Hochschule für Technik
und Wirtschaft Berlin
University of Applied Sciences

Seminar Paper

in the Master Program:
Financial Mathematics, Actuarial Science and Risk Management

Forecast of stocks in the ESG market using Recurrent Neural Networks (RNN)

Course: Deep Learning

Annalena Fink

Lecturer: Dr. Alla Petukhina

submitted on February 28, 2023

Contents

1 Introduction

In the first section of the paper, the study's problem statement is presented. Consequently, the aim of this paper is established, along with its structure. Moreover, this chapter covers the basic concepts that are necessary to understand the paper's content. Specifically, the notion of ESG is introduced, which plays a crucial part in the subsequent analysis. Lastly, the primary methods for time series prediction are discussed, and in this context, recurrent neural networks are introduced.

1.1 Objective

For many years, trading in shares has been a popular method for achieving high financial returns in a short or long-term period. The success of this type of investment is contingent upon selecting the right stock at the right time, which is expected to increase in value over the future. The global concern for sustainability has led to the emergence of a new investment trend, namely investing in companies that demonstrate high environmental, social, and governance (ESG) standards. The growth of ESG investing has been remarkable in recent years, as it provides investors with a way to align their investments with their values and contribute to a sustainable future [21] [20]. However, investing in ESG stocks requires a proper understanding of the market and a reliable prediction of the stock's future performance as the future course of share prices remains uncertain. Consequently, trading in shares carries a high risk of losing staked assets in the event of a decline in value. In response to this challenge, methods have been developed to predict share price trends, thereby reducing uncertainty, preventing significant financial losses, and achieving high profits.
Various procedures and forecasting methods are available today, which are intended to enable reliable predictions based on the historical development of share prices. These methods include classical autoregressive models, artificial neural networks, and combined forecasting methods, which increase the selection of support tools for investment decisions [13]. However, the reliability of these forecasts varies, and choosing the best method is critical to avoid wrong decisions regarding share pur-

chases or sales. Market participants are faced with a variety of choices and by comparing different methods, investment decisions can be made and executed early based on reliable forecasts.

The study seeks to address the challenge of identifying the most effective forecasting method for stock price predictions. The selected methods are both classical and machine learning techniques, allowing for a comprehensive analysis of different approaches. The classical models, including linear regression and ARIMA, are established techniques with a history of successful implementation in the field of stock price forecasting. The selection of these models serves as a benchmark for evaluating the performance of the Long Short Term Memory (LSTM) model, which represents a new and promising approach to stock price prediction. The use of historical stock price data as the basis for all models ensures a reliable and valid comparison of their respective forecasting capabilities. By individually predicting two stock prices for each method, the study aims to provide a detailed analysis of the strengths, limitations, and disadvantages of each method. This analysis will enable investors to make informed decisions regarding the selection of a forecasting method that best suits their investment goals and risk tolerance. The suitability of the LSTM model as a new technique for stock price forecasting is a key area of investigation in the paper.

1.2 Literature Review

Time series forecasting is a crucial aspect of stock price prediction and investment decision-making. Over the years, various techniques have been developed and used for time series forecasting, including Linear Regression and ARIMA. In recent years, the application of deep learning techniques, particularly LSTMs, has gained significant attention due to their ability to model complex, non-linear relationships in time series data.

Several studies have compared the performance of these techniques in time series forecasting, with varying results. For example, Jadhav and Wadhai (2020) [12] aimed to compare the performance of several machine learning techniques in predicting stock prices. The authors used data from the National Stock Exchange of India and compared the accuracy of the models in terms of mean absolute error (MAE) and root mean squared error (RMSE). The results showed that support vector regression (SVR) and long short-term memory (LSTM) were the most accurate methods, with SVR performing slightly better. In contrast, Chong, Han and Lee [12] compared the performance of multiple linear and non-linear models, including ARIMA and LSTMs, and found that the ARIMA model was the most accurate for stock price forecasting. One study by Sandra Litz (2019) [13] examined the effectiveness of using predictive neural networks (PrNN) for stock price prediction. Litz compared the PrNN approach to traditional time series forecasting methods, including Linear Regression and ARIMA, and found that the PrNN model outperformed these methods in terms of accuracy and ability to capture non-linear patterns in the data. Litz's study also highlighted the importance of properly selecting and preprocessing input features for the deep learning model to achieve optimal results. Overall, while numerous studies have compared the performance of different time series forecasting techniques, the results have been inconsistent. Therefore, further research is needed to identify the most suitable forecasting technique for different types of stock price data, taking into account the impact of model parameters and preprocessing techniques. The present study aims to contribute to this research by comparing the performance of Linear Regression, ARIMA, and LSTM models on price data of two stocks in the ESG market, using R and Python programming languages.

