**Abstract**

In the ever-evolving landscape of financial markets, predicting stock returns remains a challenging and vital task for investors seeking informed decision-making. This research endeavors to advance stock return predictability by strategically applying machine learning algorithms, specifically Random Forest, Linear Regression, and Decision Tree models. Recognizing the inherent complexities of forecasting stock behavior, especially with reliance on publicly available data using some Macroeconomic Factors and Technical Indicators, the study explores the intersection of machine learning and stock return predictability. The emphasis lies not only on traditional statistical metrics like RMSE and MAE but also on essential investment metrics such as the Sharpe Ratio, cumulative sum of return, and Percentage of Direction Agreement. This holistic evaluation approach provides valuable insights for practical investment decision-making. The findings contribute to the ongoing dialogue on leveraging advanced algorithms, offering a structured and succinct exploration suitable for an end-of-studies report.

+to add

Dediction

To my dearest family,

I dedicate these words of gratitude to express my deepest appreciation for your unwavering support and encouragement throughout my journey. Your boundless belief in my aspirations and constant encouragement have been the driving force behind my endeavors. Your willingness to stand by me in every decision, cheering me on as I pursued my goals, has been a source of strength and inspiration. Thank you for being the pillars of support that have enabled me to reach new heights.

To my cherished friends,

I extend my heartfelt thanks for being the uplifting companions on this journey. Your camaraderie, shared laughter, and understanding have added immeasurable joy to my experiences. Thank you for standing by me through thick and thin, for celebrating my successes, and offering solace during challenges. Your friendship has been a priceless gift, and I am grateful for the shared moments that have enriched my life.

A special thank you to Hamza and Malek for their invaluable assistance and efforts during this internship. Their support has made this period smoother and more enjoyable for me. I deeply appreciate their generosity and kindness throughout this journey.

With heartfelt appreciation,

[Your Name]

Acknowledgements

I am deeply grateful to several individuals whose support and guidance were instrumental in the successful completion of this internship. I would like to express my sincere gratitude to each of them:

First and foremost, I would like to express my sincere appreciation to my professional supervisor, Mr. Wajdi Tekaya, for giving me the opportunity to work on this exciting topic at Quant-Dev. His invaluable mentorship, unwavering encouragement, and the chance to delve into such a compelling project have been crucial to my growth and learning during this internship.

I would like to thank the entire Quant-Dev team for their generous support and warm welcome into their team. Their help has been invaluable and I appreciate the collaborative spirit that has made my experience truly enriching.

I am sincerely grateful to my academic supervisor, Prof. Tristan Cazenave. It has been a pleasure to be his student and to benefit from his efforts to impart insightful knowledge. His accessibility, guidance and invaluable advice were instrumental in the success of this project.

I would like to acknowledge the support of Paris Dauphine University Campus Tunis for providing me with a scholarship that made this internship possible. Your commitment to education has played a pivotal role in shaping my academic and professional journey.

Finally, my sincere thanks to the members of the jury who took the time to evaluate my work. Your insights and feedback are greatly appreciated.

This internship has been a remarkable and enjoyable experience thanks to the collective contributions of these individuals. Their support has been instrumental in my personal and professional development, and I am truly grateful for the opportunities they have provided.

General Introduction :

As time progresses, Technology keeps showing up in the environment around us, making it more and more helpful. The development of artificial intelligence, which has attained a remarkable degree of proficiency across numerous fields, makes this tendency particularly clear. In our pursuit, we turn our attention to a particularly groundbreaking facet: artificial intelligence's transformative impact on economies, notably exemplified within stock markets.

In the ever-evolving landscape of today's information age, the value of data has transcended its conventional significance, echoing the sentiment of British mathematician Clive Humby's proclamation in 2006 that "Data is the new oil." The pivotal role of data in decision-making processes is irrefutable, yet its mastery, confidentiality, and precision remain formidable challenges, particularly in complex domains such as trading and stock markets.

Traditionally, the stock market has been a domain that evokes apprehension in those who either eschew risk or find the intricacies of trading daunting. However, the contemporary landscape is distinct from the past. Propelled by the advancements in Artificial Intelligence and Machine Learning, a new era has emerged where the once-daunting prospect of entering the stock market has become considerably less intimidating.

In essence, the synergy of artificial intelligence and stock market dynamics is changing how people make investments. Investors are in a position to turn uncertainty into opportunity, optimize their strategies, and increase their chances of success by taming the natural volatility through predictive skill. This significant change highlights how technology can alter markets as well as democratize access to previously closed-off areas, ushering in a new era of knowledgeable and assured investment.

The stock market's unpredictable nature presents a significant obstacle for investors. The objective is not only to predict future stock prices but also to identify potential volatility in advance to strategically minimize risk and maximize returns.

To align with this objective, our graduation project seeks to implement a machine-learning model. The purpose is to aid investors in making well-informed decisions about whether to buy or sell stocks today while minimizing risk, therefore improving their comprehension of possible outcomes for the following day. To accomplish this goal, we will build a predictive regression model, opting for an approach that reduces risk and maximizes return on investment.

Our primary objective is to identify the most efficient model while minimizing risk and guiding investors toward maximizing their gains.

This report encompasses the progress we have made since July, which marks the initiation of our internship at Quant-Dev.

This report has the following structure:

* **Chapter 1: General Framework of the Project:** In this section, an overview of the project's context and the financial markets and data we'll be using. We also describe the project goal and present elements of the literature review on the topic.
* **Chapter 2: Mathematical Background** This chapter focuses on understanding the modeling tools utilized and the metrics employed to assess their performance. Initially, we briefly overview the models employed and introduce the cross-validation technique, which leverages available data. Subsequently, we present the metrics used to evaluate results, utilizing two main types of evaluations: Statistical Evaluation and Investment Evaluation.
* **Chapter 3: Modeling and Empirical Results:** This chapter introduces the programming language utilized in our project. We then proceed to discuss the exploratory data analysis and feature engineering. Finally, we present an interpretation of the results and critique the suggested models.
* **Chapter 4: Conclusion and Perspectives:** This section provides an overview of our efforts and offers some perspectives.

**Chapter 1**

**General Framework of the Project**

**1.1 Introduction**

In the first chapter, we introduce Quant-Dev, the host company, and provide a contextual background for this internship. We then outline the financial terms and data that are essential to our project, as well as the macroeconomic factors critical to our study. After that, we articulate the problem statement and finally delve into the explicit objectives guiding our work.

**1.2 Presentation of the host company: Quant-Dev :**

Founded in 2014 and headquartered in Tunis, Tunisia, Quant-Dev is a dynamic consulting company specializing in developing data-driven quantitative tools and models. With over a decade of expertise in mathematical modeling and optimization, the company has positioned itself as a key player in the realm of investment management, quantitative trading, and high-frequency trading.

Quant-Dev's main client is a multi-strategy investment fund based in New York City, engaging in global trading across equities, futures, fixed income, and derivatives. This collaborative project exemplifies Quant-Dev's commitment to delivering innovative solutions that leverage advanced quantitative methodologies to optimize trading strategies and enhance overall portfolio performance. As we delve into the details of this project, we witness how Quant-Dev's expertise unfolds in addressing real-world financial challenges, reflecting its pivotal role in the evolving landscape of quantitative development.

**1.3 General Context**

**1.3.1. Machine Learning:**

Developing algorithms and statistical models that allow computers to learn from data and make predictions or judgments based on them is the main goal of the branch of artificial intelligence known as Machine Learning. It encompasses various techniques and approaches, from supervised learning to unsupervised learning and reinforcement learning.

For this, ML methods have been developed in many research papers to evaluate the prediction power of AI in the stock markets. The ML algorithms implemented for this purpose mostly try to figure out patterns of data, measure the investment risk, or predict the investment future. Nowadays, the power of ML strategies in addressing the stock market prediction problem is strengthening rapidly in both fundamental and technical analyses.[Effectiveness of Artificial Intelligence in Stock Market..]

