other hand, recurrent neural networks can not only transmit information in one direction but also backwards. The number of hidden layers in a network must also be considered because it improves a network's representational power and assists in learning highly abstract characteristics. A network with more than one hidden layer is often referred has a DNN.



Figure 3.2 - DNN architecture

The training of an ANN can be described in a few steps. First, all the weights are randomly initialized. Then each node receives data through the inputs, calculates the output and progresses it forwards. After a value is calculated from the last node, the result is compared to the actual target in the training data and the error is determined using a loss function. One of the most popular loss functions used to solve binary classification problems is the binary crossentropy:

$$
Loss = -\frac{1}{n} \sum_{i=1}^{n} y_{true} \cdot log(y_{predicted}) + (1 - y_{true}) \cdot log(1 - y_{predicted}) \quad (3.2)
$$

Where  $\boldsymbol{n}$  is the number of samples (batch),  $\boldsymbol{y}_{true}$  the true value and  $\boldsymbol{y}_{predicted}$  the predicted value. Finally, using backpropagation, the error is passed through every individual neuron and the gradient is computed with respect to all the weights. Then a step is taken in the opposite direction of the gradient and the size of this step is determined by the learning rate. The use of optimizers like the stochastic gradient descent and the adaptive moment estimation can help prevent the training process from staying in a local minimum. A complete loop over the data using the described procedures is an epoch.

#### **3.2. Support Vector Machine**

A SVM is a supervised learning algorithm that can solve regression and classification problems, although it is most commonly used to solve classification problems. In essence, SVM seeks the hyperplane in the original input space capable of correctly separating a given dataset and maximizing the margin between the support vectors and the hyperplane. The hyperplane is a decision boundary that separates classes, and the support vectors are simply the datapoints closer to the hyperplane. These points have a direct impact in the position of the hyperplane.



Figure 3.3 - Linear SVM

In addition to solve linear problems, SVMs can perform a non-linear decision boundary by mapping input data onto high-dimensional feature spaces and calculating the largest hyperplane margin in that space. This mapping is induced by the kernel functions method. In figure 3.4 it is possible to observe how this method helps the SVM algorithm to draw a hyperplane between two different classes.



Figure 3.4 - Non-linear SVM using radial basis function kernel

#### **3.3. Decision Tree**

The Decision Tree is another supervised learning algorithm with the ability to solve regression and classification problems. It reflects a decision-making process through a tree like structure. The algorithm starts from the root of the tree, which represents the entire data sample population, and progressively splits the dataset into smaller subsets as the tree grows. The outcome is a tree comprising decision nodes and leaf nodes. Decision nodes are sub-nodes that split the data into additional sub-nodes and show a decision to be made. Leaf nodes are nodes that do not split and represent a classification. A branch refers to a subsection of the tree.



Figure 3.5 - Definitions inside a DT

The tree split decision is based on a specific criteria. The two most common criteria are information gain and gini index. In information gain, the features with the least entropy are selected as the ones that provide the most information about a class. Entropy quantifies the randomness of the information being processed:

$$
E = \sum_{i=0}^{c} -p_i \cdot log_2(p_i)
$$
 (3.3)

Where  $c$  is the number of different classes in the dataset and  $p_i$  the probability of class  $i$  and  $E$  the calculated entropy. As the gini index quantifies the probability of an element in a dataset being mislabeled if randomly classified, it is the features with the lower gini index that determine the optimal split:

$$
G = 1 - \sum_{i=0}^{c} p_i^2 \tag{3.4}
$$

#### **3.4. Random Forest**

The RF is also a supervised learning algorithm capable of solving regression and classification problems. The algorithm uses the bagging ensemble method to combine multiple DTs in order to get a more accurate prediction. This method allows each individual tree to be trained with a different training subset from the training data and the output is then calculated based on majority voting. Compared to DTs, RFs are slower algorithms that tend to have fewer overfitting problems (Prajwala, 2015). (For more information about the machine learning algorithms consult Burkov, 2019)



Figure 3.6 - RF overview

# **Chapter 4: Methodology**

#### **4.1. Data**

For this study, all data were obtained from https://finance.yahoo.com/ using a data mining program written in Python. This program collected the daily OHLCV historical performance data of each company included in the S&P 500 index from 01-01-2010 through 30-06-2022 (3144 records). These requirements were met by 440 companies in total (Annex A). The distribution of industries among the selected companies is shown in Figure 4.1. The datasets were divided into three subsets: the training set (01-01-2010 through 30-06-2021), the validation set (01-07-2021 through 31-12-2021), and the test set (01-01-2022 through 30-06- 2022). During the test set period, the average return was -15.43%, 79.1% of the companies had a negative return (Figure 4.2), and energy was the best performing sector (Figure 4.3).