1.3 ESG

Environmental, social, and governance (ESG) factors have become increasingly important for investors seeking to make informed decisions about where to allocate their capital. An increasing number of companies are incorporating ESG principles into their business operations, recognizing the importance of sustainability, social responsibility, and good governance. This trend is driven by a number of factors, including pressure from investors, customers, and regulatory bodies. Investors and stakeholders have recognized the potential benefits of investing in companies with strong ESG practices. For example, companies that prioritize environmental sustainability may be better positioned to mitigate climate change risks and take advantage of opportunities in the transition to a low-carbon economy. Similarly, companies with strong social practices may be better equipped to attract and retain top talent and maintain positive relationships with customers and other stakeholders. To demonstrate their commitment to ESG principles, many companies are now disclosing information about their ESG practices, such as carbon emissions, employee diversity, and community engagement. This information is typically included in annual reports, sustainability reports, and other corporate disclosures.
Moreover, rating agencies play a key role in the ESG market by providing evaluations of companies' ESG performance. These ratings can be used by investors to make informed decisions about which companies to invest in based on their ESG practices. Rating agencies such as MSCI, Sustainalytics, and Bloomberg have developed their

own ESG ratings and provide reports on companies' ESG performance. Companies that score well on ESG ratings are often seen as more attractive to investors, as they are perceived to be better managed and more sustainable over the long term.
Overall, the ESG market represents a significant shift in how investors evaluate companies and allocate capital. By incorporating ESG factors into their investment decisions, investors have the potential to promote positive change and drive improvements in corporate sustainability practices. Likewise, companies with strong ESG practices may be better positioned to attract and retain capital from ESG-focused investors.

1.4 Statistical Methods and Models

A time series is a set of observtions y_t, each one being recorded at a specific time t. Time series models for observered data y_t is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables Y_t, of which y_t is postulated to be a realization [3]. In statistical time series analysis, the assumption is made that the mathematical model of a time series follows a stochastic process [22]. A stochastic process refers to a sequence of random variables Y_t ordered over time. In general, the random variables of the stochastic process are represented by $Y_t, t \in T$, where T denotes the observation period and t represents a specific time point [5]. An observed value y_t of the time series is subject to a random variable Y_t. Thus, a time series can be seen as the realization of a random variable. The current time series is only a realization of a sample with a limited time frame of the stochastic process [5].
Time series models have diverse applications, one of which is predicting a series using observations of another, particularly forecasting future closing values using historical pricing data. Financial time series, such as stock prices, are subject to investigation using different time series analysis techniques to facilitate the prediction of future developments based on the outcomes of the analysis.
A possible approach to model a time series and make predictions is through the use of a linear regression model. This involves conducting a regression analysis to explore the relationship between two features. Another popular time series analysis method is ARIMA (Autoregressive Integrated Moving Average), which models a time series as a combination of autoregressive (AR) and moving average (MA) processes. Further explanation of those models are provided in chapter 2.3 and 2.4.
Recurrent Neural Networks (RNNs) are a type of neural network that are particularly used for time series analysis. RNNs can take into account the sequential nature of time series data, making them more effective at capturing patterns and

making predictions. They are characterized by their ability to maintain an internal state or memory that allows them to process sequential data. This is achieved through the use of recurrent connections between the neurons in the network, which allow information to be passed from one time step to the next. This paper utilized one speicific type of RNN - the Long Short Term Memory (LSTM). LSTMs are specifically designed to handle the problem of vanishing gradients, which can occur in traditional RNNs when trying to learn long-term dependencies. It is capable of bridging time intervals, even when dealing with noisy, complex input sequences, without sacrificing its ability to capture short time lag capabilities. This is accomplished through the use of an optimized algorithm that ensures a constant error flow through specialized units, preventing the issues of vanishing or exploding gradients. However, to prevent issues with gradient computation, the algorithm truncates at certain architecture-specific points, without affecting the long-term error flow [11]. A mathematical explanation and deeper insight to LSTM will be provided later on in chapter 2.5.

2 Forecasting Model

The purpose of this chapter is to discuss the process of selecting data and the basic concepts of the three forecasting methods and how they are implemented.

2.1 Stock Selection

The present study utilizes data pertaining two stocks traded on the ESG market. Prior to commencing data collection, a suitable metric was identified to gauge the performance of companies with respect to environmental, social, and governance factors. The S&P CSA emerged as the chosen metric after an extensive review of available literature and analysis.
The S&P Global Corporate Sustainability Assessment (CSA) [19] is a widely recognized and comprehensive annual evaluation of companies' sustainability practices and performance. It assesses companies based on their performance across various environmental, social, and governance dimensions, such as carbon emissions, labor practices, and board diversity, among others. The CSA utilizes a rigorous methodology that includes both quantitative and qualitative analysis, incorporating publicly available information and direct engagement with the assessed companies. The resulting scores are used by investors, corporations, and other stakeholders as an important reference point for evaluating sustainability performance and identifying potential risks and opportunities. A growing number of sustainability focused investor use the CSA as a reference tool to compare sustainability performance which directly influences their investment decision-making process. Such ratings can help investors to align their investments with their values and beliefs. As of January 21, 2022, the CSA drew participation from 7,554 companies across the globe [18].
During the data collection process for this project, the S&P Sustainability Yearbook of 2022 was consulted to identify top-performing companies in each industry. The Yearbook classifies companies into three categories - Gold Class, Silver Class, and Bronze Class - and includes the top 15% of companies from each industry. A filter was applied to identify German companies within the Gold Class, resulting in a shortlist of four companies. From this list, two companies were selected at random

for this project - ALLIANZ SE with an ESG score of 93 and SAP SE with a score of 79 [18].
Allianz SE is a leading German financial services company operating in the insurance and asset management sectors. The company operates in over 70 countries worldwide and has a strong presence in Europe and Asia. Allianz is committed to sustainability and has implemented various initiatives to reduce its environmental impact and promote social responsibility [2]. SAP SE is a multinational software corporation based in Germany that specializes in enterprise software solutions. The company's software solutions are designed to help businesses optimize their operations and improve productivity. SAP is present in over 180 countries and has a strong focus on sustainability, including reducing its carbon footprint and promoting sustainable business practices among its customers [16].