**1.3.2 Stock Market**

Among this intricate web of data-driven dynamics, the stock market emerges as a commanding realm, embodying the essence of economic fluctuations and investor aspirations. The stock market often likened to a vibrant financial tapestry, operates as a conduit for trading shares and various financial instruments of publicly listed companies. Within its domain, the price of shares takes center stage, symbolizing the tangible valuation of a company's worth. This enigmatic world, brimming with potential riches and intricate strategies, has been the bedrock of countless fortunes and economic milestones.

The Stock Market is an important place for trading that affects individuals and even whole countries. The idea is simple: companies offer their stocks, which are like tiny parts of their ownership, to raise money. This happens at the start of something called an Initial Public Offering (IPO). The IPO price is what the company first asks for its stocks. People who buy these stocks can then sell them to others at places like the Stock Exchange. The stocks belong to the owners, and they can decide how much to sell them for. If the buying and selling go well, the stock's value can go up. But if the company makes more stocks available at a lower starting price (IPO), the stock's value can drop, causing losses for traders. This is why some people worry about investing in stocks and why prices go up and down.

The Stock Market is an ancient way for people and businesses to trade ownership in companies, invest money, and make profits. Companies sell small pieces of themselves, known as stocks, on this platform to raise funds. It can be a good way to invest if you're careful and follow a plan. But the prices and how easily you can buy or sell stocks can change a lot and are hard to predict. This is where technology helps. Machine learning is one tool that can assist us.

**1.3.3. S&P 500 index**

When one explores the stock market's relevance in-depth, one discovers a network of indices that have an impact on both national and international economies. The **S&P 500**, **Dow Jones Industrial Average**, and **Nasdaq** Composite are three significant economic indicators that have emerged in the United States and have come to be closely reviewed by the media and discerning investors. Not to be disregarded, a vast selection of more than 5,000 other indices embraces the varied aspects of the US equities market and provides a broad perspective of financial landscapes. Thus, this explains why the stock markets today represented a crucial turning point for the American economy.

Due to its tough selection criteria set by the Index Committee, the S&P 500 index will be the primary focus of our research. This index plays a key role in measuring both the U.S. and 3 global economies. To cut, companies must meet a series of demanding requirements, including being primarily U.S.-based, having a market capitalization surpassing USD 8.2 billion, maintaining high liquidity shares, trading over 50% of their outstanding shares publicly, and reporting positive earnings for the most recent quarter, as well as a positive aggregate for the preceding four quarters' earnings. Fulfilling these conditions is a testament to the exceptional success of the included companies, given the rigorous nature of the prerequisites.

Additionally, the S&P 500 Index, which represents around 83% of the total market capitalization of regularly traded stocks across major stock exchanges, is frequently used as a benchmark for assessing the performance of large U.S. corporations. The index, which was first created in 1957, is still dynamic and undergoes frequent modifications to maintain a representative selection of top businesses from diverse industries. The S&P 500 index has consistently outperformed most active money managers and mutual funds throughout its existence, enhancing its importance.

In addition to all of that, there are some pitfalls for the S&P 500 index is it's float-adjusted, and it’s explained that the index is weighted toward large-cap companies. The weighted average market capitalization is calculated for each component by dividing the market capitalization of the company by the index’s total market cap. As an example, we will choose the symbol of Apple (AAPL) which represents the largest market cap among stocks, with a value of $3 trillion on July 9, 2023, and we should know that, on June 30, 2023, the total market cap of all the companies in the index is equal to $39.13 trillion so the Apple’s weight in the index represented 7.70% approximately. This leads to the mega-cap stocks having an outsized impact on the index. Sometimes, this index structure can mask strength or weakness in smaller companies if large-cap companies are diverging. Despite these considerations, the S&P 500 index continues to be favored due to its historical significance and broad representation of the market.

**1.3.3.4 Financial Data:**

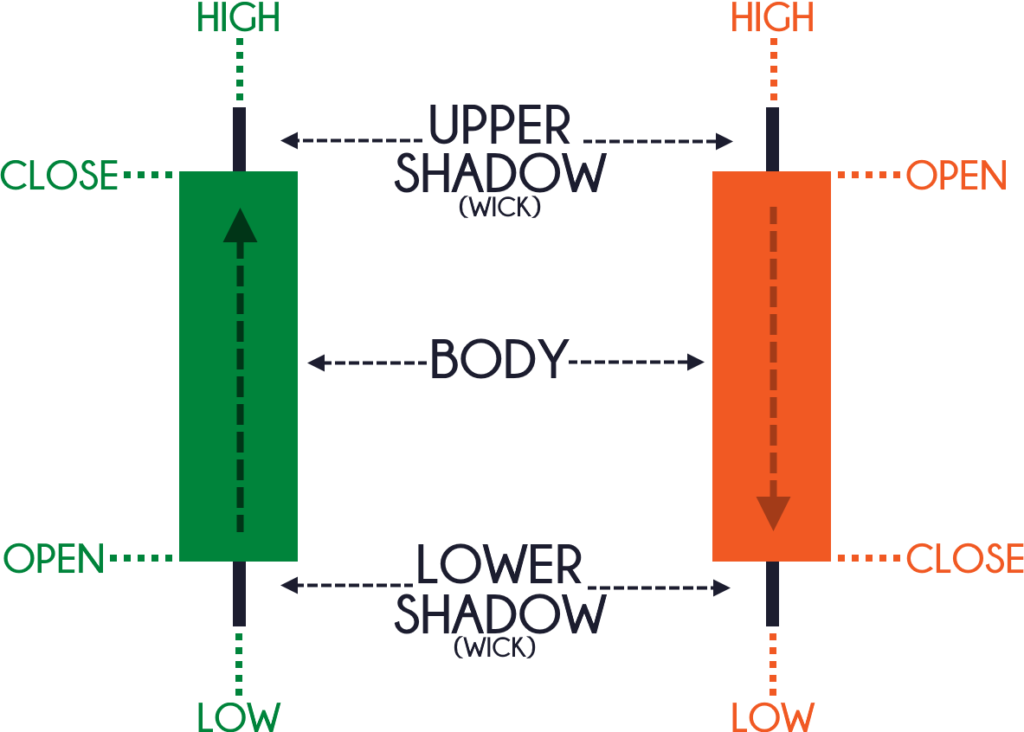
The Financial Data contains crucial information necessary for the understanding and evaluation of a financial instrument’s performance. Two important types of financial data are Technical Data and Fundamental Data.

* **Technical Data:** involves analyzing historical market data and price movements to predict trends. It uses charts, patterns, and technical indicators to identify potential entry and exit points for traders.
* **Fundamental Data:** explores the underlying financial health and performance of a company or financial instrument. It includes financial reports and economic indicators that provide valuable insight into the intrinsic value and long-term sustainability of an asset.

In our project, we’re going to focus on the Technical Data.It’s a data featuring historical stock prices that provides crucial information, such as daily or minute-by-minute prices, opening and closing values, as well as trading volumes. Investors and analysts use this data to analyze trends, conduct technical analysis, and make informed investment decisions. The dataset serves as a fundamental base for creating financial models and predictive algorithms intended for forecasting future stock prices.

To elucidate the significance of these features, let us delve into the core elements that comprise this financial dataset:

* **Open, High, Low, and Close (OHLC) Prices**: These four prices encapsulate a wealth of information about the asset's trading activity.



* **Volume**: represents the total number of shares of the asset that were traded on the given trading day, its measure is a critical indicator of market interest and liquidity.
* **Adj Close** or "**Adjusted Closing Price**", is a financial term used in stock market analysis that takes into account various corporate actions and events that can influence the stock’s value over time. These events include:
* **Dividends**: When a company pays out dividends to its shareholders, the stock price tends to decrease by the amount of the dividend. The adjusted closing price accounts for this reduction, allowing investors to see the true price movement of the stock.
* **Stock Splits**: A stock split occurs when a company divides its existing shares into multiple new shares. For example, in a 2-for-1 stock split, each existing share becomes two new shares. The stock price is adjusted to reflect this change. The adjusted closing price ensures that the historical performance of the stock remains consistent despite the split.
* **Reverse Splits:** Reverse splits are the opposite of stock splits. They involve consolidating multiple shares into one share. The adjusted closing price accounts for this change as well.

