MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

Sumy State University

Academic and Research Institute of Business, Economics and Management

Department of Economic Cybernetics

«Admitted to the defense»

Head of Department

_____ Vitaliia KOIBICHUK
(signature) First and LAST NAME)

_____ 2023 p.

QUALIFICATION WORK

to obtain an educational degree bachelor
(bachelor / master)

from the specialty 051 «Economics»,
(code and name)

educational-professional programs Business Analytics
(educational-professional / educational-scientific) (the name of the program)

on the topic: Economic and Mathematical Modeling of Financial Asset Returns Using Python

Winner(s) of the group: AB-91a.an Dun Vadym Romanovich
(group cipher) (full name)

The qualification work contains the results of own research. The use of ideas, results and texts of other authors are linked to the corresponding source

Head: Assistant of the Department of Economic Cybernetics, PhD Serhii MYNENKO _____
(position, academic degree, academic title, Name and SURNAME) (signature)

Consultant: Leading economist of the VKIARFOP UKRSIBBANK BNP's, Eleonora MARTYNENKO _____
(position, academic degree, academic title, Name and SURNAME) (signature)

Sumy – 2023

SUMMARY

of Bachelor's level degree qualification thesis on the theme
«Economic and Mathematical Modeling of Financial Asset Returns Using Python»

Student: Dun Vadym Romanovich

(full name)

The analysis of stock prices and the prediction of their changes are topical issues that constantly attract the attention of researchers in the field of finance and economics. The stock market plays a key role in stimulating economic growth and wealth creation, so accurate forecasts of its state are important for the overall stability and efficiency not only of financial markets, but also of national economies. In today's environment, when investing is becoming accessible to a wide audience, interest in forecasting financial assets is growing. The study and development of modeling methods, stock value forecasting is a practically significant task for all market participants and those who just want to enter the market, allowing to make informed decisions to managing an investment portfolio.

The purpose of the bachelor's qualification work is to develop economic and mathematical model of financial asset returns using Python.

The object of the study is the return on financial assets.

The subject of the research is the economic and mathematical methods and models for studying the return on financial assets.

In accordance with the set tasks, the following was done: deepening of theoretical knowledge in the field of stock markets and indices, review of modern approaches to modeling and forecasting financial assets. Data cleaning/validation, determination of moving statistics, test of stationarity by Augmented Dickey-Fuller method, seasonal decomposition, logarithmic transformations, search for model parameters and implementation of ARIMA model for forecasting S&P 500 index, analysis of results and verification of accuracy.

The study was tested at the International Virtual Conference "Cybersecurity Challenges Facing the Financial Services Industry" – presentation on "Analysis of the Impact of Major Cyber Incidents on the company's stocks" [45].

Keywords: Time series forecasting, stock market, investments, Python, S&P 500 index, stationarity test, ARIMA model, moving statistics

The content of qualification work is presented on 40 pages. The references consist from 48 names, placed on 5 pages. The work contains 26 figures, 2 appendices.

Year of performance of qualification work – 2023.

Year of protection of work – 2023.

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

SUMY STATE UNIVERSITY

Academic and Research Institute of Business, Economics and Management

Department of Economic Cybernetics

APPROVED BY

Head of Department Candidate

of Economics, Associate Professor

_____ Koibichuk V. V.

“__” _____ 2023.

TASKS FOR BACHELOR’S LEVEL DEGREE QUALIFICATION THESIS

(specialty 051 “Economics” (Study Programme “Business Analytics”))
Student of IV course, group’s code AB-91.a.an.

DUN VADYM ROMANOVICH
(student’s full name)

1. The theme of the work is “Economic and Mathematical Modeling of Financial Asset Returns Using Python” approved by the order of the university from “23” May, 2023 year № 0553-VI
2. The term of completed paper submission by the student is “20” June 2023 year.
3. The purpose of the bachelor's qualification work is to develop economic and mathematical model of financial asset returns using Python.
4. The object of the study is the return on financial assets.
5. The subject of the research is the economic and mathematical methods and models for studying the return on financial assets.
6. The qualification paper is based on various scientific sources in the field of financial asset modeling, historical data for the S&P 500 Index taken from the Yahoo Finance website, and documentation on the Python programming language.7. The indicative plan of qualification work, terms of submission of the chapters to the research advisor, and the content of tasks for the performance of the set purpose is as follows:
Chapter 1. Theoretical foundations of stock market analysis and modeling
In Chapter 1 it is necessary to conduct a comprehensive study of the theoretical aspects and contemporary structure of stock exchanges, encompassing

their historical background, key characteristics, functions, and the profound impact they have on the economy. Review methods of modeling the stock market and model selection for further forecasting.

Chapter 2. Analysis and forecasting of the S&P 500 index using the ARIMA model. In Chapter 2 it is necessary to cover all aspects related to the practical part of building an accurate model. Namely, the determination of moving statistics, stationarity tests, the search for model parameters, and the implementation of the chosen model for forecasting future values and verifying accuracy.

8. Supervision on work:

Chapter	Full name and position of theadvisor	Date, signature	
		task issued by	task accepted by
1			
2			

9. Date of issue of the tasks: “03” April 2023 year

Research Advisor _____ Mynenko S.V.
(signature) (full name)

The tasks has been received _____ Dun V.R.
(signature) (full name)

CONTENT

INTRODUCION

In today's environment, where investing has become accessible to the masses, it is hard to find a person who does not have an investment portfolio. When we talk about stock markets, it is important to note that they play a key role in stimulating economic growth and wealth creation. Consequently, there is a growing interest in forecasting financial assets.

One of the main goals of econometric modeling of the money market is the study of time series in finance. For a long time, financial market researchers assumed that financial assets follow a normal distribution and are completely unpredictable. However, the application of new approaches to financial market modeling has shown that real time series of financial data are not only devoid of randomness, but also have a long memory. This means that past events have a strong influence on the future returns of financial assets.

However, the analysis of financial markets is complicated and there is no model or indicator that can predict the price. But, if we remove the factor of force majeure from the equation, it is possible to predict the trends and directions of the markets, and many specialists from all over the world are involved in it. One of the most interesting assets is the S&P 500 index, which is one of the main indicators of the American economy. Successful and accurate forecasting of the index allows analysts and economists to draw conclusions about the trends in the economy and take appropriate actions not only within an individual investment portfolio, but also within countries.

The purpose of the bachelor's qualification work is to develop economic and mathematical model of financial asset returns using Python.

The object of the study is the return on financial assets.

The subject of the research is the economic and mathematical methods and models for studying the return on financial assets.

The purpose of the work determined the following tasks: deepening of theoretical knowledge in the field of stock markets and indices, review of modern

approaches to modeling and forecasting financial assets. Data cleaning/validation, determination of moving statistics, test of stationarity by Augmented Dickey-Fuller method, seasonal decomposition, logarithmic transformations, search for model parameters and implementation of ARIMA model for forecasting S&P 500 index, analysis of results and verification of accuracy.

The study was tested at the International Virtual Conference "Cybersecurity Challenges Facing the Financial Services Industry" – presentation on "Analysis of the Impact of Major Cyber Incidents on the company's stocks" [45].

The Python programming language was chosen to perform the purpose of the bachelor's qualification work.

1. THEORETICAL FOUNDATIONS OF STOCK MARKET, ANALYSIS AND MODELING

1.1 The role of the stock market in the world economy

The idea of trading goods can be traced back through history to the earliest civilizations. For centuries, early businesses cooperated and pooled their resources to conduct international trade transactions. These transactions were conducted by trading groups as well as individuals. In the Middle Ages, merchants gathered in the center of cities to exchange goods from different countries. However, since they represented different nationalities, it became necessary to establish a system of currency exchange to ensure fairness in trade transactions. Thus, over time, trading mechanisms and institutions evolved, contributing to the development of international trade and laying the foundation for modern stock markets.

The stock exchange («purse» derived from the Latin word «bursa») means the organizer of commodities, securities and labor-powered wholesale sales on the basis of supply and demand in the economy, as well as for the sale of financial and trading transactions to sellers and buyers place [1].

A stock exchange is an organized marketplace where securities such as stocks, bonds, funds, and other financial instruments are traded. An exchange is where buyers and sellers come together to make transactions based on supply and demand. The exchange provides transparency and liquidity in the securities market, allowing companies to raise capital and investors to purchase assets. In addition, the stock exchange sets the rate of return (in the form of dividends or interest), the interest rate on a loan. And stock market transactions attract and redistribute equity capital across many industries. Trading on the stock exchange is conducted according to certain rules and procedures, and all transactions are registered and monitored by the relevant regulatory authorities [2].

At the end of the 15th century, the city of Antwerp in present-day Belgium became an excellent center for international trade and is considered the birthplace of

the stock market. At that time, some merchants began to practice a strategy of buying goods at a certain price in the hope of selling them later for a profit. Wealthy merchants made high-interest loans to those who needed financing. These merchants then sold bonds backed by these loans and paid interest to other people who bought these bonds. In 1611, the first modern stock exchange was established in Amsterdam. The Dutch East India Company played a special role in its formation, as it was the first public company and for a long time remained the only company whose shares were traded on the stock exchange. It attracted investors by offering them a share in the profits and successes associated with colonial expansion and trade with the East. This event was fundamental in the history of the development of stock markets and laid the foundation for the further expansion and diversity of financial instruments and companies traded on the stock exchange [3].

The stock market continues to evolve and is influenced by a number of important events. One such event that had a significant impact on the development of the stock market was the Industrial Revolution which took place in the 18th and 19th centuries. The Industrial Revolution was accompanied by rapid advances in industry and the creation of new companies. The shares of these companies were traded on stock exchanges, which led to a significant expansion of the market itself and attracted a large number of investors. The growth of industry and the emergence of new companies created new investment opportunities and stimulated the development of the stock market. However, with this development came the need to establish regulatory measures and protect the interests of investors. In response to these needs, rules and regulations were introduced and specialized organizations responsible for the control and supervision of stock exchanges and companies were established. Such organizations play an important role in ensuring stability and creating confidence in the market by ensuring that trading rules are followed, investor rights are protected, and fraud is prevented [4].

Today, the stock market continues to undergo active technological innovations that are changing the very process of conducting transactions and accessing information. Electronic trading, process automation and the development

of online platforms have created opportunities for investors to trade stocks and other financial instruments with greater efficiency and convenience [5]. Now the exchange market is a complex system where various financial assets, such as stocks, bonds, commodities and derivatives, are traded. It is an electronic platform where buyers and sellers make transactions based on supply and demand. Modern exchange markets are characterized by the following features [6,7]

– Electronic Trading: Most modern exchange markets are conducted through electronic platforms where trading is automated and instantaneous. This allows buyers and sellers to match offers quickly and provides high market liquidity.

– Global accessibility: Exchange markets are becoming increasingly global, allowing investors and traders from different countries to trade. Modern technology allows trades to be executed remotely, giving participants around the world broad access to the market.

– Diversity of financial instruments: Today's exchange markets offer a wide range of financial instruments that allow investors to diversify their portfolios and manage risk. These include equities, government and corporate bonds, commodities, currencies and derivatives.

– Regulation and oversight: Today's exchange markets are subject to strict regulation and supervision by financial regulators. This includes setting rules and regulations for conducting transactions, protecting investors interests, and maintaining integrity and transparency in the marketplace.

Prominent examples in the literature on financial markets and their analysis are the following works: Malliard "Technical Analysis of Financial Markets" describes the basic tools and methods of technical analysis used to predict the behavior of the stock market. Fisher in his book "Random Walk in Economics and Finance" examines random walk and its application to financial economics, including the stock market. Koch, "Modeling Financial Markets," presents various models and methods for modelling financial markets to predict price changes and risks. Kleiner's Mathematical Methods in Finance presents mathematical methods and models used in financial analysis and the stock market. Kaminski and Lopez de

Silva, in their book Analysis and Risk Management in Financial Markets, explore methods for analyzing and managing risk in financial markets in light of current requirements and tools.