Figure 4.1 - Distribution of industries among the studied companies



Figure 4.2 - Distribution of companies returns in the test set period



Figure 4.3 - Average return by industry in the test set period

Based on the OHLCV collected data, all technical indicators and candlestick pattern identification variables were calculated with the TA-Lib Python package using a five-day time window. In total, eighteen technical indicators and eighteen candlestick patterns were used as inputs to the ML models (Annex B). Before the training process began, all the technical indicators were standardized:

$$
z = \frac{X - \mu}{\sigma} \tag{4.1}
$$

Where z is the calculated standardized value, X the observation,  $\mu$  the mean and  $\sigma$  the standard deviation. Candlestick pattern identification variables were not standardized since they are categorical variables where 0 represents the absence of a pattern, 1 represents the identification of a bullish pattern, and -1 represents the identification of a bearish pattern.

#### **4.2. Training and Tunning**

Due to the large number of stocks used in this study, we opted to use grid search with only a few hyperparameter options (Table 4.1). The Scikit-learn Python package was applied to train the SVM and the RF. The Keras Python package was utilized to train the DNN. Regarding the DNN, there were a set of parameters that weren't used in the grid search and in turn had a predefined value (Table 4.2). After the training process was completed for a specific ML model, the validation set accuracy was calculated for all the trained models. The one that achieved the highest accuracy was chosen as the optimal and used to predict the test set. A naive predictor, which predicted randomly if a stock would increase or decrease the next day, was also implemented to help compare the models results.

Table 4.1 - Tunning parameters used for each model during the training phase

Model	<b>Tunning Parameters</b>		
<b>Random Forest</b>	{"n estimators": [100,150,200], "criterion": ["gini", "entropy"]}		
<b>Support Vector Machine</b>	{"C": [0.9, 0.95, 1], "kernel": ["rbf", "poly"]}		
<b>DNN</b>	{"hidden_layers": [2,3], "layers_neurons": [20,25]}		

Table 4.2 - Default parameters used for the DNN during the training phase



# **Chapter 5: Results**

Table 5.1 shows the average accuracy achieved by each ML model and the naive predictor for all the company's stocks during the test period (Annex C). All ML models achieved similar accuracy values and were close to the naive model's value. As a result, the model's inputs provided no insight into the next-day price direction. In addition, Table 5.2 supports this idea since the models achieved an average accuracy close to 50% in each sector.

	<b>DNN</b>	RF	<b>SVM</b>	<b>Naive</b>
Average Accuracy	50.04%	50.22%	!9.21% 49	50.17%

Table 5.1 - Average accuracy obtained by each model during the test period

Table 5.2 - Average accuracy obtained by each model in different sectors during the test period

Sector	<b>DNN</b>	<b>RF</b>	<b>SVM</b>	Naive
<b>Communication Services</b>	51.42%	48.23%	48.75%	51.16%
<b>Consumer Discretionary</b>	50.72%	51.62%	50.03%	50.15%
<b>Consumer Staples</b>	50.29%	50.26%	51.37%	50.69%
Energy	49.61%	46.77%	49.13%	51.16%
<b>Financials</b>	49.21%	49.79%	48.07%	49.63%
<b>Health Care</b>	49.45%	49.40%	48.58%	50.20%
Industrials	50.10%	49.96%	49.40%	50.73%
<b>Information Technology</b>	50.63%	51.72%	48.69%	49.44%
<b>Materials</b>	50.19%	49.03%	49.70%	49.37%
<b>Real Estate</b>	49.71%	51.82%	48.55%	50.13%
<b>Utilities</b>	49.82%	50.23%	50.56%	50.47%

Besides the accuracy values, the returns were also compared between the models and with a buy-and-hold strategy (Annex D). In light of the fact that the ML models outperformed the buy-and-hold strategy (Table 5.3), it is possible to argue that they might be able to anticipate significant market movements. The naive model, however, refutes this idea as it also outperformed the buy-and-hold strategy and its returns are close to those obtained by the other ML models. Furthermore, although almost all models outperformed a buy-and-hold strategy in bearish sectors, none of them outperformed the energy sector which had positive returns (Table 5.4).