⇒ The adjusted closing price provides a clearer picture of how the stock has performed over time and helps investors assess the true historical performance of a stock while accounting for events.

* **Daily Return:** From one trading day to the next, daily returns represent the percentage change in the price of the stock. This vital number quantifies the level of increases or decreases in an individual stock's value over a day and, consequently, gives useful indications as to its performance against different stocks. It is determined by deducting the previous session's closing price from the present session's closing price and dividing this result into a value expressed in fractions:

**(price[t] - price[t-1]) / price[t-1]**

**1.3.3.5 Macroeconomic Factors:**

Macroeconomic indicators are essential for comprehensively assessing a nation's economic well-being and performance. They offer valuable insights into the collective behavior of economic actors on a large scale, allowing analysts and businesses to comprehend the wider trends that shape an economy.

As the stock market becomes increasingly interconnected, the macroeconomic context plays an even more critical role in forecasting stock returns. The impact of one stock's performance on others reinforces the need for careful consideration of influential factors. Therefore, we focus on macroeconomic factors that have economically significant effects on daily returns for the next trading day. Our selection includes crucial indicators, namely the return of the S&P 500 Index, returns of various assets such as gold, crude oil, and Bitcoin, and the volatility index (VIX). These carefully selected features serve as valuable predictors of stock returns, highlighting essential aspects of market dynamics and economic conditions.

* **Market Sentiment and Momentum:** The **S&P 500 Index** return serves as a general market sentiment indicator. Positive returns on the S&P 500 generally indicate a bullish market outlook and momentum. Investors frequently refer to recent S&P 500 trends to make short-term predictions. Forward momentum in the S&P 500 may encourage bullish sentiment among investors, resulting in increased buying activity and positive stock yields.
* **Safe-Haven Status:** **Gold** is frequently thought of as a secure investment during periods of market instability or economic uncertainty, as it is considered a safe-haven asset. When investors witness positive returns on gold, this usually suggests an inclination toward safety among investors and indicates a market environment that is averse to risk. In such scenarios, demand for stocks may decrease as investors shift their focus to safer assets, causing stock returns to be adversely affected.
* **Commodity Price Trends:** Returns on **crude oil** and **Bitcoin** offer valuable insights into economic activity and market speculation. Positive returns in these commodities may indicate economic growth and increased investor confidence, positively impacting the stock market. Conversely, negative returns may signify economic concerns, potentially leading to lower stock returns.
* **Volatility and Risk Perception:** The **Volatility Index (VIX)** measures market volatility and is often considered an indicator of risk perception. A rise in the VIX suggests increased market volatility and risk perception, which can create uncertainty and potentially lead to lower stock returns as investors become more risk-averse. This hesitancy to make large investments among traders can impact market liquidity and stock prices.

**1.3.3.6 Technical Indicators:**

Technical indicators are mathematical calculations applied to historical price, volume, or open interest data to help traders make decisions. Common indicators include moving averages (SMA, EMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), and Relative Daily Prices (RDP). These tools help traders identify trends, momentum, and potential reversal points in the financial markets. Traders often use a combination of indicators for a more comprehensive analysis.

**1.4. Problem Statement and Project Goals:**

The stock market is a significant part of every country. The flow of the stock market influences the virtual economy crucially. The stock price shows the expectations of the investors towards the market. Many researchers have done a lot of work to try to predict the stock price and stock market trend with different dimensions [1-3].

Though some of those researches shows nice predictions of stock price, only using last day’s stock price as the prediction also gives a nice result. This phenomenon is caused by the small price difference between two neighbor days. As an individual investor, the favorite question is how much money will be earned daily from each dollar invested. However, forecasting the daily return should be much harder than predicting stock price. The daily returns are always represented by percentages and depend on various factors, including numerical variables (close price, volume, turnover, etc.) and qualitative features such as policy and financial news. Predicting stock returns remains to be one of the most demanding tasks in the machine learning field.[Stock Daily Return Prediction of Amazon and Alibaba using Linear Regression and LSTM]

As previously noted, there has been extensive research focused on forecasting tomorrow's closing prices. However, our current research is focusing on predicting daily returns. There are compelling reasons for this strategic shift.

* **Artificial Variability:** Closing prices may be subject to external factors such as market sentiment and speculation, which can introduce fluctuations and noise.
* **Challenges in Prediction:** Predicting closing prices can be challenging due to the non-normalized and artificial variability inherent in financial datasets.
* **Investor Focus:** Investors prioritize understanding potential gains or losses over predicting specific closing prices.
* **Practical Insight:** Predicting daily returns offers a clearer view of relative performance and assists investors in evaluating risk for informed decision-making.
* **Actionable Decisions:** Investors seek guidance on actionable decisions for the next day, including whether to buy, sell, or hold assets based on the projected daily returns.
* **Informed Choices:** By focusing on daily returns, investors can make well-informed decisions about their investment actions, considering potential outcomes and maximizing their understanding of market dynamics.

The graduation project, situated in a specific context, aims to develop a comprehensive understanding of daily return forecasting. A pursuit prompted by practical implications, the necessity of efficient risk management tactics, and an acknowledgment of the dynamic nature of the market. The definitive aim is to furnish investors with practical and relevant findings.

To accomplish this goal, a regression model using machine learning will be created. This model has been meticulously developed to meet rigorous statistical criteria, as well as align with key investment metrics. Its ultimate goal is to provide a reliable forecasting tool that can effectively navigate the intricacies of financial markets, thus offering valuable insights for informed investment decision-making. The tool will address the complexities of financial markets, helping investors make sound decisions based on detailed analysis.

**1.5 Literature review**

The prediction of daily stock returns is a pivotal aspect of financial market analysis, garnering considerable attention due to its potential impact on investment decisions and financial strategies. Several studies employed diverse methodologies to forecast daily stock returns. These studies leverage various data sources and techniques, including the integration of online sources such as Wikipedia and Google News, financial ratios, and macroeconomic indicators among others. By examining these approaches, we aim to gain insights into the effectiveness of various prediction models and their implications for the financial industry.

In [Stock Daily Return Prediction of Amazon and Alibaba using Linear Regression and LSTM] , the study examines the forecasting of the two heads of e-commerce ,Amazon and Alibaba, using a combination of Linear Regression (LR) and Long Short-Term Memory (LSTM) models. The input factors for the study comprise the date and closing price, targeting the forecast of the following day's daily return. The LR model incorporates the closing prices of the last 10 days, and applies cross-validation for dimensionality reduction. The results are as follows: For Alibaba, the Root Mean Squared Error (RMSE) is **0.027** and the R² is **-0.559**. For Amazon, the RMSE is **0.024** and the R² is -**0.457**. The LSTM model, which is a more complex approach, yields the following results: For Alibaba, the root mean square error (RMSE) is **0.031**, with a coefficient of determination (R²) of **-0.651**. Similarly, for Amazon, the RMSE is **0.032**, with an R² of **-1.031**.

The study yielded unexpected results, demonstrating that the LR model surpasses the LSTM model in both RMSE and R² measures for both Alibaba and Amazon. These findings emphasize the importance of model selection and the difficulties in capturing the nonlinearity and instability of stock returns. The research yields valuable insights into the intricacies of forecasting stock returns and emphasizes the necessity for further exploration to refine the precision and practicality of predictive models in financial markets.

In **[Forecasting Daily Stock Market Return with Multiple Linear Regression [2020]**] The author focused on a five-year dataset comprising eleven crucial features, including the SPY return, T-bill rates, and economic indicators. They adeptly handled missing data through cubic spline interpolation and managed outliers using the clip method. The study's data analysis involved dividing the dataset into training and testing sets and applying various model selection techniques such as Forward Selection, Backward Selection, and Bidirectional Selection. Ultimately, it was determined that the forward selection model exhibited the smallest mean squared error of **4.29654**, signifying an optimal model choice. This paper underscores the paramount importance of feature selection and effectively demonstrates the prowess of multiple linear regression in the realm of stock market forecasting.