One of Henry's seminal works, "Stock Market Liberalization, Economic Reform and Stock Prices in Emerging Markets", emphasizes that stock markets play a crucial role in facilitating the relationship between savers and producers in society. Savers, who have accumulated a surplus of funds, seek to invest their savings in profitable and ambitious projects. On the other hand, producers, representing the productive sectors of the economy, need financial resources to fuel their activities and promote economic growth. Stock markets act as intermediaries, allowing the transfer of funds from savers to producers. This process allows productive sectors to access the necessary capital for expansion and development. The productivity and functions of the stock market play an important role in redirecting funds from those who have excess resources to those who need them, thereby facilitating economic activity and development [8].

Figure 1.1 shows that creditor depositors can finance their expenditures by borrowing money from borrower-lenders in two main ways. The first is an indirect transfer through financial intermediaries such as banks and commercial institutions. The second method is a direct transfer, where borrowers receive funds directly from lenders through the sale of securities [9].

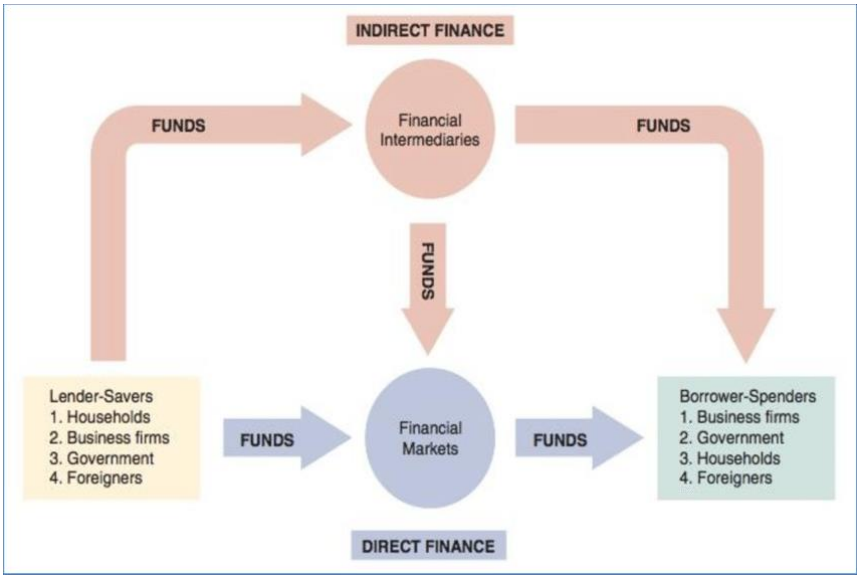


Figure 1.1 – Flows of Funds Through the Financial System [10]

Financial intermediaries play an important role in reducing risk to the economy because they own most of the investments. Lower interest rates encourage investment, and intermediaries are more efficient at transferring funds from lenders to borrowers. Financial intermediaries are better able to own assets and investment instruments, allowing them to diversify their portfolios and manage investment risk. Thanks to the expertise of their financial professionals, intermediaries can make significant profits from managing the purchase and sale of these investment instruments [9].

The stock market performs a number of economic functions that can be defined as follows [11, 12]:

– To provide access to financial resources: The stock market allows lenders and investors to invest their money in various financial assets, such as stocks and bonds. This increases the amount of financial resources available and encourages investment activity.

– Provides financial information: The stock market plays an important role in providing information about the financial condition of companies and projects related to various financial assets. This reduces the cost of accessing such information, making it more accessible and allowing for more informed investment decisions.

– Provides liquidity: The stock market provides liquidity to holders of financial assets. Owners of stocks and bonds can sell their investments in the market when they need funds or want to reallocate their investments. This creates the ability to quickly convert assets into cash.

– Development of financing methods: The stock market serves as a platform for the development and evolution of various methods of financing projects. It provides an opportunity for companies and organizations to raise capital to finance their projects through the issuance of stocks and bonds of various types and maturities.

From the above, we can conclude that the stock market plays a critical role in the economy by providing a transfer of funds from owners who do not have

investment opportunities to those who can use these funds effectively. This helps to increase production, economic efficiency and welfare of the society. Financial intermediaries, such as commercial banks, investment banks, insurance companies and pension funds, play an important role in transferring funds from lenders to borrowers. The common factor that unites these intermediaries is the ability to access funds by creating debt obligations and borrowing funds from the public for subsequent investment in instruments such as stocks and bonds. This ensures efficient use of funds and stimulates economic growth and development.

1.2. The significance of stock market indices: exploring the S&P 500

Stock market indices are an integral part of the analysis and understanding of the financial system. They provide macroeconomists and financial economists with important tools for studying and forecasting market behavior and economic development. Without reliable and consistent indices, it becomes more difficult to identify long-term patterns, evaluate the performance of financial companies, and make comparisons between different markets.

Financial indices are an important source of information for traders and investors. They provide a quick summary of the state of stock markets, evaluate their performance and make informed investment decisions. Indices allow traders and investors to easily track changes in the market, identify trends and predict possible price movements. However, despite the importance of indices, not enough attention is always paid to their methodology and proper use. Currently, information on index methodology is rarely included in economics or business curricula and remains available only to a limited number of specialists. This can lead to misinterpretation of data and potentially misleading index-based decisions[13].

Historically, there have been many different indexes, offering different calculation methodologies and focusing on different aspects of the market. Some have become widely known and used, such as the Dow Jones Industrial Average (DJIA), the S&P 500, and the NASDAQ Composite [14]. However, each index has its own characteristics and purpose, and the choice of a particular index depends on your analysis or investment strategy. Given these factors, it is clear that a thorough understanding of indices and their methodology is key to the proper use and interpretation of data. The correct use of indices allows you to draw more accurate and meaningful conclusions about the state of the stock markets and to make more informed decisions.

According to the calculation method, the indices are divided into groups, the most common of which are as follows [15]:

– Price indices: Calculated by averaging the prices of index components with their weights. Examples include the Dow Jones Industrial Average (DJIA) and the Nikkei 225.

– Market-cap-weighted indices: Calculated by taking into account the market value of the index components. The larger the market capitalization of a company, the greater its weight in the index. Examples include the S&P 500 and the NASDAQ Composite.

– Balanced indices: All index constituents are equally weighted, regardless of size or market capitalization.

– Factor indices: These are calculated based on specific factors such as value factor (price/earnings), capitalization factor (small, mid, or large capitalization), asset value factor, and others. Examples of factor indices include the Fama-French three-factor model and the MSCI Minimum Volatility Index.

The importance of indices in the stock market includes [16-18]:

– The first function of indices in the stock market is their ability to reflect the overall performance of the market. By combining several stocks or other financial instruments into a single index, indices allow investors to measure overall market movements and changes. Indices such as the S&P 500 or the Dow Jones Industrial Average provide valuable information about the market and its long-term trend.

– The second function of indices in the stock market is to be used as a benchmark to compare the performance of investment portfolios, mutual funds or individual stocks. Indices allow investors to gauge how well their investments are performing relative to the overall market. Comparison to an index can help determine the effectiveness of an investment strategy and whether it needs to be adjusted.

– The third function of indices in the stock market has to do with their ability to serve as a market trend indicator. Changes in the index can indicate current market trends. For example, an increase in an index can be a signal of a strong economy and investor confidence, while a decrease in an index can indicate economic problems

or uncertainty in the market. Such signals can help investors make decisions to buy or sell assets based on current market trends.

– The fourth function of indices in the stock market is their ability to guide investment strategies. Indices provide information about industries, geographic regions, or other market segments. This information can be used by investors to develop their investment strategies and make asset allocation decisions. For example, an investor may decide to focus on an industry sector that is performing well relative to the overall market.

Let's take a closer look at the S&P 500, which is considered one of the most important stock market barometers and an indicator of the health of the US economy. Its movements and changes affect other markets and indices around the world. The index is designed to measure the performance of eligible stocks listed on the NYSE and Nasdaq. It is weighted by float-adjusted market capitalization and incorporates liquidity and tradability criteria in the constituent selection process [19]. The S&P 500 Index measures the value of the stocks of the 500 largest companies by market capitalization listed on the New York Stock Exchange or Nasdaq. The intent of Standard & Poor's is to have a price that provides a quick look at the stock market and the economy [20].

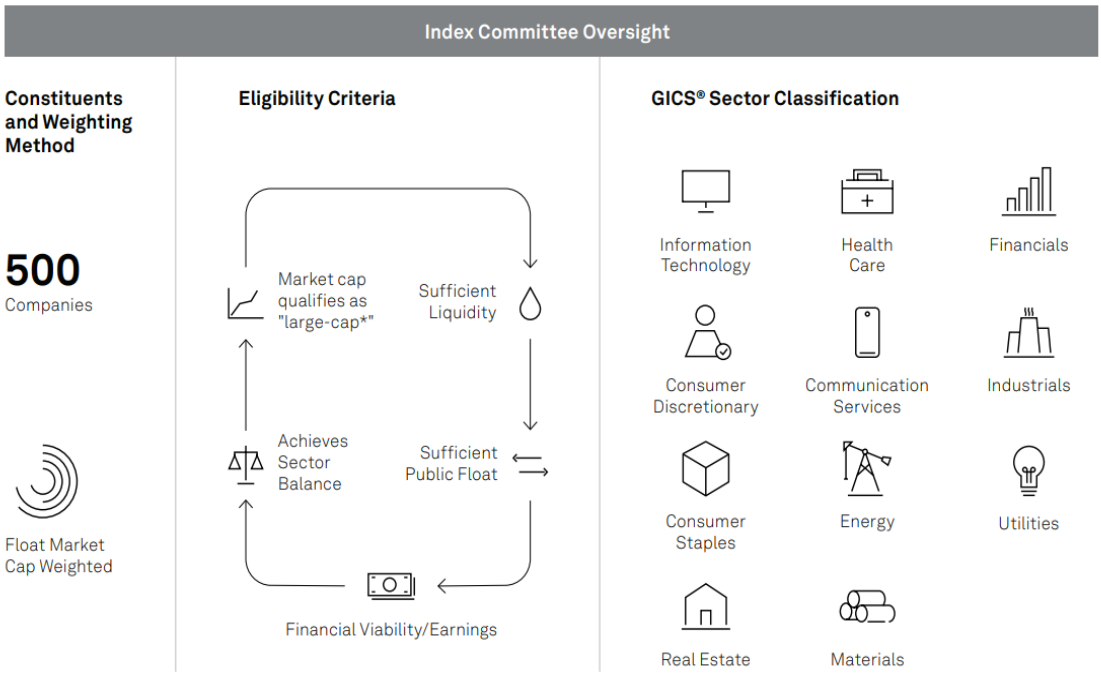


Figure 1.2 – Index Committee Oversight [21]

In order to be selected by the Index Committee and to be included in the S&P 500 Index, a company must meet certain criteria [22]:

– Geographic location: The company must be incorporated and headquartered in the United States.

– Market Capitalization: The company must have a market capitalization of at least $8.2 billion. Market capitalization is calculated by multiplying the company's current share price by the total number of shares outstanding.

– Stock Liquidity: A company's stock must be highly liquid, meaning that it is actively traded on the stock exchange. This makes it easy to buy and sell a company's stock in the market.

– Public availability: At least 50 percent of a company's outstanding shares must be publicly traded. This makes the company's stock widely available in the marketplace and allows investors to include it in their portfolios.