2.2 Data Exploration

Prior to transitioning to the adopted models, this chapter will provide a concise overview of the datasets utilized and an exploration of certain features. Following the identification of the two stocks, the data acquisition process involved accessing Yahoo Finance [23]. This platform provided cost-free access to historical stock data. The selected timeframe for data collection spans six years, starting on January 4th, 2016, and concluding on December 30th, 2021. The historical stock prices were downloaded from Yahoo Finance in CSV format, which enabled seamless processing by the utilized programs Python and R. Notably, the selected timeframe corresponds to 1,522 trading days, given the unavailability of trading activities on weekends and public holidays, as it is typical for most public trading platforms. Upon importing the data into R and Python, two datasets were generated, which exhibit the format as seen in figure 2.1.

	Date	Open	High	Low	Close	Adj.Close	Volume
1	2016-01-04	160.15	160.15	155.50	156.00	111.40583	2742295
2	2016-01-05	158.00	158.40	154.30	156.50	111.76291	1521606
3	2016-01-06	157.15	158.15	155.95	157.00	112.11997	1686090
4	2016-01-07	153.15	154.95	151.55	153.75	109.79903	2454430

Figure 2.1: This figure shows an excerpt of the datasets implemented in R.

The illustration above indicates the presence of seven distinctive features, namely: Trading Date, Open and Close prices, High and Low values, Adjusted Closing (Adj Close) prices, and Volume. Open and Close values denote the initial and concluding

trading prices for the same duration, respectively. High and Low values pertain to the maximum and minimum stock prices within a specified period. The Adj Close prices account for corporate actions like dividends and stock splits. The Volume feature represents the overall extent of trading activities.
In order to evaluate the effectiveness of the algorithms employed, a division of the available data into test and training sets was necessary. The conventional ratio of 80% for training data and 20% for test data was deployed for this purpose. Thus, there are 1,218 observations for the training data and 304 observations for the test data. Preceding the construction of the regression models, pre-processing of the raw data was undertaken. This involved the inclusion of timesteps for the regression model, with the first timestep denoted as $t(1) = 1$. Following this, the data was partitioned into test and training sets according to the aforementioned ratio, resulting in the creation of new data frames. The closing price feature was designated as a time series, as it served as the primary target for modeling. Consequently, two datasets were produced, exhibiting the structure seen in figure 2.2.

	Date	Open	High	Low	Close	Adj.Close	Volume	observation	timeseries
1	2016-01-04	71.50	71.76	70.10	70.58	63.26058	4580799	1	70.58
2	2016-01-05	71.74	71.74	70.17	71.43	64.02243	2771848	2	71.43
3	2016-01-06	73.50	73.78	71.24	72.05	64.57813	3767568	3	72.05
4	2016-01-07	69.60	71.94	69.50	71.34	63.94176	4636705	4	71.34

Figure 2.2: This figure shows an excerpt of the new data frames implemented in R including a timeseries.

As illustrated, two new columns were appended to the datasets, with one consisting of the time series and the other representing the closing price as a time series.
To provide a brief graphical overview of the two stocks used in the model, a plot of the closing prices of both stocks over the preceding five-year period is presented in figure 2.3. The plots reveal an overall upward trend in the stock prices. Notably, in the year 2020, significant fluctuations were observed in both stocks. Specifically, Allianz experienced a substantial drop of approximately 80€, resulting in closing prices slightly above 100€ in the initial stages of the year. On the other hand, SAP witnessed an increase in prices, peaking at 140€ during the same period. These discruptions can be attributed to the outbreak of the COVID-19 virus, which led to significant losses for insurance companies like Allianz, while other companies, such as SAP, were able to rapidly recover and attain new all-time highs. As of the end of 2021, both stocks have regained closing prices close to their pre-pandemic levels.

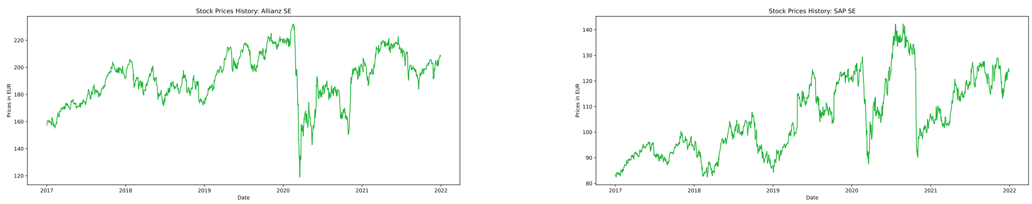

Figure 2.3: This figure shows the historic closing prices of Allianz SE and SAP SE over a period of five years.