In the context of stock market prediction, R. Seethalakshmi's[**Analysis of Stock Market Predictor Variables Using Linear Regression (R. Seethalakshmi**] research stands out for its focus on data preprocessing techniques. Notably, it addresses the essential task of converting date information into a format suitable for feature analysis, a critical step in analyzing temporal financial data. Leveraging statistical methods in the R environment, this study examines the performance measures using the S&P 500 Index and demonstrates its superiority over existing approaches. By emphasizing the use of linear regression, it aims to identify the key independent variables that have the most significant impact on predicting closing stock market prices. This research offers valuable insights into the factors crucial for accurate stock market forecasting.

The study[**Stock Daily Return Prediction Using Expanded Features and Feature Selection**] aimed to enhance stock daily return prediction by expanding the feature set to include dollar-gold features (DG), BIST100 stock features (BIST100), and technical indicators. To address the challenging and noisy nature of stock market prediction, the authors considered not only the stock to be predicted but also other stocks and currencies. They computed twenty-five different indicators on daily stock prices, creating feature vectors for each trading day. These feature vectors were labeled based on daily close returns. Feature selection methods, specifically gain ratio and relief, were applied to select the most informative features. Experimental results demonstrated that using gain ratio feature selection with a gradient boosting machine (GBM) achieved an accuracy of **0.599** for GARAN stock, while relief feature selection with GBM yielded an accuracy of **0.558** for THYAO and gain ratio feature selection with logistic regression resulted in an accuracy of 0.581 for ISCTR. Overall, the inclusion of BIST100 stock features enhanced classification accuracy for all stocks, highlighting the effectiveness of feature expansion and selection in stock return prediction.

In[**Macroeconomic Attention and Stock Market Return Predictability**], the authors evaluated the predictive capabilities of novel macroeconomic attention indices (MAI) introduced by Fisher et al. (2021) concerning stock market returns. They employed shrinkage methods and dimension reduction techniques to assess MAI's effectiveness. The findings demonstrated that MAI had a strong predictive ability for stock market returns. Components derived from MAI, particularly through methods like partial least squares (PLS) and the least absolute shrinkage and selection operator (LASSO), showcased superior potential in enhancing return prediction accuracy compared to traditional macroeconomic variables. Moreover, shrinkage methods proved to be the most effective in forecasting stock market returns. The research highlighted MAI's robust predictive performance, especially during the COVID-19 pandemic and over extended periods. This study provides valuable insights into stock market return prediction, emphasizing the significance of MAI and the effectiveness of shrinkage methods.

In [**Machine Learning Algorithms for Predicting the Stock Market Daily Returns**], the authors delved into the domain of stock market daily returns prediction using machine learning algorithms. The research collected data spanning from November 2008 to November 2019, encompassing 1939 records for training purposes. The data included open, high, low, close prices, and close prices for the next day (t+1), which served as the output.

To prepare the data for analysis, a Min-Max normalization technique was applied. The study then employed three prediction algorithms: Support Vector Regression (SVR), Auto Regression and Integrated Moving Average (ARIMA), and traditional Artificial Neural Network (ANN).

The findings highlighted the power of machine learning in forecasting daily returns, with a particular focus on the SPDR S&P 500 ETF. The results demonstrated that ANN outperformed SVR and ARIMA in terms of classification accuracy, emphasizing the effectiveness of neural network-based approaches in stock market prediction.

This paper[**Is the Distribution of Stock Returns Predictable?**] investigates the predictability of different quantiles of stock returns using quantile regression. It explores the influence of economic state variables on various parts of the return distribution, such as lower and upper quantiles. While predicting the central part of the distribution remains challenging, the study demonstrates that economic state variables can predict upper quantiles. Additionally, it shows the economic benefits of using time-varying quantile forecasts for portfolio selection and options trading.

The “Stock Return Prediction study employs Voting Regressor Ensemble Learning” to predict profits from share trading through investment returns calculation. Reliable algorithm development is the main focus of the study, using steps such as data preprocessing, exploratory data analysis (EDA), and modeling in Python. The study found that the Voting Regressor was highly effective, with a low error rate of **0.032523** as measured by Root Mean Square Error (RMSE). Its potential for automated predictions of stock return values in future analyses adds significant value to this research.

The paper “**Stock Prediction with Random Forests and Long Short-term Memory**” explores the application of machine learning and deep learning techniques, specifically Long Short-Term Memory (LSTM) and Random Forests (RF), in predicting stock market returns. By leveraging real historical data, the study examines multiple approaches, including ensemble learning methods and recurrent neural network architectures. The research sheds light on the potential of these technologies for forecasting future stock returns.

“**Predicting Stock Prices Using the Random Forest: In the realm of stock market prediction**”, this study delves into the use of regression models to predict stock price movements. It distinguishes between regression models and rating models, challenging the prevailing trend of employing classification models. By focusing on accurately predicting stock prices using regression models, the paper contributes valuable insights into the algorithms employed for this purpose.

**“Fundamental Quantitative Investment Research based on Machine Learning”**: Despite the simplicity of traditional linear models, this study emphasizes their effectiveness in quantitative investment research. The traditional linear regression model is hailed as a comprehensive tool that can autonomously identify accurate relationships between predicted outcomes and influencing factors. The research underscores the enduring relevance of these models in the landscape of quantitative investment.

In the context of stock market return prediction, this extended literature review provides a deeper understanding of each paper's objectives, methods, and results. The review covers a broad range of techniques and approaches, from traditional regression models to advanced Machine Learning and Deep Learning techniques, providing a comprehensive overview of the research landscape in the field.

The results also highlight the importance of selecting features, reducing dimensionality, and adapting models to specific financial contexts. As we embark on our research in daily stock return prediction, these studies provide valuable reference points and methodologies to consider, ensuring a comprehensive and informed approach to the task.

**Conclusion**

In concluding this chapter, our objective is to furnish our readers with a comprehensive understanding of the overarching framework of this graduation project. We provided an overview of the general context, delved into the stock market's environment, and explained the financial data that is integral to our analysis. Additionally, we conducted a concise analysis of the Machine Learning principles underpinning this project, laying the groundwork for a more in-depth discussion in the following chapters.

Moving forward, our goal is to explore the intricacies of our methodology and analysis in the next two chapters, providing a detailed examination of the machine learning models employed and their application to stock market forecasting.

Additionally, we highlighted the importance of conducting a literature review as a critical step in developing a comprehensive understanding of our project. The reviewed articles have clarified the complex world of stocks, revealing how macroeconomic factors affect stock behavior, highlighting the central role of stock returns in informing investor decisions and emphasizing the indispensable contribution of machine learning models in predicting stock returns. This knowledge gathered from the literature review lays the groundwork for the next stages of our research.

**Chapter 2**

**Mathematical Background**

**1 . Introduction**

In this chapter, we concentrate on the mathematical background needed to achieve our specified objectives. First, we present the Regression Models that we’re going to use.Then, we focus on the evaluation methods that we’re going to use to evaluate the performance of our models. Finally presenting the steps to follow for the data pre-processing, going through Cross Validation, and finishing withfeature selection.

**2. Regression Models:**

A regression model is a statistical tool utilized for examining the correlation between one or more independent variables and a dependent variable. The primary aim is to comprehend how variations in the independent variables are linked to alterations in the dependent variable. In other words, its purpose is to establish a mathematical relationship. This relationship can be used for prediction or inference.

Stock return prediction is carried out using a regression algorithm because the regression algorithm can produce output in the form of a numeric value. Machine learning algorithms that can be used in the process of determining stock return values are machine learning algorithms,[Stock Return Prediction Using Voting Regressor Ensemble Learning ].

In our project, we will choose three regression models for our predictions: Linear Regression, Decision Tree, and Random Forest. Ultimately, the model with the most favorable results in both statistical and investment evaluations will be selected.