– Financial performance: The company must have positive earnings in the most recent quarter and positive earnings in the previous four quarters. This indicates that the company is financially stable and successful.

It is important to note that while the S&P 500 Index was originally intended to be an index of 500 companies, it actually contains 505 stocks. This is because some companies, such as Google (now Alphabet Inc.), Facebook, and Berkshire Hathaway, have multiple share classes that are still considered separate components of the index [20].

The S&P 500 Index is an important part of the stock market and its connection to the U.S. economy is obvious. This index is considered by most analysts and investors as an indicator of the overall health of the stock market. It includes the 500 largest publicly traded U.S. companies representing various industries and sectors of the economy. Therefore, changes in the index effect not only the performance of individual companies, but also the collective performance of the economy as a whole. Also an important tool for passive investors seeking access to the states economy through index funds. Index funds, such as ETFs or index mutual funds,

track the performance of the index and allow investors to diversify their portfolios by investing in all the companies in the index [23].

The relationship between the S&P 500 Index and the US economy is manifested in several ways. First, the index includes the largest and most representative companies in various sectors of the economy. Therefore, its performance and movement reflect conditions and trends in a wide range of sectors and industries. The index is an important indicator of investor confidence and expectations about economic conditions in the United States. Rising prices in the index indicate optimistic investor sentiment and belief in continued economic growth. This can stimulate investment, business growth, and demand for labor. As mentioned above, the index is also used as a benchmark for evaluating the performance of investment portfolios and funds. Many active and passive investors use the index as a benchmark to compare and evaluate the performance of their investments. If an investment portfolio outperforms the index, it may indicate successful asset management [24]. The index considered by many analysts and economists to be one of the most important indicators of the health and prospects of the U.S. economy. Changes in the index can serve as an indicator of future economic trends, growth or decline. Analysts study the relationship between the index's movements and macroeconomic factors such as GDP, inflation, and unemployment to predict possible economic outcomes [25].

The S&P 500 Index plays an important role in reflecting and measuring the performance of stocks and the U.S. economy as a whole. Its movements and changes reflect important aspects of economic activity and investor sentiment. This makes it an indispensable tool for analyzing and forecasting economic developments and the market.

1.3 Overview of existing approaches and model selection

The growing importance of stock price forecasting has attracted considerable attention from industry experts and investors. Analyzing stock market trends is challenging due to the inherently noisy environment and significant volatility associated with market trends. The complexity of stock prices involves several factors, including quarterly earnings reports, market news, and changing investor behavior. Traders rely on a variety of technical indicators derived from daily stock market data. Despite the use of these indicators to analyze stock returns, accurately predicting daily and weekly market trends remains a challenge. Accurately predicting stock trends is a fascinating and challenging task in an ever-changing industrial world. Several aspects that influence stock market behavior are both non-economic and economic factors that are taken into consideration. Thus, stock market forecasting is considered a major challenge for increasing production. Traditional methods show that stock market returns are predicted based on past stock returns, other financial variables, and macroeconomics. Predicting stock market returns has led investors to investigate the reasons for predictability. Forecasting stock market trends is a complex process because it is influenced by many aspects, including traders' expectations, financial circumstances, administrative events, and certain factors related to market trends. Moreover, the stock price list is usually dynamic, complex, noisy, non-parametric and non-linear in nature. Financial time series forecasting becomes problematic due to certain complex characteristics such as volatility, irregularity, noise and changing trends [26].

Statistical methods include a large number of methods, such as methods of valuation theory, factor analysis, regression and correlation analysis, etc. With the help of these methods, investors can conduct a comprehensive statistical research of the financial market, make forecasts of market processes and, on the basis of these forecasts, make more reasonable investment decisions. However, working with such systems for forecasting short-term price movements, rapidly changing intraday information is associated with some difficulties both in the selection of an analysis

method and in the interpretation of results. This seems to be a significant drawback, because the speed of forecasting intraday trading is very important [27].

Below we are available to see the brief review of the main approaches:

One of the simplest and most effective methods is the Autoregressive Integrated Moving Average (ARIMA): This model is one of the most common and simplest time series forecasting models. It is based on the assumption that the future values of a series depend on its past values and forecast errors. It is based on a combination of three main components: autoregression (AR), integration (I) and moving average (MA), as follows:

– Autoregression (AR): The model assumes that the future values of a time series depend on its past values. Autoregression uses the lags (previous values) of the series to predict its future values. Autoregression (AR order) determines the number of past values used in the model.

– Integration (I): Integration is used to ensure that the time series is stationary. If the original series is non-stationary (has trends or seasonality), it can be transformed into a stationary series by using the differences between successive observations. The integration order (I-order) determines the number of differences applied to the series.

– Moving Average (MA): The moving average assumes that the current value of the series depends on random forecast error at previous times. The model uses smoothing of the forecast errors to account for their effect on future values of the series. The MA order determines the number of past errors [28].

The next model that can be distinguished - GARCH (General Autoregressive Conditional Heteroscedastic) - is a model used to model and predict the volatility of time series, such as the prices of financial instruments, including the SP500 index. The GARCH model is based on the assumption that the variance of a series varies over time and depends on the previous values of the series. The model is widely used in financial econometrics and time series analysis to model and predict the volatility of financial data. It has advantages in modeling variability and accounting for variance structure in time series. However, like any model, GARCH has its

limitations and requires proper parameter selection and estimation to achieve good results in predicting volatility [29].

Random Forest is a machine learning algorithm that combines multiple decision trees to perform classification and regression tasks. It uses randomness to select features and data samples, and combines tree predictions to improve model accuracy and stability. Random forest has high performance, the ability to estimate the importance of features, robustness to overlearning, and a wide range of applications in a variety of domains [30].

LSTM (Long Short-Term Memory) is a tool for stock market forecasting and modeling. Unlike traditional models, it is able to handle long-term dependencies in time series and capture complex time patterns. It has a built-in ability to remember and forget information over time, allowing it to account for long-term trends and seasonal fluctuations in the market. LSTM also has the ability to model non-linear dependencies and adapt to changing market conditions. It can use different types of data, including stock prices, trading volumes, macroeconomic indicators, and other factors to make more accurate predictions. Numerous studies and publications demonstrate the successful application of this model in stock market forecasting and describe various methods and approaches to its use [31].

VAR (Vector Autoregression) is a model that is widely used for stock market forecasting and modeling. It allows you to analyze the relationships between multiple time series, taking into account the impact of one series on others. A VAR model is a system of simultaneous equations in which each variable depends on its past values and the past values of other variables. This allows for complex interactions and dependencies between various factors such as stock prices, trading volumes, market indicators, and economic performance. The VAR model has the ability to capture dynamics and long-term trends in time series and to predict future values based on past data. Its advantages include the ability to analyze historical relationships, estimate impulse response, and perform scenario analysis. There are many papers in the financial econometrics literature and research studies that apply

the VAR model to stock market forecasting and modeling, and describe methods for estimating and interpreting the results [32].

XGBoost (Extreme Gradient Boosting) - Combines weak models, such as decision trees, to improve predictions. Works with different types of features, automatically selects important features, and is resistant to overfitting. It has several important advantages. First, it provides high speed and efficiency due to its optimized implementation of gradient binning, making it ideal for dealing with large amounts of financial data. Second, can handle both numeric and categorical attributes, allowing it to account for a variety of factors in stock market analysis, providing models with flexibility and accuracy. Third, the algorithm automatically determines the importance of the attributes and selects the most important ones, improving the quality of forecasts and simplifying the model by removing unnecessary attributes. XGBoost has built-in regularization mechanisms that prevent model overlearning and provide more reliable stock market forecasts [33].

Exponential smoothing is a time series forecasting method that uses a weighted average of past observations with decreasing weight as you move away in time. The basic idea of exponential smoothing is to give more weight to more recent observations and less weight as you move away from the current moment. This allows us to model the impact of newer data on predictions while taking into account the decreasing importance of older data [34].

Our choice of the ARIMA model for the following parts of the paper is based on the following considerations:

– Simplicity and interpretability: it is relatively easy to use and understand. It has a set of parameters, such as autoregression orders (p), difference (d), and moving average (q), that can be chosen based on data analysis and statistical metrics. This makes the model accessible to a wide range of users.

– Flexible model specification: Models are very flexible and can be customized to simulate different types of time series. This versatility is useful when working with multiple time series, as one type can be applied to different data sets.

– Time Dependency Accounting: The model accounts for time dependencies in the data, taking into account previous values in the series. This allows for trends, cyclicality, and seasonality in stock market time series. The model can capture long and short term dependencies, making it effective for forecasting financial time series.

– Suitable for all datasets: Can be trained on relatively small datasets due to fewer parameter requirements compared to neural networks or deep learning models. This makes them suitable when working with limited data availability.

– Robust performance: typically provide robust performance comparable to other time series statistical methods. While they may not always be the most efficient models, they provide consistent and reliable results, making them a good choice when time is limited for extensive experimentation.

– Prevalence and Availability: This is one of the most common time series forecasting models. It is well studied in the literature and has extensive support in statistical packages and software tools. This makes the model accessible and usable in practical stock market forecasting tasks.

– Proven efficiency: The model has proven efficiency in forecasting time series, including financial data. Numerous studies and practical applications show that the model can be an effective tool for stock market forecasting.

These advantages contribute to the attractiveness and usefulness of ARIMA models in forecasting and analyzing time series data, including our selected stock market index, the S&P 500.

1.4 Setting research objectives

The following objectives were set for the future construction of a mathematical model to predict the dynamics of the S&P 500 Index

– EDA and Data Cleaning/Validation: Perform exploratory data analysis to understand the characteristics of the time series.

– Determine Moving Statistics: Compute moving statistics, such as moving averages or moving standard deviations, to identify trends and seasonality.

– Test for stationarity: Apply a stationarity test, such as the Augmented Dickey-Fuller test, to ensure that the time series is stationary. If the time series is not stationary, we will perform additional manipulations to achieve stationarity.

– Apply seasonal decomposition: Apply the seasonal decomposition method to decompose the time series into its components: trend, seasonality, and residuals. This will allow us to better understand the contribution of each component to the overall index dynamics.

– Applying the logarithmic transformation: Apply the logarithmic transformation to the time series to smooth the extreme values and reduce their impact on the model.

– Finding the optimal model parameters: Search for the optimal parameters for the ARIMA model - this will allow us to build the most accurate and appropriate time series model.

– Implement an ARIMA model to predict the S&P 500 index: Divide the data into training and test data and implement the ARIMA model using the optimal parameters to predict the future price.

– Analyzing the Results and Checking the Accuracy of the Model: Analyze the results of the analysis and compare the predicted values with the actual data to evaluate the accuracy and reliability of our model. This will allow us to draw conclusions about the applicability of the model for predicting the future dynamics of the index.

2. ANALYSIS AND FORECASTING THE PRICE OF THE S&P 500 INDEX USING THE ARIMA MODEL

2.1 Dataset overview and preprocessing

In this section, we will be look at a dataset taken from the Yahoo Finance website [35], that contains historical data for the S&P 500 Index. This daily dataset covers the period from January 3, 1990 to May 31, 2023 and provides information about the opening price, the maximum and minimum price, the closing price and the adjusted closing price.

This case study will provide a basic understanding of the data structure and provide reasonably accurate future price predictions, which can be a great tool to minimize the risk of losing money in the market.