2.3 Linear Regression

Linear Regression Model is a statistical approach for modelling the relationship between a dependent variable denoted as y and one or more independent variables denoted as x. Let X and Y be two continuous variables from which a two-dimensional sample $(X_t, Y_t), t = 1, ..., n$ is obtained. The objective of regression analysis is to capture the relationship between these variables through a simple model. A relationship between the variables X and Y is described by the following simple linear regression model [17]:

$$Y_t = \phi_0 + \phi_1 X_t + \epsilon_t \tag{2.1}$$

where Y_t is the dependent variable, X_t is the independent variable, ϕ_0 is the intercept of the population regression line and ϕ_1 is the slope of the population regression line. ϵ_t is the error term with $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma^2$, $\forall t$, $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. It is assumed that X is not stochastic.

The Ordinary Least Squares method is utilized to estimate the parameters of the regression line. The square of the distance between the point and the line in the perpendicular direction (residual) is chosen as a measure of the goodness of fit of the line to an individual point (X_t, Y_t):

$$r_t := y_t - (\phi_0 + \phi_1 x_t) \tag{2.2}$$

The ordinary least square estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared errors made in predicting Y given X. The method of least squares is employed to determine the optimal parameters for linear regression, where the objective is to minimize the square of the distance[8]:

$$\min_{\phi_0, \phi_1} \sum_{t=1}^{n} (y_t - \phi_0 + \phi_1 x_t)^2. \tag{2.3}$$

For the purpose of forecasting a time-series of a stock index, histroical data on Y is utilized to construct the OLS estimators ϕ_0 and ϕ_1. In general, Y_{T+1} will denote the forecast of Y_T based on information through period T using a model estimated with data through period T.
As part of the project, a linear regression line was determined to represent the closing price of the stock as a function of time. Using this line, the stock price trend over the entire period can be described, allowing for the prediction of future price movements. Within the regression model, the closing prices of the stock represent the dependent variable to be examined in relation to time, which is the independent variable. Time is recorded as individual time steps, denoted by $t = 0, 1, ..., n$, where each time step corresponds to one trading day. Based on the observed values, the linear regression model is determined using Formula (2.1). However, first, the unknown ϕ_0 and ϕ_1 of the regression line must be estimated. Using the criterion of the least square method in Formula (2.3), these values are determined so that the regression line best approximates the values of the stock price. For the closing price of Allianz SE, the given model determines the parameters $\phi_0 = 149.47846$ and $\phi_1 = 0.04749$. With regard to the current stock price, this means that the price of the stock at t_0 is 149.47846 €. Since the regression coefficient ϕ_1 describes the increase of the trend line, the estimated closing price increases by ϕ_1 with each time step t. Thus, the linear regression model for the stock price of Allianz SE is given by formula (2.4):

$$\hat{y}_t = 149.47846 + 0.04749t. \tag{2.4}$$

The linear regression model for the stock price of SAP SE is subsequentially given by formula (2.5):

$$\hat{y}_t = 69.18595 + 0.04594t. \tag{2.5}$$

Subsequently, the regression model will be tested for its predictive accuracy and reliability. To evaluate the model, the test dataset will be used. Based on the established regression models in formula (2.4) and (2.5), predictions regarding the development of the stock price will be created. The number of prediction values corresponds to the size of the test data. The prediction of a value at a desired time is achieved by substituting the trading day to be predicted for t in formula (2.4) and (2.5). The evaluation of the predictions and the assessment of the model's quality will be conducted in chapter 3.

2.4 ARIMA

As previously described, the Autoregressive Integrated Moving Average (ARIMA) model is a commonly used approach for forecasting time-series data and was introduced by Box and Jenkins in 1970 [15]. In the ARIMA model, the future value of a variable is a linear combination of past values and past errors, expressed as follows:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ... - \theta_q \epsilon_{t-q} \tag{2.6}$$

where, Y_t is the actual value and ϵ_t is the random error at t, ϕ_i and θ_j are the coefficients, p and q are integers that are often referred to as autoregressive and moving average, respectively [1].
In research involving time-dependent variables, it is crucial to test the data for stationarity. A stationary series is one in which there is no correlation between previous values. Regression analysis using trended (non-stationary) variables can yield misleading and spurious results. Various formal methods have been proposed by researchers to test for stationarity. In this project, the test developed by Dickey and Fuller (1979, 1981) was employed, which is considered the most significant and widely used method in time series analysis. This test examines the null hypothesis of an ARIMA against stationarity and alternatively [14]. The calculated p-value for the time series used in this project was less than 0.05, leading us to reject the null hypothesis and conclude that the underlying series is stationary.