**2.1 Linear Regression:**

Linear regression is a fundamental model of a machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. Despite its simplicity, still has a great deal of power and can produce significant results in a wide variety of cases. Additionally, it boasts the advantage of producing easily interpretable results.

In our work, we’re going to use the Ordinary least-squares model. The OLS model is based on linear regression. After inputting the dataset, the model will generate data points based on all the variables. Then, it draws a best-fit line that minimizes the residuals (distance between the data points and the best-fit line). The best-fit line gives an equation for prediction. [Bitcoin Return Prediction based on OLS, Random Forest, LightGBM, and LSTM ]

The model equation extends to :

y = β0x0 + β1x1 + β2x2 + · · · + βnxn + e = β T x + e

Where:

• y is the dependent variable.

• x = [x0, x1, x2, . . . , xn] are the independent variables or called also the predictors or features. Generally, x0 is a constant variable that is filled by one.

• β = [β0, β1, β2, . . . , βn] are the regression coefficients to be estimated.

• e is the error term.

Therefore:

yˆ = βˆT x

where βˆT is the estimated regression coefficient.

The OLS method estimates the values of the regression coefficients. Minimize the sum of squared residuals (the vertical distances between the observed and predicted values) to calculate the optimal line of best fit through the data points.

**2.2 Decision Tree**

A decision tree is a powerful and versatile supervised machine-learning algorithm utilized for both classification and regression tasks. This algorithm operates by recursively partitioning the dataset into subsets based on the values of input features, ultimately forming a tree-like structure. At each step, decisions are made, leading to the creation of branches and leaves in the tree.

* **Decision Tree Regression:**

Decision tree regression is a tree-based structure used to predict numerical outcomes for a dependent variable **[Rathore, S. S., Kumar, S. A Decision Tree Regression-based Approach for the Number of Software Faults Prediction. ACM SIGSOFT Software Engineering Notes, 2016, 41(1).]**. Decision tree regression was chosen for this study because, unlike traditional decision trees, it predicts numerical outcomes for the dependent variable.**[Compare Linear regression, Decision Tree Regressor, and Random Forest Regressor based on Python, a restaurant company on Kaggle as a case]**

* **Parameters of Decision Tree Regression:**

When using a decision tree regression model, it's important to fine-tune its parameters to enhance its effectiveness. Several key parameters can influence the behavior and performance of the model:

**Criterion**: The criterion parameter determines the function to measure the quality of a split. "MSE" (Mean Squared Error) is commonly used for regression tasks. It measures the variance reduction in the target variable.

**Splitter**: The splitter parameter dictates the strategy used to choose the split at each node. "Best" selects the best split, and "Random" selects the best random split. "Best" is typically preferred for regression tasks.

**Max Depth**: Max depth restricts the maximum depth of the decision tree. It controls the size of the tree and helps prevent overfitting. A deeper tree may capture more details in the training data but is prone to overfitting.

**Min Samples Split:** Min samples split sets the minimum number of samples required to split an internal node. It prevents the tree from making splits that only capture noise in the data.

**Min Samples Leaf:** Min samples leaf sets the minimum number of samples required at a leaf node. It prevents the creation of leaves with very few samples, which can lead to overfitting.

**Max Features:** Max features determine the maximum number of features considered for a split. It helps control the diversity of the trees in an ensemble and can be crucial in reducing overfitting.

**Random State:** Random state ensures reproducibility by fixing the random number generator's seed. Setting a random state allows for consistent results when re-running the model.

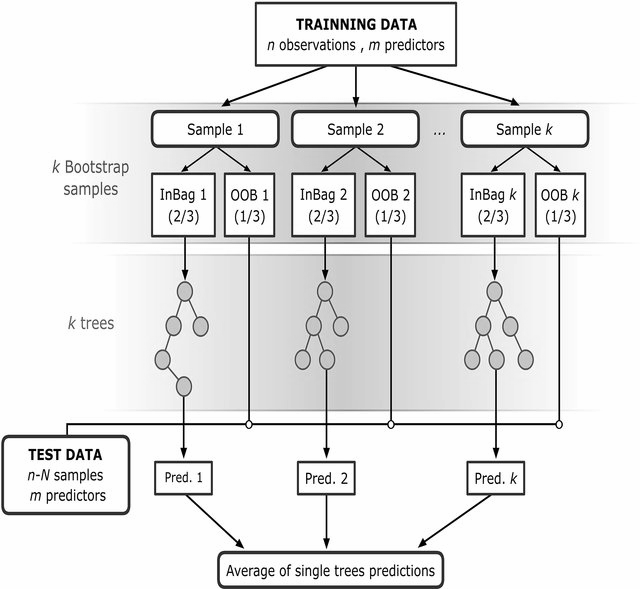
**2.3 Random Forest**

Random Forests is a machine-learning algorithm that uses a collection of decision trees. This algorithm combines the bagging method and the random subspace method. It is usually used for classification, regression, and clustering [11]. The main idea is to use a large collection of decision trees.In our project we’re going to use the Random Forest Regressor

The Random Forest Regressor (RFR) is a Machine Learning algorithm that uses ensemble learning techniques, namely combining several machine learning models into one more powerful model. In the Random Forest Regressor algorithm, several decision trees that are formed randomly will be combined to produce a prediction [L. Breiman, “A Data Mining Based System for Transaction Fraud Detection,” 2021 IEEE Int. Conf. Consum. electrons. Comput. Eng. ICCECE 2021 , pp. 542–545, 2021, doi: 10.1109/ICCECE51280.2021.9342376]. The Random Forest Regressor takes a random sample from the dataset for each created tree and uses the sample to construct a decision tree. Each decision tree that is formed will provide predictions independently and the results will be combined to produce more accurate predictions [A. Cutler, DR Cutler, and JR Stevens, “Random forests,” Ensemble Mach. Learn. Methods Appl. , pp. 157–175, 2012].

Random forest regressor (RFR) has the advantage of generating many decisions based on a decision tree algorithm that can affect better accuracy and making, but this algorithm requires hyperparameter selection for better performance and makes the model overfitting if the model is too complex [L. Breiman, “A Data Mining Based System for Transaction Fraud Detection,” 2021 IEEE Int. Conf. Consum. electrons. Comput. Eng. ICCECE 2021 , pp. 542–545, 2021, doi: 10.1109/ICCECE51280.2021.9342376].

Here's a representation of how a random forest for regression works:



[Source :[Victor F. Rodriguez-Galiano](https://www.researchgate.net/profile/Victor-Rodriguez-Galiano?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoicHVibGljYXRpb24ifX0) ]

* **Parameters of Random Forest Regressor:**

It's often beneficial to start with a reasonable set of values for these key parameters and to avoid overfitting in our mode The most important parameters to consider when tuning a Random Forest Regressor are:

**1. n\_estimators:** Number of trees in the forest. Increasing this value generally improves performance, but there are diminishing returns beyond a certain point.

**2. max\_depth:** Maximum depth of the trees. Controls the complexity of individual trees. A higher value allows for capturing more complex relationships but increases the risk of overfitting. We tried to limit our tree depth.

**3. min\_samples\_split:** Minimum number of samples required to split an internal node. A higher value helps prevent the creation of small nodes capturing noise, and controlling tree growth.

**4. Min\_samples\_leaf:** Minimum number of samples required to be at a leaf node. Controls the size of terminal nodes and helps prevent overfitting, we increased the number to ensure larger, more generalized lead nodes.

**5. max\_features:** Number of features to consider when looking for the best split. A lower value can reduce overfitting by introducing diversity among trees.