We will do our work using the Python programming language version 3.10.12. You should start by importing all the necessary components (Fig. 2.1). These libraries and modules play a crucial role in data analysis, visualization, time series modeling, and evaluation of model performance on the dataset. They provide a wide range of functions and tools that simplify various aspects of data analysis and forecasting in the context of financial markets. They are also divided into sections, and you can see a description of them in the comments

```
!pip install pmdarima

import warnings
warnings.filterwarnings('ignore')

# Data manipulation and analysis libraries
import numpy as np
import pandas as pd

# Data visualization libraries
import matplotlib.pyplot as plt
import seaborn as sns

# Time series analysis libraries
from statsmodels.tsa.stattools import adfuller
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.arima.model import ARIMA
from pmdarima.arima import auto_arima

# Evaluation metrics libraries
from sklearn.metrics import mean_squared_error, mean_absolute_error
```

Figure 2.1 – Import Libraries

All necessary libraries have been downloaded, now you need to import and read the data set (Fig. 2.2).

```
1 df = pd.read_csv("S&P 500 Historical Data.csv")
2 df
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2000-01-03	1469.250000	1478.000000	1438.359985	1455.219971	1455.219971	931800000
1	2000-01-04	1455.219971	1455.219971	1397.430054	1399.420044	1399.420044	1009000000
2	2000-01-05	1399.420044	1413.270020	1377.680054	1402.109985	1402.109985	1085500000
3	2000-01-06	1402.109985	1411.900024	1392.099976	1403.449951	1403.449951	1092300000
4	2000-01-07	1403.449951	1441.469971	1400.729980	1441.469971	1441.469971	1225200000
...	...	...	...	...	...	...	...
5885	2023-05-24	4132.959961	4132.959961	4103.979980	4115.240234	4115.240234	3739160000
5886	2023-05-25	4155.709961	4165.740234	4129.729980	4151.279785	4151.279785	4147760000
5887	2023-05-26	4156.160156	4212.870117	4156.160156	4205.450195	4205.450195	3715460000
5888	2023-05-30	4226.709961	4231.100098	4192.180176	4205.520020	4205.520020	4228510000
5889	2023-05-31	4190.740234	4195.439941	4166.149902	4179.830078	4179.830078	5980670000

5890 rows × 7 columns

Figure 2.2 – Import S&P 500 Historical Data

And first, we need to check all the information about the data, check the data types and also check for missing data in the set (Fig. 2.3).

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5890 entries, 0 to 5889
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       5890 non-null   object
 1   Open       5890 non-null   float64
 2   High       5890 non-null   float64
 3   Low        5890 non-null   float64
 4   Close      5890 non-null   float64
 5   Adj Close  5890 non-null   float64
 6   Volume     5890 non-null   int64
dtypes: float64(5), int64(1), object(1)
memory usage: 322.2+ KB
```

```
1 # Check for null values
2 null_values = df.isnull().sum()
3 print(null_values)
```

```
Date         0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

Figure 2.3 – Dataset information and check for null values

In this paper, we only need the closing price for the entire study. To focus specifically on the closing price, we can extract the 'Close' column from the dataset and store it in a new variable called df_close(Fig. 2.4). This will allow us to perform further analysis and calculations on the closing prices only.

```
df_close = df['Close']
```

Figure 2.4 – Extract the 'Close' column from the dataset

For further successful data visualization, let's set the 'Date' column as an index, it gives us the opportunity to replace the serial number of the rows with a specific date of observation (Fig. 2.5).

```
df['Date'] = pd.to_datetime(df['Date'])
df = df.set_index(df['Date']).sort_index() # setting date feature as our index
print(df.shape)
df.sample(5)
```

```
(5890, 7)
```

Date	Date	Open	High	Low	Close	Adj Close	Volume
2022-05-19	2022-05-19	3899.000000	3945.959961	3876.580078	3900.790039	3900.790039	5113550000
2016-05-31	2016-05-31	2100.129883	2103.479980	2088.659912	2096.949951	2096.949951	4514410000
2014-06-03	2014-06-03	1923.069946	1925.069946	1918.790039	1924.239990	1924.239990	2867180000
2001-01-23	2001-01-23	1342.900024	1362.900024	1339.630005	1360.400024	1360.400024	1232600000
2000-01-03	2000-01-03	1469.250000	1478.000000	1438.359985	1455.219971	1455.219971	931800000

Figure 2.5 – Setting the 'Date' column as an index

Now we can look at the full chart of our price - it will create a line chart of closing prices, where the x-axis represents the dates and the y-axis represents the corresponding closing values (Fig. 2.6).

```
df_close.plot(figsize=(18,8))
plt.title('S&P 500 Historical Data')
plt.grid(True)
```

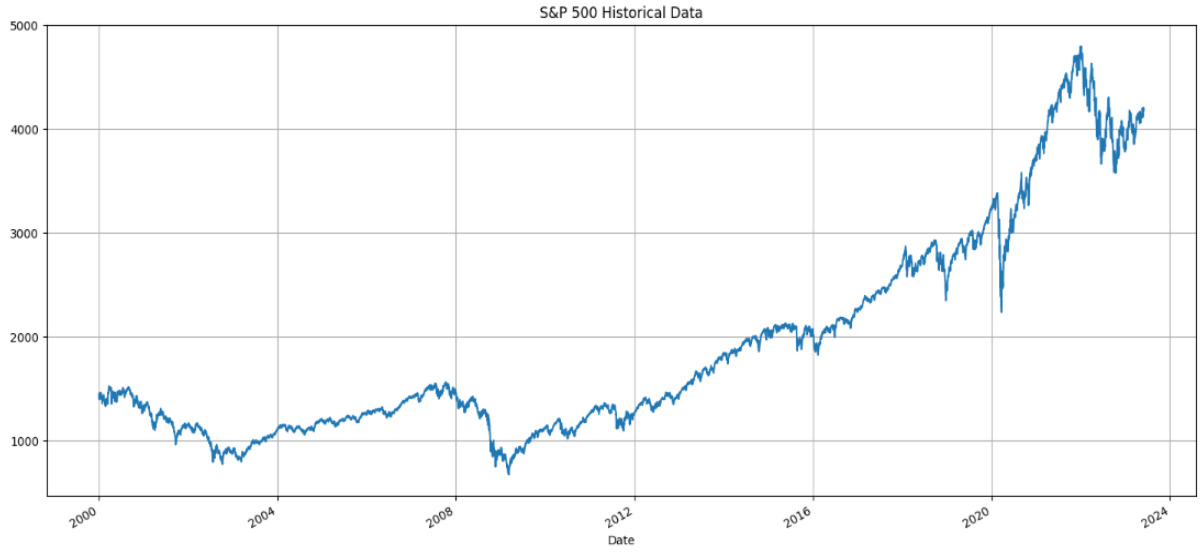


Figure 2.6 – Plot S&P 500 Historical Data

Next, we'll create a Kernel Density Estimation (KDE) graph that estimates the probability density function of the data, providing insight into the distribution and shape of the data. This KDE chart allows you to visualize the distribution of closing prices. The resulting curve provides an estimate of the probability density function, with higher peaks indicating areas of higher density and lower troughs indicating areas of lower density. The shading below the curve provides a visual representation of the estimated probability density function (Fig. 2.7).

```
import seaborn as sns
sns.kdeplot(df_close, shade=True)
plt.show()
```

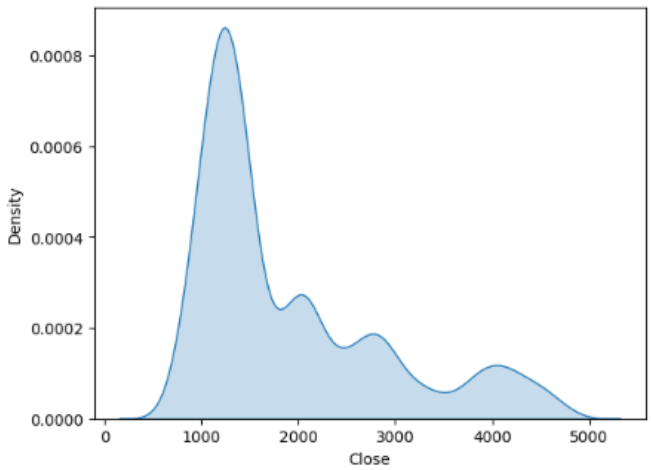


Figure 2.7 – Kernel Density Estimation graph

Kernel density estimation (KDE) is a method used to estimate the probability density function of a random variable from a given data set. The graph can provide insight into the central trend of closing prices. The location of the highest peak on the KDE chart corresponds to the mode of the distribution that represents the most frequent closing price. This can be useful for traders in identifying potential support or resistance levels - from this point of view we can identify four support levels, the main one being around 1200. It is also important to note that the KDE chart can help identify potential outliers in the closing prices. Outliers are data points that deviate significantly from the overall distribution pattern. These outliers may represent important events or anomalies that have affected the closing prices - of which we do not observe any [36].

We can move on to more basic things like the Dickey-Fuller test for stationarity. Data are stationary if they have no trend or seasonal effects. And if the data is non-stationary, we need to convert it to stationary before we can fit it to an ARIMA model. But before we do that, construct a rolling mean and a rolling standard deviation, which are statistical metrics that help to estimate the mean and the dispersion of the values in the time series in a given window(Fig. 2.8).

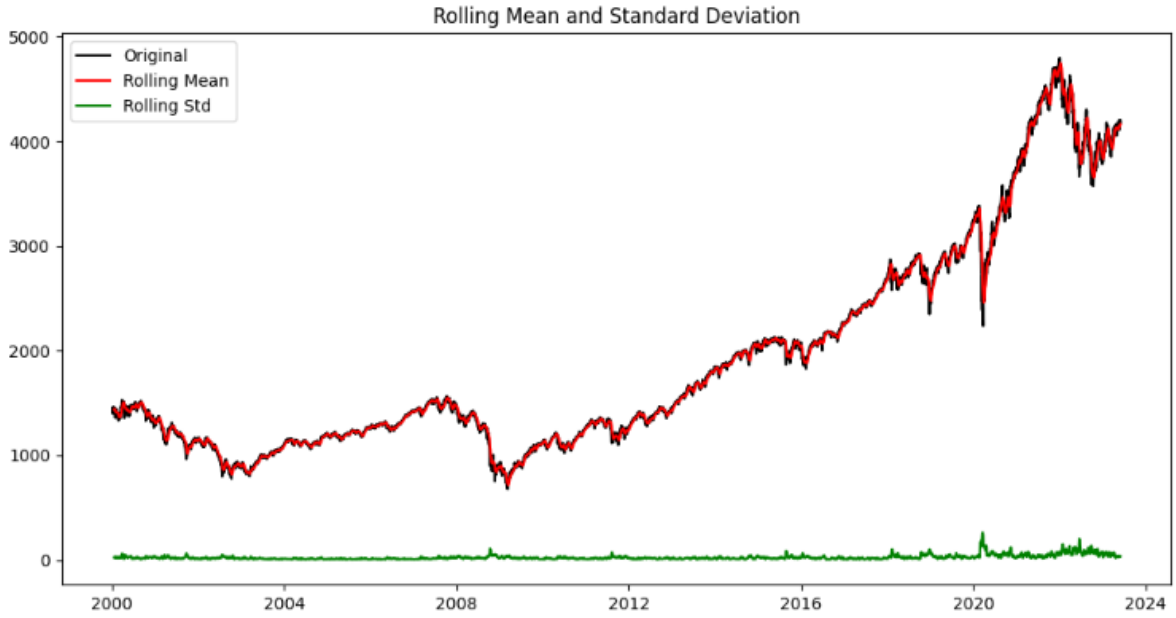


Figure 2.8 – Rolling Mean and Standard Deviation

The moving statistic chart created in the code (Fig. 2.8) helps us to visually assess the stationarity of the time series before applying the Dickey-Fuller test. Drawing these curves allows us to visually assess whether the data has a trend (change in mean value over time) or the presence of a changing variance (dispersion of values). If the graph shows clear trends or changes in the variance of the data, this may indicate non-stationarity of the time series.

This function def test_stationarity (Fig. 2.9) is useful for analyzing time series and their stationarity. Stationary time series have a constant mean and variance, which makes forecasting and modeling easier. If a time series is not stationary, additional transformations or modeling may be required to achieve stationarity and more accurate analysis.