2.5 LSTM

Recurrent Neural Networks (RNNs) have a key advantage over traditional timeseries modelling in their ability to utilize contextual information in mapping input and output sequences. However, this benefit is limited in standard RNN architectures due to the constraint on the range of context that can be practically accessed. The issue arises from the exponential decay or amplification of the influence of a specific input on the hidden layer, and ultimately on the network output, as it cycles through the recurrent connections of the network. This phenomenon is commonly known as the vanishing gradient problem. As introduced by Alex Graves [10], LSTMs are designed to address this problem by introducing a memory cell that can store information over multiple time steps and gating mechanisms that control the flow of information into and out of the cell. These gates are trained to selectively allow or block the flow of information based on the current input and the previous hidden state, enabling the network to remember important information and discard irrelevant information

over time. Moreover, LSTMs use a set of specialized activation functions, known as the hyperbolic tangent and the sigmoid functions, that are less susceptible to the vanishing gradient problem. These functions are bounded between -1 and 1 or 0 and 1, respectively, ensuring that the gradients do not become too small or too large during backpropagation. Overall, the memory cell and gating mechanisms in LSTMs, combined with the use of specialized activation functions, allow the network to model long-term dependencies in sequential data without suffering from the vanishing gradient problem. This makes LSTMs particularly effective for tasks such as speech recognition, natural language processing, and time series prediction, where long-term dependencies are crucial for accurate predictions.
Sepp Hochreiter and Jürgen Schmidhuber introduce the mathematical LSTM architecture in their book [11] as follows:
Let x_t denote the input at time step t, h_{t-1} denote the hidden state at time step $t-1$, and y_t denote the output at time step t. The LSTM cell consists of a memory cell c_t that stores information over time, and three gating mechanisms: an input gate i_t, a forget gate f_t, and an output gate o_t.
The LSTM cell is governed by the following equations:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2.7}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2.8}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{2.9}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{2.10}$$

$$h_t = o_t \times \tan(c_t) \tag{2.11}$$

$$y_t = W_y h_t + b_y \tag{2.12}$$

where σ is the sigmoid function, tan is the hyperbolic tangent function, W and U are weight matrices, and b is the bias vector. The input gate i_t controls how much of the input x_t should be stored in the memory cell, while the forget gate f_t controls how much of the previous memory cell value c_{t-1} should be retained. The forget gate takes as input the current input x_t and the previous hidden state h_{t-1} and outputs a value between 0 and 1 for each element of the memory cell. The output gate o_t controls how much of the current memory cell value c_t should be used to compute the output y_t.

The forget gate, input gate, and output gate are all trained through backpropagation to optimize the overall performance of the LSTM architecture on a given task. By selectively controlling the flow of information, these gating mechanisms enable LSTMs to model long-term dependencies in sequential data.

For the purpose of this project, a LSTM network architecture was implemented in Python using the Keras deep learning library [6]. The presented algorithm [9] implements a sequential model that permits stacking layers in a sequential manner. The architecture includes a LSTM layer with 100 network units, where the return_sequence parameter is set to true. As a result, the output of this layer corresponds to a sequence of equal length as the input sequence. The input_shape parameter specifies the dimensions of the input data, which consists of a 2D array with dimensions of (number of time steps, 1). Another LSTM layer with 100 network units is added, where return_sequence is set to false, meaning that only the last output of the sequence is returned. Additionally, a fully connected neural network layer with 25 network units is included and is connected to the previous layer. Furthermore, a densely connected neural network layer with one network unit is appended as the output layer of the model. Overall, this LSTM network architecture consists of two stacked LSTM layers with 100 network units each, followed by two fully connected layers with 25 and one network units respectively. The use of LSTM layers allows the model to effectively capture long-term dependencies in time series data, while the fully connected layers allow for non-linear relationships to be learned between the input and output data. The details of the LSTM model's training and results will be discussed in chapter 3.1.3.

3 Results

3.1 Findings and Interpretation

In this section of the paper an evaluation of the predictions utilizing the earlier presented timeseries methods is discussed. The objective is to ascertain the most dependable form and model for forecasting stock prices through comparison. Ultimately, a comprehensive assessment of the predictions is performed to determine the most reliable model and forecasting method for achieving accurate predictions.

3.1.1 Linear Regression

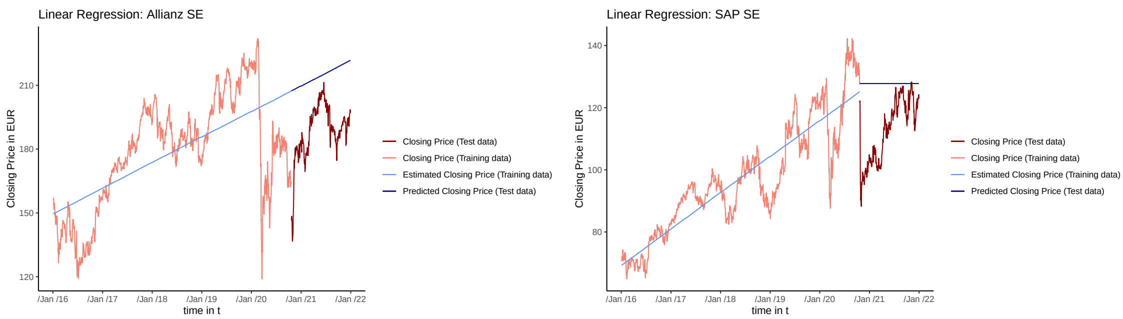

Figure 3.1: This figure shows the results of the linear regression model for Allianz SE and SAP SE.