**2.4 Comparison between Models**

In the realm of regression analysis, choosing the right model is pivotal for accurate predictions. This comparative study delves into three major regression techniques: RFR, DTR, and LR. We'll explore their characteristics, considering factors like overfitting, predictive accuracy, robustness, interpretability, and computational efficiency.

| **Aspect** | **RFR** | **DTR** | **LR** |
| --- | --- | --- | --- |
| **Nature** | Ensemble of multiple decision trees | Single decision tree | Linear model |
| **Bias-Variance Trade-off** | Lower variance, reduced overfitting | Higher variance, prone to overfitting | Balancing act, controlled by regularization parameters |
| **Predictive Accuracy** | Generally higher due to ensemble | Prone to overfitting, may vary | Moderate, depends on linearity and feature independence |
| **Robustness** | More robust to outliers and noise | Sensitive to outliers and noise | Sensitive to outliers, may require data preprocessing |
| **Training Time** | Slower due to multiple tree construction | Faster as it builds a single tree | Faster compared to tree-based models |
| **Interpretability** | Less interpretable due to ensemble | More interpretable as a single tree | More interpretable, provides explicit coefficients |

This analysis seeks to enhance our comprehension of the distinctions among these three regression models. It aims to provide deeper insights into the results generated by each model concerning the prediction of the next day's return. The ensuing interpretation of these results will be expounded upon in the next chapter.

**3. Model Performance Evaluation**

**3.1 Statistical Evaluation**

Statistical evaluation entails quantitatively assessing models, analyses, or hypotheses through a range of statistical metrics. Its objective is to ascertain the degree of fit between the model or analysis and the data and to measure the dependability of the relationships or patterns identified.

Common statistical evaluation metrics include:

**3.1.1 The Coefficient of Determination R²**

The R² metric functions as a statistical indicator that defines the percentage of variance in a dependent variable, clarified by one or more independent variables within a regression model. Also known as the coefficient of determination, or for multiple regression, the coefficient of multiple determination, R² is a crucial measure in regression analysis.The following equation presents the formula for calculating the R2 metric: 

where yi , and yˆi are the i th actual and predicted value, respectively, and y¯ shows the mean of actual values.

R² is a statistical measure between 0 and 1.A higher R² is always desirable, indicating greater capacity to elucidate changes in the outcome variable. Assessing the strength of an effect size involves interpreting R² values. A value below 0.3 is commonly considered fragile, whereas a range between 0.3 and 0.5 denotes a low effect size. Conversely, if R² exceeds 0.7, it indicates a robust effect size

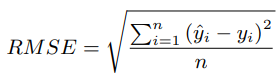
**3.1.2 Explained Variation**

The explained variance is used to measure the discrepancy between a model and actual data. In other words, it’s the part of the model’s total variance that is explained by factors that are actually present and are not due to error variance. The explained variation is the sum of the squared of the differences between each predicted value and the mean of actual values**.[Effectiveness of Artificial Intelligence in Stock Market Prediction Based on Machine Learning ].**The equation below demonstrates how to calculate the EV metric:



**3.1.3 Root Mean Squared Error**

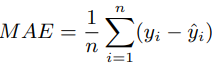
The RMSE is a performance indicator for a regression model. It measures the average difference between the predicted and actual values. It provides an estimation of how well the model can predict the target value. In other definitions, it shows how concentrated the data is around the line of best fit , this explains why the smaller value of this metric represents a better prediction of the model.Here’s its equation:



where n is the number of experiments, yˆi is the predicted value, and yi is the actual value for the i th experiment.

**3.1.4 Mean Absolute Error**

The MAE metric calculates the average magnitude of errors between the predicted and actual values in a regression model. It is the sum of the absolute differences between the predicted and actual variables, and similar to the (RMSE), a smaller MAE value indicates a better prediction model. The mathematical formula for MAE is presented below.

****

**3.2 Investment Evaluation**

In the field of investment assessment, the primary aim is to evaluate the practical utility of a given model or analysis, particularly in the context of making informed investment decisions. Vital metrics utilized in this assessment encompass important measures like the **Sharpe Ratio**, **Cumulative Sum**, **Maximum Drawdown**, and **Direction Agreement Percentage**. The main goal of this assessment is to measure the effectiveness of the model in producing lucrative investment decisions whilst managing risk effectively. In essence, investment evaluation is critical in identifying the practicality and efficiency of financial models in guiding strategic investment decisions.

**3.2.1. Sharpe Ratio**

* **The Sharpe ratio or Risk-Adjusted Return** means how well the return of an asset compensates the investor for the risk taken. When comparing two assets against a common benchmark to calculate the risk-adjusted return, the one with a higher Sharpe ratio provides a better return for the same risk, it can only be accurate if the data has a normal distribution.

The Sharpe ratio is calculated by dividing the mean daily return by the standard deviation of the daily returns. The higher the Sharpe ratio, the higher the return, and the lower the standard deviation or volatility, the better the trading strategy.[Paper : Forecasting daily stock market return using dimensionality reduction]

**=> Sharpe Ratio = (Return - Risk-Free Rate) / Standard Deviation of the Return.**

**3.2.2. Cumulative Sum of Return**

* **The cumulative sum,** also known as the **running sum** or **cumulative total,** is a simple mathematical calculation that shows the sum of values in a sequence as you move through that sequence. It helps track the total accumulation of values over time.

**3.2.3 Percentage of Direction Agreement**

* **This percentage of the Direction Agreement** helps us assess whether the actions predicted by the model align with the actual actions to be taken. This percentage provides more insight into the accuracy of predicting the direction of the "return" rather than the actual return value.

This metric allows us to evaluate the correctness of the direction prediction, helping us understand if the model is effectively forecasting the direction of the "return."

**4. Model Optimization**

* **Cross Validation**

Cross-validation is a resampling method utilized in machine learning to evaluate the performance of a predictive model. Cross-validation can be used in many scenarios and for several objectives. It should be used when we are dealing with small data sets and also when we are building a statistical model with one or more unknown parameters. So, the idea is to optimize those parameters in order to get a model that matches the data as soon as possible.**[An Introduction to Statistical Learning]**

CV divides the entire dataset into two sets: a training set and a testing set. Each observation in the complete dataset belongs to one, and only one. This is done to prevent leakage from one set to another, which would defeat the purpose of testing invisible data. There are many alternative CV methods, of which one of the most popular is the k-fold CV. [7] This method follows the following process:

1. The dataset is divided into k subsets.

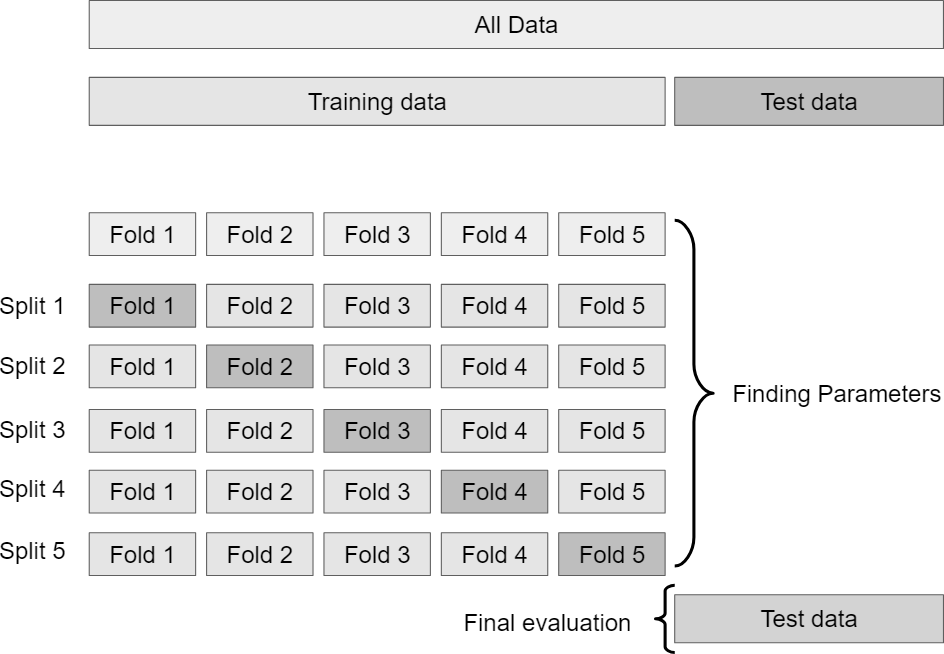
2. For i ∈ {1, . . . , k} :

(a) The model is trained on all the subsets excluding the i th subset.