```
def test_stationarity(timeseries):
    # Definition of rolling statistics
    rolmean = timeseries.rolling(window=12).mean()
    rolstd = timeseries.rolling(window=12).std()

    # Create a graph of moving statistics
    plt.figure(figsize=(12, 6))
    plt.plot(timeseries, color='black', label='Original')
    plt.plot(rolmean, color='red', label='Rolling Mean')
    plt.plot(rolstd, color='green', label='Rolling Std')
    plt.legend(loc='best')
    plt.title('Rolling Mean and Standard Deviation')
    plt.show()

    # Test for stationarity with the Dickey-Fuller test
    print("Results of Dickey-Fuller test:")
    adft = adfuller(timeseries, autolag='AIC')
    output = pd.Series(adft[0:4], index=['Test Statistics', 'p-value', 'No. of lags used', 'Number of observations used'])

    for key, value in adft[4].items():
        output['Critical value (%s)' % key] = value

    print(output)

test_stationarity(df_close)
```

Figure 2.9 – Def test_stationarity with definition of rolling statistics

And also, in addition to the graph with the moving statistics, it gives us the results of the test - Augmented Dicky Fuller test - the unit root test Let us first see the formula for the Dickey Fuller test which is the origin of the Augmented Dickey Fuller test, and that is (2.1)

$$Y_t = c = \beta_t + aY_{t-1} + \phi\Delta Y_{t-1} + e_t \qquad (2.1)$$

where,

Y_t = value in the time series at time t or lag of 1 time series.

$\phi\Delta Y_{t-1}$ = first difference of the series at time (t-1)

The formula for Augmented Dickey Fuller test, and it goes as follows (2.2):

$$\mathrm{Yt} = \mathrm{c} + \beta\mathrm{t} + aY_{t-1} + \phi\Delta Y_{t-1} + \phi_2\Delta Y_{t-2} \ldots + \phi_p\Delta Y_{t-p} \qquad (2.2)$$

The formula for ADF is the same equation as the DF with the only difference being the addition of differencing terms representing a larger time series. Fundamentally, it has a similar null hypothesis as the unit root test. That is, the coefficient of Y(t-1) is 1, implying the presence of a unit root. If not rejected, the time series is taken to be non-stationary. If null hypothesis is rejected, then Test statistic < Critical Value and p-value < 0.05, the time series is stationary [37-40]. The result of the function (Fig. 2.9) is available to see in (Fig. 2.10).

```
Results of Dickey-Fuller test:
Test Statistics                  1.015948
p-value                          0.994431
No. of lags used                32.000000
Number of observations used    5857.000000
Critical value (1%)              -3.431467
Critical value (5%)              -2.862034
Critical value (10%)             -2.567033
dtype: float64
```

Figure 2.10 – Results of Augmented Dicky Fuller test

Analysis of test results (Fig. 2.10):

– Test Statistics has a positive value of 1.015948. Compared with the critical values, this indicates that the test statistic is far from zero and is not negative enough to confirm the stationarity of the series.

– The p-value is 0.994431, which is a high value close to 1. This means that there is a high probability of obtaining such or more extreme results even if the null hypothesis of non-stationarity of the series is true. Therefore, we do not have sufficient evidence to reject the null hypothesis.

– Critical values) are shown as -3.431467 (1%), -2.862034 (5%), -2.567033 (10%). They are threshold values compared to the test statistic. If the test statistic is less than the critical value, the null hypothesis of non-stationarity is rejected. In this case, the value of the test statistic is not low enough compared to the critical values, which confirms the lack of stationarity in the series.

In summary, based on the results obtained, we can conclude that the time series is not stationary, since we have not rejected the null hypothesis of non-stationarity. This may indicate the presence of a trend, seasonal fluctuations or other systematic changes in the data.

The next logical step is to separate the seasonality from the trend before analyzing the time series. Such an approach will lead to stagnation of the resulting series. (Fig. 2.11).

```
# Apply seasonal decomposition
result = seasonal_decompose(df_close, model='multiplicative', period=365)

# Visualization of decomposition results
with plt.rc_context({'figure.figsize': (18, 8)}):
    fig = result.plot()

# Differentiation to achieve stationarity
df_close_diff = df_close.diff().dropna()
```

Figure 2.11 – Seasonal decomposition

The result of the code (Fig. 2.11) is available to see in (Fig. 2.12).

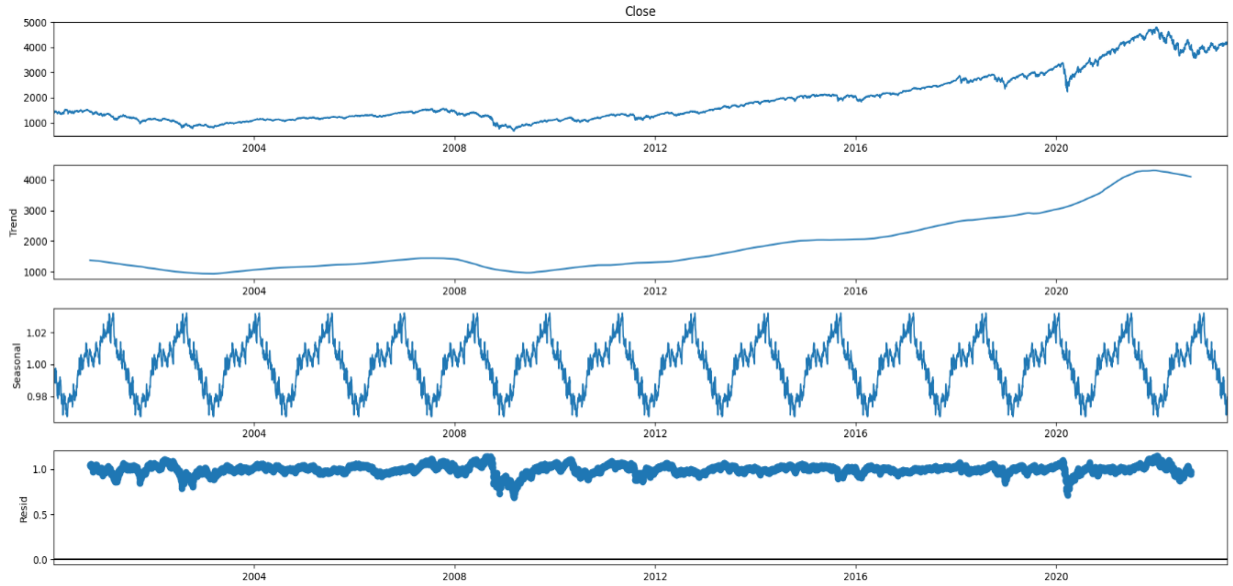


Figure 2.12 – Results of seasonal decomposition

Also consider the second version of stationarity - to reduce the magnitude of the values and the increasing trend in the series, we first take a log of the series. Then, after obtaining the logarithm of the series, we compute the rolling average of the series. A rolling average is calculated by taking data from the previous 12 months and calculating an average consumption value at each subsequent point in the series. The following code is used to smooth the time series and analyze its variability. The logarithmic transformation helps to reduce non-stationarity and smooth fluctuations in the data, and calculating the moving average and standard deviation allows you to assess the overall trend and variability of the series (Fig. 2.13).

```
# Set the size of the chart shape
plt.figure(figsize=(10, 6))

# Apply a logarithmic transformation to the time series
df_log = np.log(df_close)

# Calculate moving average and standard deviation
window = 12
moving_avg = df_log.rolling(window).mean()
std_dev = df_log.rolling(window).std()

# Plot moving average and standard deviation charts
plt.plot(std_dev, color="black", label="Standard Deviation")
plt.plot(moving_avg, color="red", label="Mean")

# Chart legend and title settings
plt.legend(loc="best")
plt.title("Moving Average and Standard Deviation")

plt.show()
```

Figure 2.13 – Moving average and standard deviation

The plot of the code (Fig. 2.13) is available to see in (Fig. 2.14).

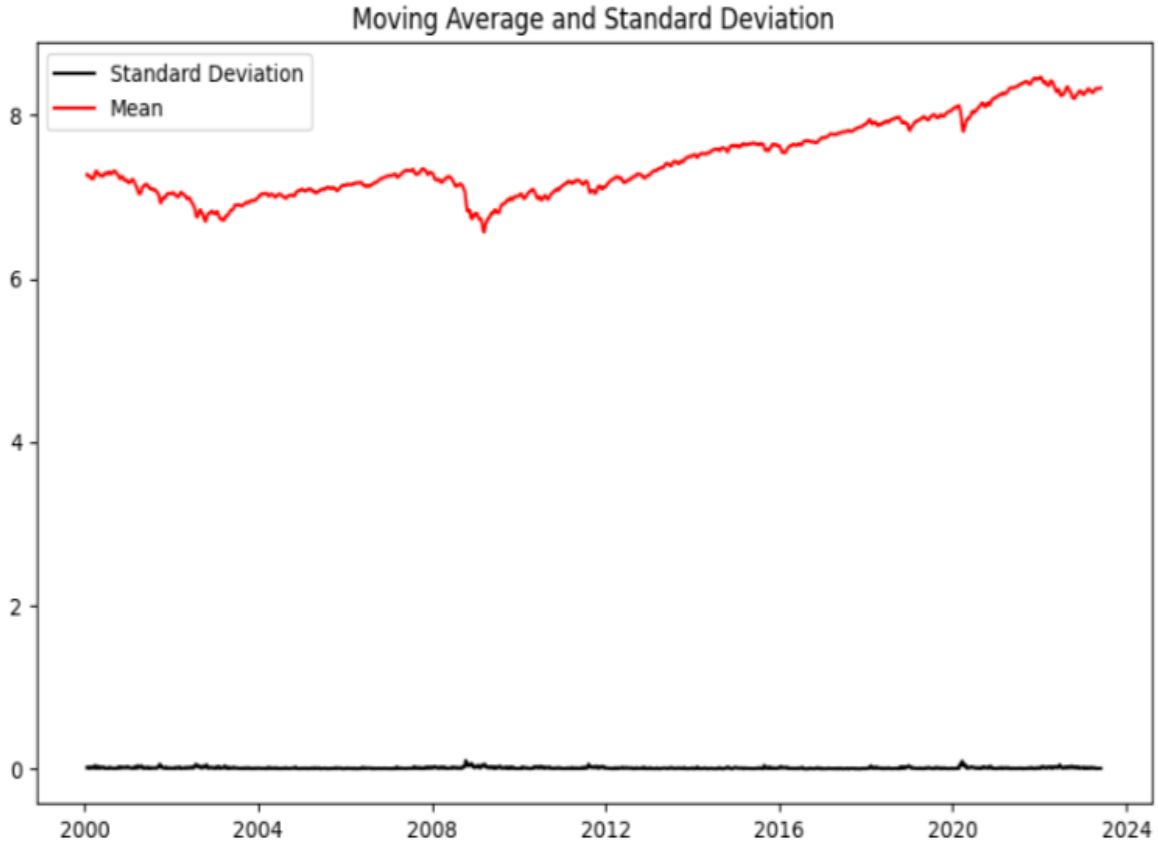


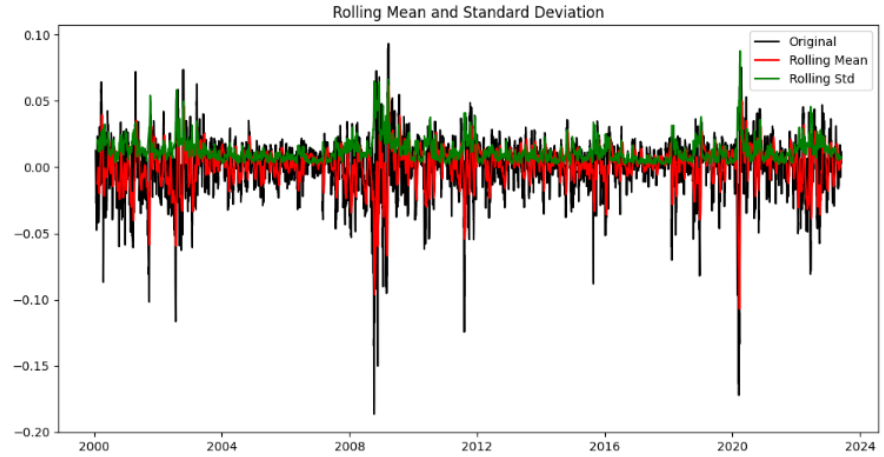
Figure 2.14 – Plot of Moving Average and Standard Deviation

The process described in the code (Fig. 2.15) is called "trend removal" or "detrending" a time series. It is an important step in time series analysis and can help reveal hidden features and patterns in the data. Trend removal is performed by subtracting the moving average from the original time series. This highlights shorter-term fluctuations, such as business cycles and seasonal patterns, and makes it easier to analyze these components.