The results obtained from the linear regression model and its predicted stock prices are displayed in figure 3.1 above. The light blue line illustrates the estimated closing price of the training data, the orange line corresponds to the actual closing price of the training data, the red line depicts the actual closing price of the test data, and the dark blue line represents the predicted closing price of the test data. It is evident that the regression model adequately predicts the training data for both models, as the lines consistently follow a similar pattern and overlap seamlessly.
According to the linear regression model established in section 2.3, a predicted stock price of Allianz and SAP is obtained for 1,218 training data, as shown in figure 3.1. Both derived models have a positive y-intercept and a constant slope, resulting in

an increase in the predicted stock price with each time step. The prediction for the 304 trading days indicates a linear increase in the predicted stock price based on the regression function. However, the actual trend of Allianz's stock price exhibits fluctuations, with a downward trend observed in mid-2020 and at the end of 2021. Although the regression line predicts an upward trend in the stock price, the real trend appears to be changing. A similar pattern can be observed for the SAP stock. The actual closing price of the SAP stock crashes at the end of 2020 whereas the predicted stock price based on the regression function predicts an upward trend.

3.1.2 ARIMA

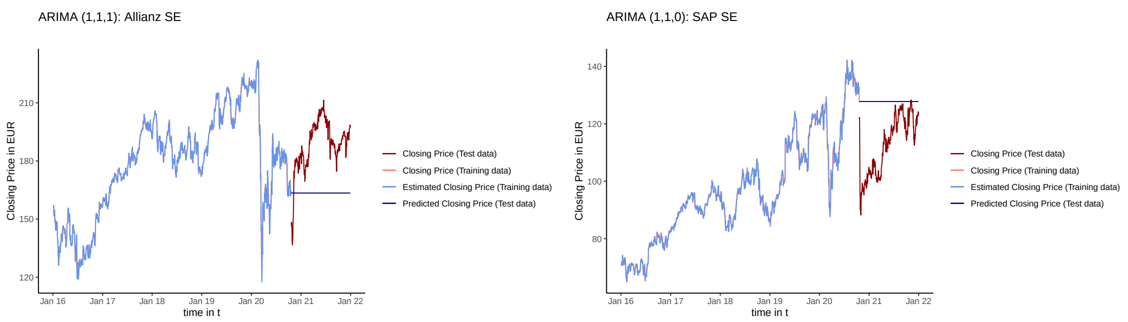

Figure 3.2: This figure shows the results of the ARIMA model for Allianz SE and SAP SE.

Figure 3.2 presented above displays the results obtained from the ARIMA model and its predicted stock prices. The light blue line represents the estimated closing price of the training data, the orange line corresponds to the actual closing price of the training data, the red line depicts the actual closing price of the test data, and the dark blue line represents the predicted closing price of the test data. It can be observed that for both models, the prediction of the training data is satisfactory with the ARIMA model. The lines follow a similar trend and overlap continuously. However, the prediction of the close value of the test data is not precise, as it is a constant price.

For Allianz, the best ARIMA model has the parameters 1, 1, and 1, indicating that the model is constructed using one lagged value and one error term. In forecasting form, the best model selected can be expressed as follows:

$$Y_t = \phi_1 Y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1} \tag{3.1}$$

where, $\epsilon_t = Y_t - \hat{Y}_t$ (i.e., the difference between the actual value of the series and the forecast value).

On the other hand, for the SAP stock, the best ARIMA model has the parameters

1, 1, and 0, implying that the model is constructed using one lagged value and no error term. In forecasting form, the best model selected can be expressed as follows:

$$Y_t = \phi_1 Y_{t-1} + \epsilon_t \tag{3.2}$$

where, $\epsilon_t = Y_t - \hat{Y}_t$ (i.e., the difference between the actual value of the series and the forecast value).

Both stock prices are non-stationary and influenced by random values, as they are a random walk. Accordingly, for both ARIMA models, the final value of the training data set is used as a constant value for the forecast. The predicted final price for all time steps to be forecasted is approximately 162€ for Allianz and approximately 128€ for the SAP stock.

3.1.3 LSTM

The training of the LSTM architecture for predicting the closing value of Allianz SE and SAP SE presented in section 2.5 is compiled by adopting the adam-optimizer and specifying the mean squared error as the loss function. The model is then trained using the training set with a batch size of one and for three epochs. To evaluate the trained model, the predict-function is used to generate predictions of the stock prices based on the test set. Finally, the resulting predictions are denormalized using the inverse_transform method to restore them to the original scale of the data.

The final neural network model comprises of multiple layers, where each layer consists of a specific number of network units and trainable parameters, that are updated with backpropagation. The initial layer is a Long Short-Term Memory network with an output shape of (none, 60, 100), where *none* indicates the batch size, *60* refers to the number of time-steps, and *100* corresponds to the number of network units. This layer involves training of 40,800 parameters. Subsequently, the second layer is also a LSTM network with an output shape of (none, 100), where *none* signifies the batch size and *100* corresponds to the number of network units. This layer involves training of 80,400 parameters. The next layers include a Dense layer with an output shape of (none, 25), where *none* indicates the batch size and *25* corresponds to the number of network units. This layer requires training of 2,525 parameters. Finally, the second Dense layer has an output shape of (none, 1), where *none* indicates the batch size and *1* corresponds to the number of network units. This layer requires training of 26 parameters. The model comprises a total of 123,751 trainable weights.