(b) The fitted model is tested on i by evaluating the model performance metrics.

3. Compute the average of each model performance metric over the k folds. We add CV as a prefix for these metrics. For instance, CV R² indicates the average of the R² metric

Here’s an example of the splitting data with k =5:

****

* **Grid Search**

Grid search is a technique utilized in machine learning to optimize hyperparameters.

There are two methods commonly used for hyperparameter tuning in ML:

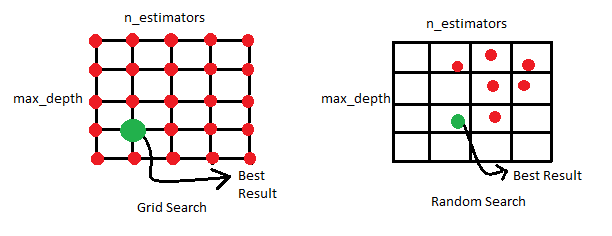
* **Grid Search**

It involves defining a grid of hyperparameter values, creating a matrix of all possible combinations, and evaluating the performance of the model for each set of hyperparameters using cross-validation. The exhaustive search ensures that no combination is overlooked. Grid search is advantageous for small hyperparameter spaces but may become computationally expensive as the number of hyperparameters or values increases.

* **Randomized Grid Search**

Randomized search adopts a more stochastic method for hyperparameter tuning. It randomly evaluates a specified number of hyperparameter combinations from predefined distributions using cross-validation to assess each set's performance. Randomized search offers a significant advantage when the hyperparameter space is expansive since it is more computationally efficient than grid search. While random sampling may not explore every possible combination, it can be useful, especially in cases where specific hyperparameters critically impact model performance.

Here’s a picture explaining these two methods:

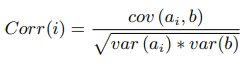
****

**5. Feature Selection**

One of the most crucial steps in the ML project is the Feature Selection or Variable Elimination. Feature Selection helps in understanding data, reducing computation requirements, reducing the effect of the curse of dimensionality, and improving the predictor performance. In our study, we’re going to present two feature selection methods: **Correlation** and **P\_value**.

**5.1 Correlation**

Correlation is a measure of the linear relationship between two or more parameters. In this method, features showing the most correlation with the target are selected to build the model. Furthermore, to avoid redundant computation, the selected features should not be highly correlated to each other.[**Effectiveness of Artificial Intelligence in Stock Market Prediction Based on Machine Learning**].To accomplish this, one of the most effective methods is the use of the Pearson correlation technique, which is outlined below:



**5.2 P\_value**

The p-value, which is derived from statistical hypothesis testing, assists in gauging the significance of the association between individual features and the target variable. When selecting features, a lower p-value indicates stronger evidence against the null hypothesis, implying a more substantial feature. This approach proves extremely advantageous in linear regression models whereby each feature's contribution can be assessed objectively for significance. It permits a focused selection of features that significantly contribute to the model's predictive capability.

This measure is a value between 0 and 1. It can be interpreted as follows:

* p-value ≤ α: It indicates that the feature is statistically significant. In this case, we reject the null hypothesis.
* p-value > α: It indicates that the feature is not statistically significant. In this case, we fail to reject the null hypothesis.

In this project, we chose α = 0.1.

**6. Conclusion**

In this chapter, we cover the mathematical background used in our project. The focal objective is to convey fundamental mathematical concepts to enable comprehension of our findings in subsequent chapters. We begin by explaining the principles behind machine learning regression models and describe the specific models that are essential to our project. We then introduce the two facets of evaluation: Statistical Evaluation and Investment Evaluation. After that, we explore model optimization techniques to identify the most effective model. Finally, we introduce the crucial aspect of feature selection and outline the techniques to be used in our project.

**Chapter 3**

**Modeling and Empirical Results**

1. **Introduction**

This chapter describes our work on the daily return predictions for the 500 symbols of our data. First, we present the work environment of this project. Then we focus on the methodology followed during the internship. After that, we turn our focus to the exploratory data analysis, feature selection, and the feature engineering. Finally, we tackle the modelling phase by including the performance evaluation of the fitted models.

1. **Software Environment**

In our project, we chose Python as the programming language that we’re going to use. Python is an Open Source Programming Language, that has become a dominant force in the field of Data Science due to its rich ecosystem of libraries, simplicity, and versatility.

Here is a list of packages that we used in this project:

* **Pandas:** Pandas is a versatile data manipulation and analysis library commonly used for cleaning, exploring, and transforming structured data using DataFrames.
* **Numpy:** NumPy is a fundamental numerical computing library, essential for handling large arrays and matrices and performing mathematical operations.
* **Datetime:** Datetime, part of the standard Python library, provides easy manipulation and arithmetic operations on dates and times.
* **Statsmodels:** Statsmodels is a statistical modeling library that provides a set of models for regression, time series analysis, and hypothesis testing.
* **Matplotlib:** Matplotlib is a comprehensive plotting library for creating a variety of static, animated, and interactive visualizations.
* **Seaborn:** Seaborn, based on Matplotlib, simplifies the creation of aesthetically pleasing statistical graphics, particularly useful for complex datasets.
* **Plotly:** Plotly is an interactive graphing library for dynamic and interactive plots, often used for web-based dashboards.
* **Scikit-learn:** is a versatile machine learning library that provides simple tools for data analysis and modeling, including various algorithms. It’s therefore very useful for statistical modeling.
* **Scipy :** Scipy is an essential scientific computing library for many scientific and engineering applications, SciPy extends NumPy by providing tools for optimization, statistics, signal processing, and other applications.
* **yfinance:** yfinance is a Python interface to the Yahoo Finance API, allowing users to retrieve historical market data and stock quotes for financial analysis.

1. **Methodology Overview**

In response to the evolving challenges of predicting stock movements in today's dynamic market, our project delves into the intricate realm of stock return forecasting, with a particular focus on the non-linear nature and pronounced fluctuations characteristic of stocks. Grounded in extensive research, particularly drawing insights from notable papers exploring stock return dynamics, our objective centers on devising a predictive model for the returns of symbols within the influential S&P 500 index.

Inspired by research such as "Predicting Stock Prices Using the Random Forest Classifier," "Macroeconomic Attention and Stock Market Return Predictability," and "Stock Daily Return Prediction of Amazon and Alibaba using Linear Regression and LSTM" by Jinxian Lyu, among others, we conducted a thorough investigation. After analyzing different methodologies, we selected a strategic approach that predicts stock returns for the next day. This method combines macroeconomic features and technical indicators to generate reliable forecasts. This choice is supported by the economic importance of these factors, as described in the introductory section.

The following section of our methodology presents a trading strategy intended to facilitate investor decision-making. We aim to optimize actions for the upcoming trading day by aligning our approach with proven strategies.

Following this, we carefully preprocess the gathered data to ensure its quality and relevance for subsequent analysis.

The essence of our methodology pertains to implementing three machine-learning regression models. We aim to identify, through rigorous testing and evaluation, the model that best fits our data and delivers optimal results in terms of statistical accuracy and investment performance. Our second chapter elaborates extensively on this meticulous selection process and the statistical and investment evaluation criteria that guide our decisions.

Our methodology is fundamentally a carefully crafted fusion of insights from the existing literature, a targeted selection of features with economic and strategic significance, and the application of advanced machine learning models. This approach, underscored by a robust trading strategy, positions us to make informed predictions regarding the stock returns of S&P 500 index symbols for the upcoming trading day.

1. **Data Description**

In our research project, we solely utilize publicly available data obtained from Yahoo Finance. The dataset originates from the S&P 500 index, providing a thorough compilation of symbols representing index constituents. Each symbol is linked to crucial attributes, which provide insights into the market behavior of the corresponding assets. These attributes comprise Open, High, Low, and Close prices, in addition to Volume and Adjusted Close, covering the period from 2016 through 2023.