```
# Calculate the difference between df_log and the moving average
data_log_minus_mean = df_log - moving_avg

# Remove any NaN values from the resulting series
data_log_minus_mean.dropna(inplace=True)

# Perform the ADF test on the modified series
test_stationarity(data_log_minus_mean)
```




```
Results of Dickey-Fuller test:
Test Statistics               -1.684634e+01
p-value                        1.127717e-29
No. of lags used               1.600000e+01
Number of observations used    5.862000e+03
Critical value (1%)           -3.431466e+00
Critical value (5%)           -2.862033e+00
Critical value (10%)          -2.567033e+00
dtype: float64
```

Figure 2.15 – Calculate the difference between df_log and the moving average and ADF results.

And also from (Fig. 2.15) available to see the updated result of the ADF test. Based on these results we can draw the following conclusions

– Test Statistic (-1.684634e+01) is less than the critical values for all significance levels (-3.431466e+00, -2.862033e+00 and -2.567033e+00). This indicates that we can reject the null hypothesis of unit root and accept the alternative hypothesis of stationarity of the time series.

– P-value (1.127717e-29) is very close to zero, which also confirms the statistical significance of the test results. Usually, if the p-value is smaller than the selected significance level [41], we can reject the null hypothesis and conclude that the series is stationary. In this case, the p-value is much less than 0.05, which confirms the stationarity of the series.

Thus, based on the results of the Dickey-Fuller test with a very low p-value and test statistics lower than the critical values, we can conclude that the time series is stationary. This ensures the stability of the model and allows the use of past data to accurately predict future values.

2.2 Applying the ARIMA model and accuracy assessment.

The Ideology of ARIMA Models. The publication by J. Box and G. Jenkins in 1978 laid the foundation for the development of new forecasting tools in econometrics. The methodology developed emphasizes the study of the stochastic properties of economic time series and carries the ideology of "letting the data speak for themselves". This ideology is based on the consideration that many economic phenomena have the property of inertia, i.e. the value of the time series "today" will be closely related (correlated) to its value "yesterday". This means that in order to build an effective model for predicting the dynamics of the time series, we should not look for any special explanatory variables (regressors). The ARIMA model is one of the most popular models for making short-term forecasts. To describe this model, three groups of parameters are used: p, d, and q are nonnegative integers that characterize the order of the model parts (autoregressive, integrated, and moving average, respectively). [42].

– Parameter p (AR - Autoregressive): Indicates the number of previous values of the time series used to predict the current value. If p=1, the model uses only one previous value. If p=2, the model uses two previous values, and so on. A larger value of p means that the model uses more previous values for prediction.

– The parameter d (I - Integration): It determines the number of differentiations needed to achieve stationarity of the series. Differentiations help to remove trends and seasonality in the series. If d=0, the series is considered stationary. If d=1, then the model uses the first difference value of the series. If d=2, the model applies the second difference value of the series, and so on.

– Parameter q (MA - Moving Average): It indicates the number of previous prediction errors used to predict the current value. If q=1, the model uses only the previous error. If q=2, the model considers two previous errors, and so on. A larger value of q means that the model uses more previous errors for prediction.

The parameters p, d, and q together define the structure of the ARIMA model. For example, an ARIMA(1, 1, 1) model means that an autoregression of order 1, a

differentiation, and a moving average of order 1 are used. The choice of the optimal values of the parameters p, d and q can be based on the analysis of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the time series, as well as on the use of statistical criteria and model evaluation methods [43]. In our work we will use the function of automatic selection of parameters (Fig. 2.16).

```
# Create and train an AutoARIMA model
model = auto_arima(df_log, start_p=1, start_q=1,
                   max_p=3, max_q=3, m=1,
                   start_P=0, seasonal=False,
                   d=None, D=None, trace=True,
                   error_action='ignore',
                   suppress_warnings=True,
                   stepwise=True)

# Obtaining optimal model parameters
order = model.order
```

Figure 2.16 – Create and train an AutoARIMA model

The auto_arima function (Fig. 2.16) performs automatic model parameter tuning and is a convenient tool that allows you to automatically select the optimal ARIMA model parameters based on statistical analysis and heuristic methods. It facilitates the process of model tuning with a simple function rather than searching for p, d and q parameters separately. The result is shown in (Fig. 2.17).

```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-34976.643, Time=2.63 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-34914.679, Time=0.50 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-34976.612, Time=1.01 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-34978.337, Time=5.80 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-34915.465, Time=0.31 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=-34976.660, Time=2.26 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=-34974.731, Time=2.53 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=-34978.785, Time=0.95 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=-34977.066, Time=2.11 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=-34977.083, Time=1.82 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=-34977.104, Time=0.68 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=-34975.161, Time=2.12 sec

Best model:  ARIMA(0,1,1)(0,0,0)[0]
Total fit time: 22.757 seconds
```

Figure 2.17 – Results of AutoARIMA model

The results (Fig. 2.17) are presented as a table, where each row corresponds to the ARIMA model with certain parameter values, and the columns show the following information:

– ARIMA(p, d, q)(P, D, Q)[m]: ARIMA model parameters, where p, d, q are order of autoregression, integration and moving average respectively, P, D, Q are order of seasonal autoregression, integration and moving average respectively, m is seasonal period (in our case have zero values due to choice of model and specifying seasonal=False parameter).

– AIC: Akaike Information Criterion (AIC), which is a measure of the relative quality of the model, where a smaller value of AIC indicates a better model.

Based on the results, the best ARIMA model is ARIMA(0,1,1)(0,0,0)[0], which has the lowest AIC value of -34978.785.

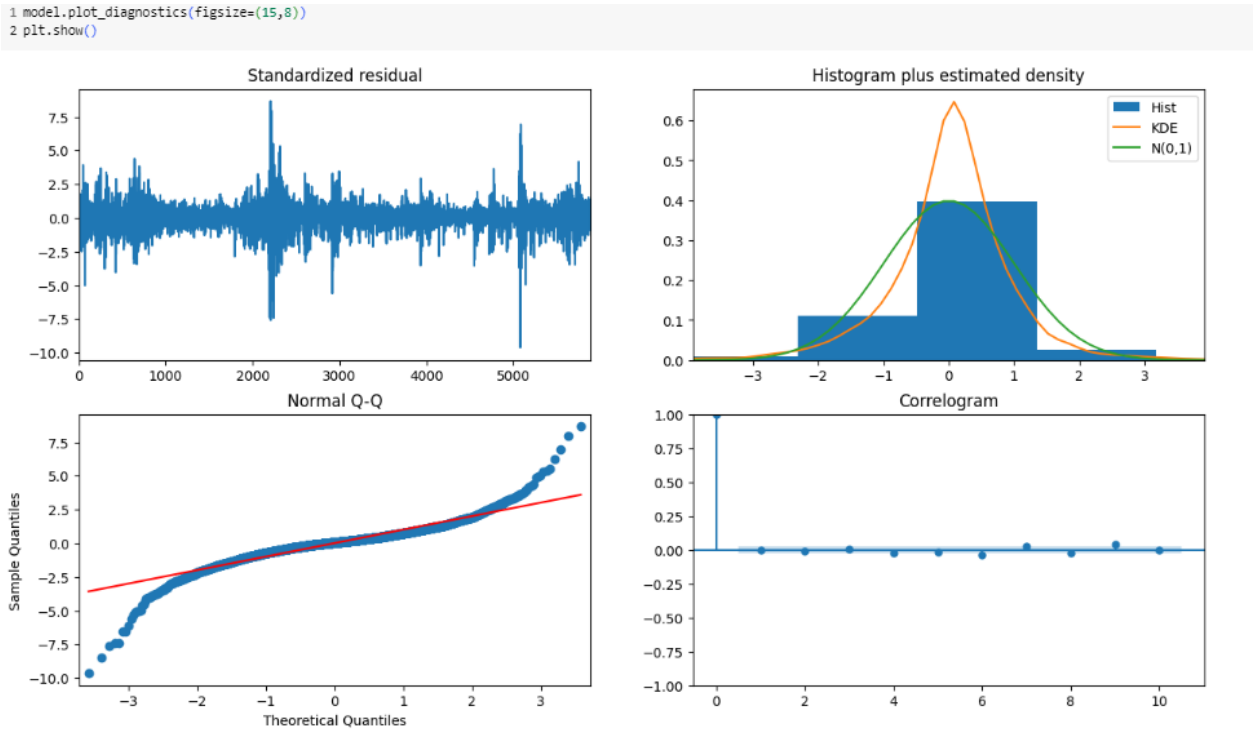


Figure 2.18 – Model plot diagnostics

The result of the (Fig. 2.18) shows to us:

– Top left: The residual errors, which are the difference between the predicted values and the actual values, are plotted. If the variance of the residuals is

homogeneous and they fluctuate around a mean of zero, this indicates that the model assumptions of constant error variance are met.

– Top right: The density plot shows the distribution of the residual errors. If the distribution appears to be approximately normal and centered around zero (mean of zero), this indicates that the model's assumption of normally distributed errors is reasonable.

– Bottom left: A red line is mentioned that should fit all points perfectly. This may refer to a plot of predicted versus actual values. If the points closely match the red line, it indicates that the model predictions are accurate and unbiased. Any significant deviation would indicate a possible systematic error or asymmetric error distribution.

– Bottom right: The correlogram, also known as the ACF (autocorrelation function) plot, shows the autocorrelation of the residual errors. If the residual errors are not autocorrelated, it means that there is no significant pattern or relationship between the errors at different delays. This is desirable because it indicates that the model is adequately capturing the temporal dependencies in the data.

Next, we need to divide our performance into test and training data at a ratio of 15:85 (Fig. 2.19).

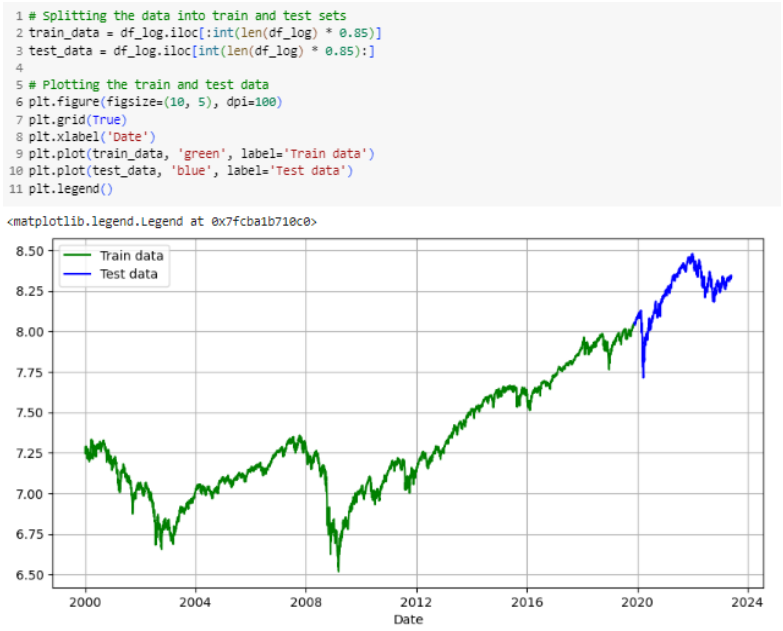


Figure 2.19 – Splitting the data into train and test sets

As a result, the Auto ARIMA model assigns the values 0, 1, and 1 to p, d, and q, respectively will input these parameters to our model (Fig. 2.20).