Regarding the Allianz stock, the model exhibits a current training set loss of 0.0067 for the first epoch, 0.0025 for the second epoch, and 0.0018 for the third epoch, respectively.
The model for the SAP stock displays a training set loss of 0.0079 for the first epoch, 0.0034 for the second epoch and 0.0024 for the third epoch, respectively.

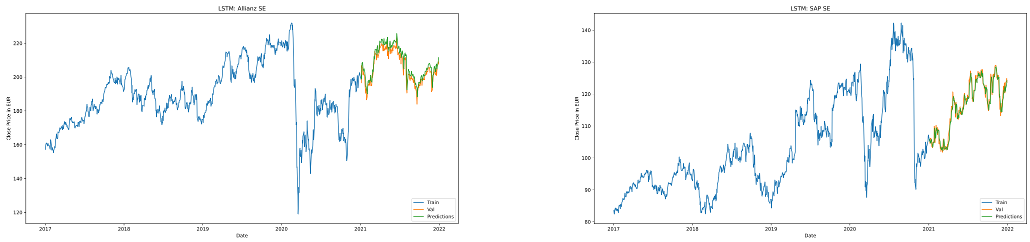

Figure 3.3: This figure shows the results of the LSTM model for Allianz SE and SAP SE.

Figure 3.3 displays the outcomes derived from the LSTM model, and its projected stock prices. The visual representation entails the blue line representing the closing price of the training data, the orange line indicating the actual closing price of the test data and the green line portraying the predicted closing price of the test data. One notable characteristic of this model is its precision in projecting the closing price of the test data, as demonstrated by the green line's resemblance to the orange line's pattern.

3.2 Comparison

In order to compare the results of the prediction of a stock price, the Root Mean Squared Error (RMSE) was utilized. The RMSE can be defined using the following mathematical expression [4]: To simplify, it is assumed that there are n samples of model errors ϵ calculated as $(\epsilon_t, t = 1, 2, ..., n)$. The RMSE is calculated for the data set as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2}. \tag{3.3}$$

The underlying assumption when presenting the RMSE is that the errors are unbiased and follow a normal distribution.
The RMSE calculates the average deviation between the estimated values and the values of the original time series and enables an assessment of the model quality. To

determine which model allows the most reliable forecast, a comparative analysis is performed on the RMSEs of the linear regression model, ARIMA model, and LSTM model. The outcome is illustrated in table 3.1 .

	Allianz SE	**SAP SE**
Model	**RMSE**	**RMSE**
Linear Regression	17.3451	19.1471
ARIMA	40.8636	16.6194
LSTM	2.5428	0.3171

Table 3.1: The table shows the RMSE scores of Allianz SE and SAP SE for each model.

Table 3.1 reveals that the average deviations in the linear regression model are 17.3451 for Allianz SE and 19.1471 for SAP SE. The RMSE in the ARIMA model is notably worse for the Allianz stock, with 40.8636, while it is superior for the SAP stock, with 16.6194. The LSTM model produces the comparatively best outcomes, with an RMSE of 2.5428 for Allianz and 0.3171 for SAP.
Overall, it is evident that the most reliable results are provided by predictions of the Long Short Term Memory model when compared to all other examined methods.
Another aspect that illustrates the advantages and also disadvantages of the respective forecasting method is the application of the methods and their complexity. Linear regression can quickly create forecasts based on the input stock price data. The use of linear regression is simple and can be executed through the established function, without requiring any special prior knowledge of model building or manual preparations by the user for forecasting stock prices. However, compared to all other presented forecasting methods, this method has significant weaknesses in terms of reliability.
In contrast, more reliable forecasts can be achieved using an ARIMA model. However, the use of ARIMA models is more complex than all other investigated forecasting methods. If the analysis of the time series and their forecast is done manually, some basic knowledge of modeling time series with ARIMA(p,d,q) models is necessary. Determining the orders of p, d, and q as well as parameter estimation of ϕ_t and θ_t are complex and require a detailed analysis of the underlying time series. Adjusting certain parameters can optimize the models, but must be done manually, requiring user analysis. With an increasing number of parameters, the complexity of the model also increases, along with the computation time.
In comparison, the LSTM allows for quick data processing and reliable forecasts. However, the downside of LSTM is that it often requires multiple runs of calcula-

tions for accurate forecasts. Nonetheless, the computation time is significantly less than that of an ARIMA model.

4 Conclusion

This chapter provides a concluding summary of the results obtained in this paper. It highlights and presents the most significant insights that have been captured in this project and its limitations.

4.1 Summary and Discussion

The objective of this study was to introduce and compare feasible models for forecasting stock prices in the ESG market. For this purpose, the linear regression model and the ARIMA-model, two traditional prediction techniques were selected as well as the Long Short Term Memory, a deep learning approach. The most effective technique to support investment choices based on the stock's future closing price was identified by computing the average deviations between the predicted and actual price trends using the Root Mean Square Error.
The utilization of various software-based prediction methods has revealed that linear regression fails to adequately represent the highly non-linear fluctuations and trends of many stock prices. The proposed method is solely suitable for stocks that follow a linear trend. Similarly, the ARIMA models that rely on the "auto.arima" function in R are also unable to accurately predict the trajectory of stock prices. This limitation stems from the fact that the Moving Average process is established based on past deviations between the previously forecasted values of the time series and the actual observations of the time series. As the real progression of the time series is yet to transpire, the deviation cannot be ascertained, rendering it impossible to incorporate this information into the forecast.
In contrast, Long Short Term Memory models have demonstrated their effectiveness in capturing long-term dependencies in sequential data and addressing the issue of vanishing gradients that often plagues conventional Recurrent Neural Networks. The inclusion of memory cells and gating mechanisms in LSTMs facilitates the retention of relevant information over multiple time steps while filtering out irrelevant data. Consequently, the network can maintain a long-term memory of the data, which is crucial for accurate prediction of time series data.