Table 5.3 - Average return obtained by each model and the buy-and-hold strategy during the test period

	DNN	<b>DE</b> ' \ L	SVM	<b>Naive</b>	Buy-and-hold
Average return	.34%	6.85%	10.04%	$^{\prime}.14\%$	15.43%
	$\overline{\phantom{0}}$	-n	$\overline{\phantom{0}}$	$-1$	- 1

Table 5.4 - Average return obtained by each model and the buy-and-hold strategy in different sectors during the test period



## **Chapter 6: Conclusion**

Predicting stock prices has always interested the research community. As traditional financial models haven't succeeded in this task, many academics have shifted their focus to ML algorithms. Throughout the literature there have been inconsistent results most probably due to different stock samples used in each research.

 This study tried to answer some of the questions related to the effectiveness of ML models in predicting stock prices by incorporating different features. The OHLCV data from 440 S&P 500 stocks was used to calculate technical indicators and identify candlestick patterns. Then 3 different types of ML algorithms were trained using that information. The average accuracy and return were calculated for each individual stock during the test set and compared between the models, a naive predictor and a buy-and-hold strategy.

 The results observed showed that the average prediction accuracy obtained by each ML model was very similar to the one achieved by the naive model. In terms of average stock returns, the three ML models outperformed the buy-and-hold strategy however the naive predictor delivered comparable results. The model's outperformance was also due to the fact that most stocks had negative returns during the test period. When the models were compared in a sector that yielded positive returns, they underperformed the buy-and-hold strategy.

 With all the results taken into account, we can conclude that all the models could not predict the stock price direction effectively. This was not due to the models weak learning capacity, but due to the non-existent relationship between technical indicators and candlestick patterns with the next day's price direction.

This dissertation is relevant in the sense that it provides a clear direction for future research. The focus of researchers should be on discovering features that have not been extensively used in previous studies, which might impact the direction of stock price movement.
## **References**

- Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. Journal of Applied Mathematics, 2014. <https://doi.org/10.1155/2014/614342>
- Agrawal, M., Khan, A., & Shukla, P. (2019). Stock price prediction using technical indicators: A predictive model using optimal deep learning. International Journal of Recent Technology and Engineering, 8(2), 2297–2305. https://doi.org/10.35940/ijrte.b3048.078219
- Altay, E., & Satman, M.H. (2005). "Stock market forecasting: Artificial neural networks and linear regression comparison in an emerging market", Journal of financial management and analysis 18(2), 18-33.
- Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications, 19(8), 1165–1195. https://doi.org/10.1007/s00521-010- 0362-z
- Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert systems with Applications, 42(20), 7046-7056. <https://doi.org/10.1016/j.eswa.2015.05.013>
- Banerjee, P., Prasad, R. K., & Singh, V. S. (2009). Forecasting of groundwater level in hard rock region using artificial neural network. Environmental Geology, 58(6), 1239-1246. <https://doi.org/10.1007/s00254-008-1619-z>
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Brownlee, J. (2020). A tour of machine learning algorithms. Machine Learning Mastery. Retrieved September 10, 2022, from [https://machinelearningmastery.com/a-tour-of](https://machinelearningmastery.com/a-tour-of-machine%20learning-algorithms/)[machine learning-algorithms/](https://machinelearningmastery.com/a-tour-of-machine%20learning-algorithms/)
- Burkov, A. (2019). The hundred-page machine learning book. Andriy Burkov.
- Casson, M., & Lee, J. S. (2011). The Origin and Development of Markets: A Business History Perspective. The Business History Review, 85(1), 9–37. <https://doi.org/10.1017/S0007680511000018>
- Chong, E., Han, C., & Park, F.C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Syst. Appl., 83, 187-205. https://doi.org/10.1016/j.eswa.2017.04.030

Cootner P. H. (1964). The random character of stock market prices. M.I.T. Press.