```
model = ARIMA(train_data, order=(0,1,1))
fitted = model.fit()
print(fitted.summary())
```

Figure 2.20 – Building and training an ARIMA model

The result of the model (Fig. 2.20) is available to see in (Fig. 2.21).

```
                               ARIMA  Results
==============================================================================
Dep. Variable:                  Close   No. Observations:                 5006
Model:                 ARIMA(0, 1, 1)   Log Likelihood               15082.405
Date:                Fri, 9  Jun 2023   AIC                         -30160.811
Time:                        15:25:27   BIC                         -30147.774
Sample:                             0   HQIC                        -30156.242
                                 - 5006
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0777      0.008     -9.649      0.000      -0.093      -0.062
sigma2         0.0001   1.24e-06    113.645      0.000       0.000       0.000
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):             15109.89
Prob(Q):                              0.94   Prob(JB):                         0.00
Heteroskedasticity (H):               0.51   Skew:                            -0.32
Prob(H) (two-sided):                  0.00   Kurtosis:                        11.49
===================================================================================
```

Figure 2.21 – ARIMA results

Based on the results of the ARIMA methods provided in (Fig. 2.21), we can make an express test [44]:

– The coefficient ma.L1 is -0.0777, which means that the model uses a lag of the difference series to predict public values. The negative sign of the hazard of detection between the variances and the current value of the relationship.

– The sigma2 is 0.0001. This is a very small value that the model is good at predicting the estimated time series.

– The Ljung-Box criterion (Q) has a value of 0.01 and the p-value (Prob(Q)) is 0.94. This indicates that the autocorrelations of the residuals in the first lag are not significant.

– The value of heteroskedasticity (H) is 0.51, which indicates an increase in the heteroskedasticity of the signs.

Thus, we can say that the ARIMA (0, 1, 1) model gives good results because it has a significant factor and a low variance of the residuals and we can proceed with the forecast (Fig. 2.22).

```
# Forecast
forecast = fitted.get_forecast(steps=884, alpha=0.05)

# Get forecast, standard errors and confidence intervals
fc_series = forecast.predicted_mean
se_series = forecast.se_mean
lower_series = forecast.conf_int().iloc[:, 0]
upper_series = forecast.conf_int().iloc[:, 1]
```

Figure 2.22 – Forecasting with a trained ARIMA model and obtaining predictive values

Obtaining forecast values is the main purpose of applying the ARIMA model. Predictions can be used to predict future values of a time series based on available historical data. The standard errors returned by the se_mean method allow you to estimate the uncertainty or scatter of the predicted values. The smaller the standard error, the more accurate and reliable the predictions will be. Confidence intervals obtained with the conf_int method provide information about the likely range in which the future values of the series will lie. Confidence intervals help assess the uncertainty of predictions and can be used to make more informed decisions based on the probability that future values will fall within a certain range. The result of the forecasting (Fig. 2.22). is available to see in (Appendix B).

The plotting of the forecast data is available to see in (Fig. 2.23).

```
# Plotting
plt.figure(figsize=(18, 8), dpi=100)
plt.plot(df_log, label='Training data')
plt.plot(test_data, color='green', label='Actual Index Price')
plt.plot(fc_series, color='orange', label='Predicted Index Price')
plt.fill_between(lower_series.index, lower_series, upper_series, color='gray', alpha=0.1)
plt.grid(True)
plt.title('SP500')
plt.xlabel('Date')
plt.legend(loc='best', fontsize=8)
plt.show()
```

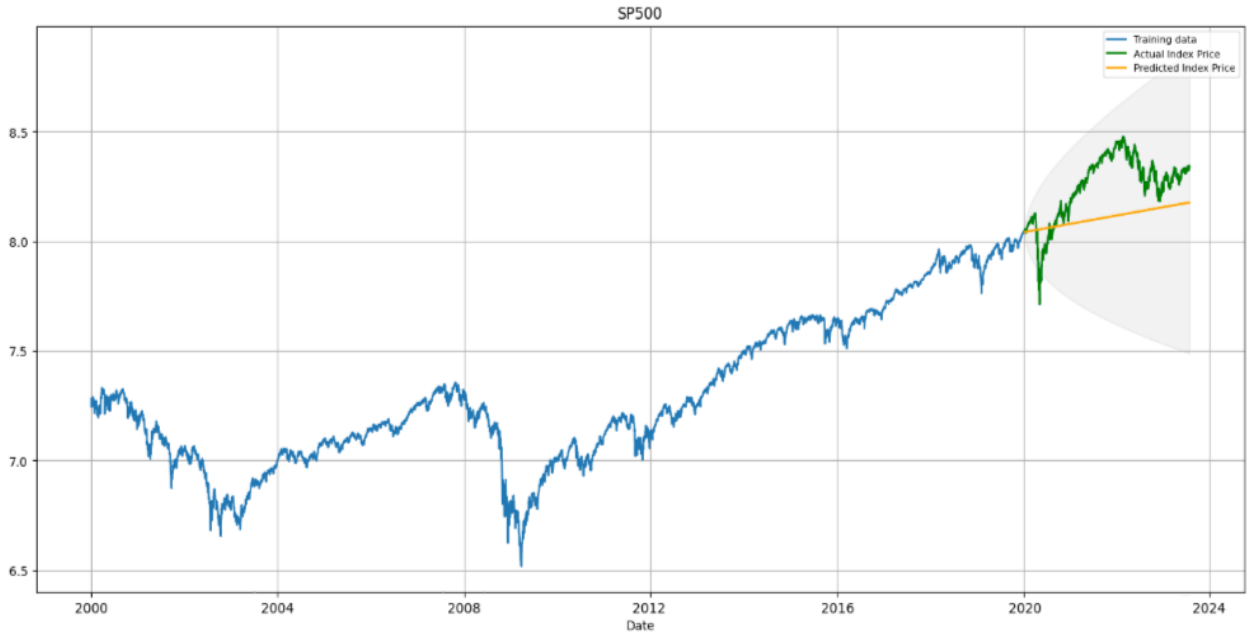


Figure 2.23 – Plot with Predicted Index price

When we evaluate the results of the forecast plot (Fig. 2.23), we can mention that it looks realistic and close to the test data. The only thing that stands out is March 2020 period, when we lost about 30% of the index value in one week, but we all know what an out-of-state situation is due to the pandemic and the general panic in the market. Such situations are extremely difficult to predict and such collapses are only possible by analyzing the news background. Similar crashes can also happen as a result of force majeure, for example, the recent incident that happened to Equifax, the large-scale cyber-attack that compromised the personal information of approximately 147 million people. As a result of the attack, Equifax's stock price plummeted and the company suffered significant financial losses - the stock price continued to decline in the following weeks, with a total loss in value of approximately 35% [45]. But it is only relevant in the context of one specific company - this is another advantage of working with indices - such events in one

company do not have a critical impact on the index as a whole. And our predictable data clearly show the data to which we returned after the correction, which gives us the opportunity to make long-term forecasts, in our case the forecast was for the next 884 steps/day.

We also need to be sure that we can call our model an accurate one. There are many metrics to evaluate the quality of a model. These (Fig. 2.24) metrics allow you to evaluate how accurately and reliably the model is able to predict the values of the S&P 500 Index under study.

```
# Calculation of MAE (Mean Absolute Error)
mae = mean_absolute_error(test_data, fc_series)
# Calculation of MSE (Mean Squared Error)
mse = mean_squared_error(test_data, fc_series)
# Calculation of RMSE (Root Mean Squared Error)
rmse = np.sqrt(mse)
# Calculation of MAPE (Mean Absolute Percentage Error)
mape = np.mean(np.abs((test_data - fc_series) / test_data))