The findings of this paper indicate that, among all the methods and forms of forecasting, Long Short Term Memory (LSTM) yields the most dependable predictions for individual stock price forecasting. Consequently, it is deemed to be the optimal choice for investment decision-making in comparison to the other models. The LSTM model exhibits the highest degree of accuracy in predicting potential future upward or downward trends in stock prices.

4.2 Limitations

There are three major limitations in this study that could be addressed in future research. First, the study's selection of only three time series forecasting methods may not be exhaustive and there may be other potentially effective models that were not included in this study. Second, time series forecasting models often require tuning of various parameters to achieve optimal performance. The parameters used in this study may not have been optimally tuned for all models, which could have affected the accuracy of the results. Finally, stock prices are influenced by various external factors such as global economic conditions, geopolitical events, and company-specific news. These factors were not explicitly considered in this study, and their potential impact on the accuracy of the forecasting models cannot be ruled out.

Bibliography

[1] **A. A. Ariyo, A. O. Adewumi and C. K. Ayo**, 2014, Stock Price Prediction Using the ARIMA Model, UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, pp. 106-112.

[2] **Allianz Group**, (n.d.), About us, https://www.allianz.com/en/about-us.html [Retrieved January 8, 2023].

[3] **Brockwell, P. J., Davis, R. A.**, 2002, Introduction to time series and forecasting, Springer, New York, pp. 16-24.

[4] **Chai, T. and Draxler, R. R**, 2014, Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, pp. 1247-1250.

[5] **Chatfield, C.**, 2001, Time-Series Forecasting, Chapman&Hall, CRC, London, pp. 24.

[6] **Chollet, F., & others.**, 2015, Keras, GitHub, https://github.com/fchollet/keras [Retrieved December 22, 2022].

[7] **Chong, E., Han, L., & Li, G.**, 2018, A comparative study of ARIMA and LSTM in symbolic financial time series predictio, IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3883-3888.

[8] **Erlwein-Sayer, C.**, 2021, Statistik 3 [Vorlesung], HTW Berlin, pp. 75-89.

[9] **Fink, A.**, 2023, Forecast of stocks in the ESG market using Recurrent Neural Networks (RNN), GitHub, https://github.com/annalenafink/DeepLearning_WS2022 [Retrieved February 27, 2023].

[10] **Graves, A.**, 2012, Long Short-Term Memory, Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, Vol. 385, Springer, Berlin, Heidelberg.

[11] **Hochreiter S., Schmidhuber J.**, 1997, Long Short-Term Memory, Neural Computation, Vol. 9, No. 8, pp. 5-7, 1735-1780.

[12] **Jadhav, S., & Wadhai**, 2020, A comparative analysis of machine learning techniques for stock price prediction, International Journal of Advanced Science and Technology, 29(8), pp. 3221-3235.

[13] **Litz, S.**, 2019, Vorhersage von Aktienkursen mithilfe rekurrenter Neuronaler Netze, Masterthesis, Hochschule Harz.

[14] **Mushtaq, Rizwan**, 2011, Augmented Dickey Fuller Test, Econometrics: Mathematical Methods & Programming eJournal.

[15] **P. Pai and C. Lin**, 2005, A hybrid ARIMA and support vector machines model in stock price prediction, Omega, Vol. 33, pp. 497-505.

[16] **SAP**, (n.d.), About SAP, https://www.sap.com/about.html [Retrieved January 8, 2023].

[17] **Stock, J., Watson, M**, 2020, Introduction to Econometrics, Fourth Edition, Pearson, Boston, pp. 144-152.

[18] **S&P Global**, February 2022, The Sustainability Yearbook 2022, spglobal.com/esg/csa/yearbook/ [Retrieved January 8, 2023].

[19] **S&P Global**, (n.d.), Corporate Sustainability Assessment, https://www.spglobal.com/esg/csa/ [Retrieved January 8, 2023].

[20] **Van Duuren, E.; Plantinga, A.; Scholtens, B.**, 2015, ESG Integration and the Investment Management Process: Fundamental Investing Reinvented, J. Bus. Ethic, 138, pp. 525-533.

[21] **Verheyden, T., Eccles, R.G., Feiner, A.**, 2016, ESG for All. The Impact of ESG Screening on Return, Risk, and Diversification, J. Appl. Corp. Financ., 28, pp. 47-55.

[22] **Vogel, J.**, 2015, Prognose von Zeitreihen - Eine Einfuehrung fuer Wirtschaftswissenschaftler, Springer Gabler Verlag, Wiesbaden, p. 22.

[23] **Yahoo!**, (n.d.), Yahoo Finance, https://finance.yahoo.com/ [Retrieved January 8, 2023].