- Edwards, R.D., Magee, J., & Bassetti, W.H.C. (2018). Technical Analysis of Stock Trends (11th ed.). CRC Press.<https://doi.org/10.4324/9781315115719>
- El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In machine learning in radiation oncology (pp. 3-11). Springer, Cham.<https://doi.org/10.1007/978-3-319-18305-> 3\_1
- Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486
- Fernandez-Rodrıguez, F., Gonzalez-Martel, C., & Sosvilla-Rivero, S. (2000). On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. Economics letters, 69(1), 89-94. [https://doi.org/10.1016/S01651765\(00\)00270-6](https://doi.org/10.1016/S0165-1765(00)00270-6)
- Folger, J. (2022). Tips for stock charts that enhance your analysis. Investopedia. Retrieved December 22, 2022, from https://www.investopedia.com/articles/trading/10/high-performance-charts.asp
- Goo, Y. J., Chen, D. H., & Chang, Y. W. (2007). The application of Japanese candlestick trading strategies in Taiwan. Investment Management and Financial Innovations, (4, Iss. 4), 49-79.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.<https://doi.org/10.1162/neco.1997.9.8.1735>
- Huang, Y., Capretz, L. F., & Ho, D. (2021). Machine learning for stock prediction based on fundamental analysis. 2021 IEEE Symposium Series on Computational Intelligence (SSCI). https://doi.org/10.1109/ssci50451.2021.9660134
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.
- Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with Applications, 38(5), 5311-5319. <https://doi.org/10.1016/j.eswa.2010.10.027>
- Kenton, W. (2022). S&P 500 index: What it's for and why it's important in investing. Investopedia. Retrieved September 10, 2022, from https://www.investopedia.com/terms/s/sp500.asp
- Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), 307-319. [https://doi.org/10.1016/S0925-2312\(03\)00372-2](https://doi.org/10.1016/S0925-2312(03)00372-2)
- Liang, B. (1999). On the Performance of Hedge Funds. Financial Analysts Journal, 55(4), 72 85. https://doi.org/10.2469/faj.v55.n4.2287
- Madge, S., & Bhatt, S. (2015). Predicting stock price direction using support vector machines. Independent work report spring, 45.
- Marshall, B. R., Young, M. R., & Rose, L. C. (2006). Candlestick technical trading strategies: Can they create value for investors?. Journal of Banking & Finance, 30(8), 2303-2323. <https://doi.org/10.1016/j.jbankfin.2005.08.001>
- Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). https://doi.org/10.1109/bigdataservice.2019.00035
- Ou, P., & Wang, H. (2009). Prediction of Stock Market Index Movement by Ten Data Mining Techniques. Mathematical Models and Methods in Applied Sciences, 3, 28. <https://doi.org/10.5539/mas.v3n12p28>
- Parmar, R. (2018). Training Deep Neural Networks. Towards Data Science. Retrieved September 15, 2022, from [https://towardsdatascience.com/training-deep-neural-networks-](https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964)[9fdb1964b964](https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964)
- Prajwala, T. R. (2015). A comparative study on decision tree and Random Forest using r tool. IJARCCE, 196–199. https://doi.org/10.17148/ijarcce.2015.4142
- Premand, S. (2021). Support Vector Machine: Beginners Guide to support vector machine. Analytics Vidhya. Retrieved September 15, 2022, from [https://www.analyticsvidhya.com/blog/2021/06/support-vector-machine-better](https://www.analyticsvidhya.com/blog/2021/06/support-vector-machine-better-understanding/)[understanding/](https://www.analyticsvidhya.com/blog/2021/06/support-vector-machine-better-understanding/)
- Qiu, M., & Song, Y. (2016). Predicting the direction of stock market index movement using an optimized artificial neural network model. PloS one, 11(5), e0155133. <https://doi.org/10.1371/journal.pone.0155133>
- Shah, D., Campbell, W., & Zulkernine, F. H. (2018). A Comparative Study of LSTM and DNN for Stock Market Forecasting. 2018 IEEE International Conference on Big Data (Big Data), pp. 4148-4155. https://doi.org/doi:10.1109/bigdata.2018.8622462
- Sruthi, R. (2021). Random Forest: Introduction to random forest algorithm. Analytics Vidhya. Retrieved September 15, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
- Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10, 988–999. https://doi.org/10.1109/72.788640
- White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns. International Conference on Neural Networks, 2, 451-458. <https://doi.org/10.1109/ICNN.1988.23959>
- Zhang, G. (2018). What is the kernel trick? Why is it important? Medium. Retrieved September 15, 2022, from [https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-i](https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-i%09mportant-98a98db0961d)  [mportant-98a98db0961d](https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-i%09mportant-98a98db0961d)