# Printing the results
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("MAPE:", mape)
```

```
MAE: 0.16072026683536186
MSE: 0.0347573081524168
RMSE: 0.1864331197840577
MAPE: 0.01934220013333761
```

Figure 2.24 – Calculation of ARIMA model evaluation metrics

Interpretation of the results [46-48]:

– MAE (Mean Absolute Error): The average difference between the predicted values and the actual values is approximately 0.161. It represents the average magnitude of the errors without considering their direction. On average, the forecasted values deviate from the actual values by around 0.161 units.

– MSE (Mean Squared Error): The average of the squared differences between the predicted values and the actual values is approximately 0.035. It measures the average squared magnitude of the errors, giving more weight to larger errors. A smaller MSE indicates a better fit between the predicted and actual values.

– RMSE (Root Mean Squared Error): The square root of MSE is approximately 0.186. It provides a measure of the average magnitude of the errors in the same unit as the original data. In this case, the RMSE suggests that, on average, the forecasted values deviate from the actual values by around 0.186 units.

– MAPE (Mean Absolute Percentage Error): The average percentage difference between the predicted values and the actual values is approximately 0.019. It measures the average magnitude of the errors as a percentage of the actual values. In other words, on average, the forecasted values deviate from the actual values by around 1.9%.

CONCLUSION

In the course of the work, the theoretical understanding of the stock market and its peculiarities was expanded, including the understanding of the nature of indices. The relationship between the S&P 500 index and the U.S. economy was analyzed. The results of the research allowed us to determine that the chosen index can serve as a reflection of the state and forecasts of economic development of the United States. A successful forecast of the index can serve not only as a key point in building an individual investment strategy, but also as an indicator of the general state of the economy.

Then, all the goals and objectives for the construction of a mathematical model for predicting the dynamics of the index were painted. Through exploratory data analysis, we gained a better understanding of the time series and its characteristics. The application of various statistical methods, such as moving statistics and stationarity tests, allowed us to identify trends and seasonality in the data. Seasonal decomposition and logarithmic transformation helped us to better understand the contribution of each component to the overall index dynamics, and special attention was paid to the stationary ADF test, where we considered not only the code but also the significant formulas. The optimal selection of parameters was done automatically. he built ARIMA model showed good results - the evaluation of the model accuracy included the comparison of the predicted values with the actual values of the SP500 index, both visually and using several metrics - MAE, MSE, RMSE, MAPE.

The result of the work is a model for predicting the dynamics of the S & P 500 index, implemented using the Python programming language with a MAPE of about 1.9%, the accuracy of the model is 98.1%. and such good results indicate the possibility of using this tool by market participants in real conditions.

REFERENCES

1. Johannes W. Flume. (2021). Constructing the Stock Exchange: On the Rise and Fall of an Iconic Place of Capitalism. URL: https://www.academia.edu/70401341/Constructing_the_Stock_Exchange_On_the_Rise_and_Fall_of_an_Iconic_Place_of_Capitalism

2. Investopedia. (2023). Stock Market. Investopedia. URL: https://www.investopedia.com/terms/s/stockmarket.asp

3. SoFi. (2023). History of the Stock Market. URL: https://www.sofi.com/learn/content/history-of-the-stock-market/

4. IndiaCharts. (2022). Stock Market History. URL: https://www.indiacharts.com/stock-market/history/

5. Malkiel, B. G. (2015). A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing. W. W. Norton & Company.

6. Wiley, J. (2017). The New Stock Market: Law, Economics, and Policy. John Wiley & Sons.

7. Fatemeh Aramian. (2021). Off-exchange Trading in Modern Equity Markets. URL: https://www.diva-portal.org/smash/get/diva2:1612002/FULLTEXT01.pdf

8. Henry, P. B. (1997). Stock market liberalization, economic reform, and emerging market equity prices. Journal of Finance, 52(3), 1227-1246.

9. Mishkin, F. S., & Eakins, S. G. (2014). Financial Markets and Institutions. Pearson; 8th edition

10. Pisano, U. A., Martinuzzi, B., & Bruckner, B. (2012). The Financial Sector and Sustainable Development: Logics, principles, and actors. ESDN Quarterly Report, URL:https://www.researchgate.net/publication/312495549_Pisano_U_A_Martinuzzi_B_Bruckner_2012_The_Financial_Sector_and_Sustainable_Development_Logics_principles_and_actors_ESDN_Quarterly_Report_N2

11. Author(s). (2016). The Impact of Stock Market Performance upon Economic Growth.URL:https://www.researchgate.net/publication/290189993_The_Impact_of_Stock_Market_Performance_upon_Economic_Growth

12. Merritt B. Fox. (2021). The Social Functions of the Stock Market. URL: https://clsbluesky.law.columbia.edu/2019/04/12/the-social-functions-of-the-stock-market-a-primer/

13. Justin Kuepper. (2022). Major World Stock Market Indexes. The Balance. URLhttps://www.thebalancemoney.com/major-world-stock-market-indexes-4148491

14. John Egan. (2023). Dow vs. Nasdaq vs. S&P 500: What's the Difference? URL: https://time.com/personal-finance/article/dow-vs-nasdaq-vs-s-p/

15. Investopedia. (2023). Composite Index: Definition, Types, and Examples. URL https://www.investopedia.com/terms/c/compositeindex.asp

16. Ellis, C. D. (2016). The Index Revolution: Why Investors Should Join It Now. Foreword by B. G. Malkiel. Wiley; 1st edition

17. Ganeshwaran Kana. (2022). Why the Stock Market Index Is Important for You. URL: https://www.nasdaq.com/articles/why-the-stock-market-index-is-important-for-you

18. Josef Novotný, Iveta Jaklová. (2022). The Importance of Global Financial Indices for Investors in Alternative Investment. URL: https://www.researchgate.net/publication/348463663_The_Importance_of_Global_Financial_Indices_for_Investors_in_Alternative_Investment

19. S&P Global. (2023). S&P 500 Brochure. URL: https://www.spglobal.com/spdji/en/documents/additional-material/sp-500-brochure.pdf

20. S&P Global. (2023). S&P 500 - Overview. URL: https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview

21. S&P Dow Jones Indices. (2023). S&P 500 Brochure, page 4. URL: https://www.spglobal.com/spdji/en/documents/additional-material/sp-500-brochure.pdf

22. Corporate Finance Institute. (2023). S&P 500 Index. URL: https://corporatefinanceinstitute.com/resources/equities/sp-500-index/

23. Author(s). (2021). Title of the article. Complexity. URL: https://www.hindawi.com/journals/complexity/2021/6645570/

24. Investopedia. (2023.). How the Stock Market Affects the Economy. URL: https://www.investopedia.com/how-stock-market-affects-economy-5296138/

25. Francisco Jareño, (2016). "US Stock Market and Macroeconomic Factors." URL:https://www.researchgate.net/publication/282292396_US_Stock_Market_and_Macroeconomic_Factors

26. Dattatray P. Gandhmal. (2019). "Systematic analysis and review of stock market prediction techniques" Computer Science Review, Volume 36. URL: https://www.sciencedirect.com/science/article/abs/pii/S157401371930084X

27. K. A. Malyshenko, M. V. Anashkina, (2014), USE OF NEURAL NETWORKS FOR THE PURPOSES OF FORECASTING THE STOCK MARKET, Efficient Economics No. 2. URL: http://www.economy.nayka.com.ua/?op=1&z=2744

28. Hyndman, R. J. (2018). Forecasting: Principles and Practice ARIMA Models. URL: https://people.duke.edu/~rnau/411arim.htm

29. Sheereen Fauzel. (2016). A Generalized Autoregressive Conditional Heteroscedastic Approach for the Assessment of Weak-form Efficiency and Seasonality Effect: Evidence from Mauritius. URL: https://www.researchgate.net/publication/301959647_A_Generalized_Autoregressive_Conditional_Heteroscedastic_Approach_for_the_Assessment_of_Weak-form-efficiency_and_Seasonality_Effect_Evidence_from_Mauritius

30. Uzakariya, M. (2021). Project: Predict Stock Prices Using Random Forest Regression Model in Python [Blog post]. URL:https://medium.com/@maryamuzakariya/project-predict-stock-prices-using-random-forest-regression-model-in-python-fbe4edf01664

31. Serafeim Loukas. (2020). LSTM Time Series Forecasting: Predicting Stock Prices Using an LSTM Model. Towards Data Science. URL https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f

32. Kris Longmore. (2020). A Vector Autoregression Trading Model. Robot Wealth. URL: https://robotwealth.com/a-vector-autoregression-trading-model/

33. Dat Tan Trinh. (2022). Stock Price Forecasting Based on XGBoost and LSTM. URL:https://www.researchgate.net/publication/357486637_Stock-price_forecasting_based_on_XGBoost_and_LSTM

34. Md Aminur Rahman. (2023). "Title of the article." International Journal of Advances in Business and Management (IJABM), Volume(Issue), Page range. URL: https://journal.formosapublisher.org/index.php/ijabm/article/view/2901/2676

35. Yahoo Finance. (2023). S&P 500 historical data 1990-2023. URL https://finance.yahoo.com/quote/%5ESPX/history?p=%5ESPX

36. Stanislaw Weglarczyk (2018). Kernel Density Estimation and Its Application. URL:https://www.researchgate.net/publication/328785939_Kernel_density_estimation_and_its_application

37. Wooldridge, J. M. (2019). Introductory Econometrics: A Modern Approach. Cengage Learning.

38. Yukai Yang. (2022). A Comparative Study of the KPSS and ADF Tests in terms of Size and Power. URL: https://www.diva-portal.org/smash/get/diva2:1668033/FULLTEXT01.pdf

39. Charles Zaiontz. (2018). Augmented Dickey-Fuller Test. URL https://real-statistics.com/time-series-analysis/autoregressive-processes/augmented-dickey-fuller-test/

40. Selva Prabhakaran. (2019). Augmented Dickey-Fuller Test. URL: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/

41. Raoof Naushad. (2020). Interpreting Results of Dickey-Fuller Test for Time Series Analysis. Medium. URL: https://medium.datadriveninvestor.com/interpreting-results-of-dicky-fuller-test-for-time-series-analysis-4bb1e98f242b

42. Chumachenko D. I. Mathematical models and methods for predicting epidemic processes: monograph / D.I. Chumachenko, T. O. Chumachenko. - Kharkiv: TOV "Planeta-Print", 2020. - 180 s

43. Hyndman, R. J. (2018). Forecasting: Principles and Practice - Identifying the numbers of AR or MA terms in an ARIMA model. URL: https://people.duke.edu/~rnau/411arim3.htm

44. Leo Smigel. (2020.). How to Interpret ARIMA Model Results. URL: https://analyzingalpha.com/interpret-arima-results

45. Dun V., Mynenko S. Analysis of the Impact of Major Cyber Incidents on ohe company's stocks // Cybersecurity Challenges Facing the Financial Services Industry: matherial of the International virtual conference, Sumy, Ukraine, June, 2 2023. Sumy: Sumy State University, 2023. P. 44-47.

46. Nochovny, O.O. (2020). Information system for forecasting stock quotes using machine learning methods. URL: https://ela.kpi.ua/bitstream/123456789/40536/1/Nochovnyi_magistr.pdf

47. Alexander Dyakonov. (2022). Small data analysis - Quality Metrics. URL:https://alexanderdyakonov.files.wordpress.com/2018/10/book_08_metrics_12_blog1.pdf

48. Nicolas Vandeput. (2019). Forecast KPI - RMSE, MAE, MAPE, Bias. Towards Data Science. URL: https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Appendix A

SUMMARY

Dun V.R. Economic and mathematical modeling of financial asset returns using Python. Qualifying bachelor's thesis. Sumy State University, Sumy, 2023.

The paper explores the features of modeling and forecasting the financial market, in particular the S&P500 index, considers the relationship between the S&P 500 index and the US economy and these roles and functions. A stationarity test was carried out. The Dickey-Fuller test. The ARIMA model for forecasting the S&P 500 index was built. The practical part was performed using the Python programming language. Model adequacy tests were carried out.

Keywords: Time series forecasting, stock market, investments, Python, S&P 500 index, stationarity test, ARIMA model, moving statistics.

АНОТАЦІЯ

Дунь В.Р. Економіко-математичне моделювання дохідності фінансових активів з використанням Python. Кваліфікаційна робота бакалавра. Сумський державний університет, Суми, 2023.

У роботі досліджено особливості моделювання та прогнозування фінансового ринку, зокрема індексу S&P500, розглянуто зв'язок між індексом S&P 500 та економікою США та в цілому - їх рол та функції. Проведено тест стаціонарності Тест Дікі – Фуллера. Побудовано модель APIMA для прогнозування індексу S&P 500. Практична частина виконана за допомогою мови програмування Python. Проведено тести адекатності моделі.

Ключові слова: прогнозування часових рядів, фондовий ринок, інвестиції, Python, індекс S&P 500, тест на стаціонарність, модель ARIMA, ковзаюча статистика.

Appendix B

Output of forecast values, standard errors and confidence intervals.

```
1 print("Forecast:\n", fc_series)
2 print("Standard Errors:\n", se_series)
3 print("Confidence Intervals:\n", lower_series, upper_series)
```

```
Forecast:
 5006    8.042497
5007    8.042538
5008    8.042423
5009    8.041701
5010    8.042268
          ...
5885    8.175310
5886    8.175876
5887    8.176458
5888    8.175910
5889    8.175348
Name: predicted_mean, Length: 884, dtype: float64
Standard Errors:
 5006    0.011890
5007    0.016126
5008    0.019461
5009    0.022303
5010    0.024821
          ...
5885    0.350065
5886    0.350294
5887    0.350523
5888    0.350751
5889    0.350979
Name: var_pred_mean, Length: 884, dtype: float64
Confidence Intervals:
 5006    8.019193
5007    8.010931
5008    8.004280
5009    7.997989
5010    7.993619
          ...
5885    7.489194
5886    7.489313
5887    7.489446
5888    7.488451
5889    7.487441
Name: lower Close, Length: 884, dtype: float64 5006    8.065801
5007    8.074144
5008    8.080566
5009    8.085414
5010    8.090916
          ...
5885    8.861425
5886    8.862440
5887    8.863469
5888    8.863370
5889    8.863254
Name: upper Close, Length: 884, dtype: float64
```