## **Annexes**

























Annex B: Table containing the features used to train the machine learning models

Symbol	Random Forest	<b>Support Vector Machine</b>	Deep Neural Network	Naive
Α	55.74%	54.92%	50.00%	50.82%
AAL	54.10%	45.90%	47.54%	55.74%
AAP	52.46%	51.64%	49.18%	50.82%
AAPL	54.92%	51.64%	50.00%	44.26%
ABC	46.72%	47.54%	45.90%	52.46%
ABMD	47.54%	51.64%	50.82%	49.18%
ABT	54.10%	52.46%	50.00%	51.64%
<b>ACN</b>	49.18%	49.18%	45.90%	51.64%
ADBE	54.92%	45.08%	52.46%	61.48%
ADI	50.82%	52.46%	50.82%	43.44%
ADM	42.62%	49.18%	55.74%	57.38%
<b>ADP</b>	57.38%	47.54%	51.64%	54.10%
<b>ADSK</b>	48.36%	40.98%	59.84%	53.28%
AEE	50.82%	49.18%	50.82%	51.64%
AEP	49.18%	59.02%	60.66%	54.10%
<b>AES</b>	42.62%	45.90%	50.00%	53.28%
<b>AFL</b>	44.26%	40.16%	53.28%	42.62%
AIG	47.54%	58.20%	50.00%	47.54%
AIZ	45.08%	52.46%	48.36%	48.36%
AJG	42.62%	46.72%	49.18%	43.44%
AKAM	49.18%	48.36%	47.54%	50.82%
<b>ALB</b>	49.18%	50.00%	46.72%	40.98%
<b>ALGN</b>	41.80%	37.70%	54.92%	54.92%
<b>ALK</b>	44.26%	45.08%	56.56%	53.28%
<b>ALL</b>	51.64%	53.28%	51.64%	49.18%
<b>AMAT</b>	50.00%	47.54%	45.08%	52.46%
AMD	56.56%	50.82%	54.10%	50.00%
AME	49.18%	50.00%	50.00%	54.92%
AMGN	51.64%	52.46%	40.16%	40.98%
AMP	53.28%	48.36%	50.00%	50.00%
AMT	46.72%	45.08%	48.36%	56.56%
AMZN	43.44%	49.18%	42.62%	45.08%
<b>ANSS</b>	55.74%	44.26%	51.64%	51.64%
<b>AON</b>	44.26%	45.90%	48.36%	48.36%
AOS	59.02%	50.82%	54.92%	54.92%
<b>APA</b>	51.64%	52.46%	53.28%	45.08%
<b>APD</b>	46.72%	45.90%	46.72%	51.64%
<b>APH</b>	57.38%	55.74%	50.82%	43.44%
ARE	53.28%	44.26%	43.44%	54.10%
<b>ATO</b>	53.28%	51.64%	52.46%	56.56%
<b>ATVI</b>	49.18%	45.08%	56.56%	52.46%

Annex C: Table containing the average accuracy obtained by each model in the test period



















		<b>Support Vector</b>	Deep Neural		
symbol	Random Forest	Machine	Network	Naive	Buy-and-hold
Α	$-0.44%$	$-10.32%$	$-10.11%$	$-8.00%$	$-24.65%$
AAL	$-1.85%$	$-35.69%$	$-32.30%$	6.08%	$-30.51%$
AAP	$-1.26%$	$-7.00%$	$-15.93%$	$-10.34%$	$-25.37%$
AAPL	$-0.17%$	0.10%	$-17.78%$	$-21.94%$	$-23.50%$
ABC	0.42%	8.85%	2.56%	13.38%	8.26%
<b>ABMD</b>	$-26.87%$	$-12.12%$	$-24.06%$	$-17.22%$	$-29.98%$
ABT	$-16.51%$	$-10.79%$	$-17.97%$	$-4.66%$	$-21.53%$
<b>ACN</b>	$-13.80%$	$-25.03%$	$-13.57%$	$-15.71%$	$-31.29%$
ADBE	2.67%	$-23.62%$	$-6.84%$	22.94%	$-34.71%$
ADI	$-6.62%$	$-12.97%$	$-0.85%$	$-26.47%$	$-16.87%$
ADM	$-0.11%$	11.76%	7.18%	17.50%	13.77%
<b>ADP</b>	13.39%	$-17.85%$	$-5.66%$	3.16%	$-13.04%$
<b>ADSK</b>	$-22.24%$	$-36.36%$	8.28%	$-6.43%$	$-37.67%$
AEE	2.57%	$-4.44%$	0.49%	$-0.80%$	1.20%
<b>AEP</b>	2.28%	10.83%	13.94%	8.23%	7.63%
<b>AES</b>	$-23.74%$	$-11.96%$	$-7.39%$	$-4.24%$	$-16.49%$
<b>AFL</b>	$-3.77%$	$-12.84%$	5.06%	$-5.85%$	$-4.82%$
AIG	$-21.56%$	13.78%	0.00%	$-12.41%$	$-11.48%$
AIZ	1.06%	12.73%	3.85%	$-2.50%$	9.34%
AJG	$-7.29%$	$-6.50%$	$-2.03%$	$-7.56%$	$-2.11%$
<b>AKAM</b>	$-8.04%$	$-18.03%$	$-14.25%$	$-7.49%$	$-21.60%$
ALB	$-9.28%$	$-11.76%$	$-13.20%$	$-29.57%$	$-10.47%$
<b>ALGN</b>	-50.99%	$-54.60%$	$-13.39%$	$-18.03%$	$-62.63%$
<b>ALK</b>	$-36.23%$	$-21.88%$	$-6.94%$	$-2.00%$	$-25.13%$
<b>ALL</b>	10.13%	2.60%	$-4.42%$	0.33%	6.05%
<b>AMAT</b>	$-24.25%$	$-40.31%$	$-46.54%$	$-13.71%$	$-42.51%$
AMD	8.68%	$-17.04%$	$-8.07%$	$-33.86%$	$-48.09%$
AME	$-15.35%$	$-7.05%$	$-17.56%$	4.13%	$-23.06%$
AMGN	3.88%	10.56%	1.97%	$-3.53%$	8.27%
AMP	0.35%	$-19.30%$	$-10.36%$	$-22.61%$	$-21.12%$
AMT	$-13.10%$	$-10.86%$	$-15.49%$	$-1.89%$	$-11.60%$
AMZN	$-39.42%$	$-14.14%$	$-29.20%$	$-31.24%$	$-36.08%$
<b>ANSS</b>	10.77%	$-37.22%$	$-15.02%$	$-11.87%$	$-38.95%$
<b>AON</b>	$-5.23%$	$-7.11%$	$-15.28%$	$-16.81%$	$-8.31%$
AOS	1.72%	$-23.21%$	$-8.92%$	$-13.43%$	$-34.43%$
APA	18.20%	30.24%	81.22%	$-2.67%$	28.65%
<b>APD</b>	$-16.40%$	$-19.26%$	$-11.15%$	0.20%	$-17.81%$
APH	$-4.43%$	$-11.16%$	$-20.12%$	$-30.43%$	$-25.58%$
ARE	$-22.49%$	$-28.04%$	$-30.17%$	$-12.17%$	$-34.25%$
<b>ATO</b>	$-1.32%$	6.53%	2.60%	12.98%	5.61%

Annex D: Table containing the average accuracy obtained by each model and the buy-andhold strategy in the test period

