Furthermore, in addition to the historical data for each symbol, we gather Macroeconomic Factors data from Yahoo Finance, as discussed in the preceding chapter. This supplementary dataset encompasses the same period as our utilized historical data.

To evaluate the performance of our three models, we follow a clearly defined timeline for training, testing, and validation. Accordingly, we structure our data as follows:

* **Training Set:** From 2016 to 2021.
* **Validation Set:** 2022.
* **Test Set:** 2023.

1. **EDA**

Exploratory Data Analysis (EDA) is a critical stage in the data analysis process, entailing the initial examination and visualization of a dataset to reveal patterns, trends, relationships, and potential outliers within the data. EDA serves as a foundational element in comprehending the nature of the dataset, directing the subsequent analyses, and aiding the selection of appropriate modeling techniques.

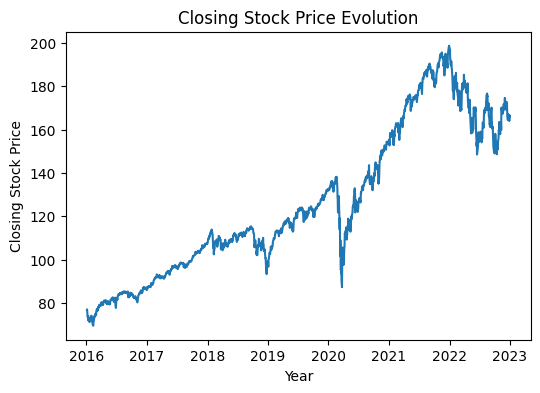
* 1. **General Descriptives Statistiques**
     1. **Missing Values**

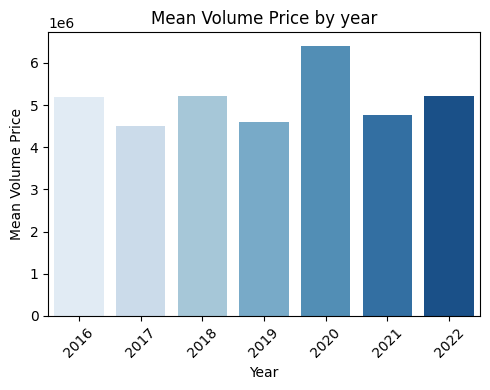
In any data science project, dealing with missing values is a key step that reveals the details of our data and is of utmost significance. Here, accuracy is crucial because even a small error can have big repercussions. In our instance, we carefully examined our data to fill in these gaps. While some missing data might appear to be unimportant when compared to the overall dataset, our domain, which is focused on the exacting world of trading and the stock market, demands absolute precision. Thus, to preserve data purity, which is essential for accurate predictions, we decided to eliminate symbols with any missing data. The number of symbols decreased from 503 to 498 as a result of this careful selection, but the dataset's integrity was strengthened.

Our goal was to create a dataset that not only satisfies the requirements of our subject but also maintains accuracy and precision, which are essential for reliable forecasts in the complex world of stock market analysis.

* + 1. **Data visualization**

Data visualization serves as the most accessible avenue for comprehending complex data, enabling us to extract insights that might otherwise remain concealed. Through compelling visual representations, we can unravel patterns, relationships, and trends, enhancing our grasp of the underlying information.

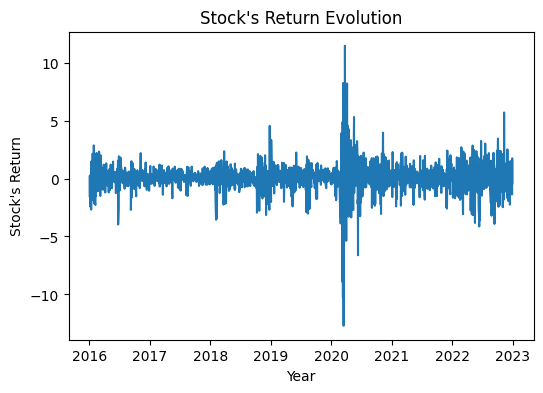




Presented here are two plots illustrating the average closing prices and average volume prices for all symbols annually from 2016 to 2022, with a notable focus on the retail factor during the pivotal year of 2020.

The first graph shows the fluctuation in closing prices, which is characterized by a significant drop in mid-2020, due to the widespread disruptions caused by the COVID-19 pandemic on the stock market. This downturn particularly affected the retail sector, which had to deal with closures, supply chain disruptions, and changes in consumer behavior during pandemic-related lockdowns. However, the subsequent robust recovery, as evidenced by the Mean Closing Prices, underscores the resilience of the market, particularly in the retail sector, beyond 2020. The Mean Closing Price nearly doubled from $75,000 in 2016 to nearly $200,000 in 2022, reflecting the adaptability and growth of the retail industry.

The second chart delves into the mean volume prices over the same period, highlighting a consistent trend with volume primarily hovering between 4.5 and 5 units. Notably, there was a noticeable spike to 6.3 units in 2020, indicating increased trading activity and volatility, particularly in the retail sector. This spike in volume underscores the dynamic response of both institutional and retail investors in the retail industry, as they adjust their strategies to meet the unique challenges posed by the pandemic. The juxtaposition of these charts provides a comprehensive view of the stock market's response from 2016 to 2022, shedding light on the challenges faced by the retail sector and its subsequent recovery and growth in the years that followed.



Concluding our analysis, we present a graph detailing the mean of daily returns over time for all symbols from 2016 to 2022. The graph highlights the percentage changes in daily returns, displaying a consistent trend of moderate fluctuations ranging from -0.03 to 0.03 between 2016 and 2020. It is an illuminating visualization. Nonetheless, a significant shift occurred in 2020, marked by a conspicuous surge in daily returns ranging between -0.1 and 0.1.

This highlights the dynamic nature of the stock market with its unforeseen fluctuations. The events of 2020 exemplify its capriciousness, deviating from conventional standards and showcasing the market's inclination towards uncommon actions. It emphasizes the ineffectiveness of relying on traditional regulations in the stock market, underscoring its natural unpredictability and deviation from the norm.

Delving deeper into the data visualization reveals subtle links among certain characteristics, exposing relationships that may have been obscured in raw data. This underscores the complexity of the stock market's evolution and underscores the need for meticulous analysis to decipher its intricate patterns.

In the following section, we'll dive into correlations, where we'll carefully look at how things are connected. This will help us understand how different pieces of data work together and give us important information.

1. **Trading Strategy**

A trading strategy is a systematic plan used in financial markets to guide decisions about buying and selling assets. Traders use various approaches, including technical and fundamental analysis, to analyze market conditions. Key elements of a strategy include entry and exit points, risk management, and position sizing. Successful strategies require discipline, adaptability, and continuous refinement based on market changes. The objective is to attain steady profits while effectively mitigating risks.

In our project, we have implemented a trading strategy to help investors make informed decisions for the upcoming day. This strategy will help us to calculate investment metrics, as outlined in Chapter 2,we aim to continuously improve these metrics. The strategy is as follows :

* **Understanding Model Actions:** To assist investors in decision-making, it is crucial to comprehend the actions of the machine learning model. We assess the effectiveness of investment decisions based on our model's predictions. Before evaluating the model's performance, we establish a predefined threshold of 0.1 to determine the optimal course of action.
* **Decision Criteria:**
  + - **If** R\*(t+1) >= 0.1 ⇒ In that case, execute a “Buy” trade today and a “Sell” trade tomorrow.
    - **If** R\*(t+1) < 0.1 ⇒ Execute a “Sell” trade today and a “Buy” trade tomorrow.
    - **Otherwise,** no action is taken.
* **Validation and Comparison**: After making decisions, we compare our actions with the actual actions to see how closely our model's predictions align with real-world market actions.
* **Calculation of Cumulative Return:** To evaluate our model’s performance, we calculate the cumulative sum of returns using the returns that we’ve calculated after choosing the actions that we’ve got with our trading strategy.
* **The Percentage of the Direction Agreement:** We can also calculate this metric too, it quantifies the overall percentage of agreement between the predicted actions and the actual market actions.It provides insights into the reliability and accuracy of our model in predicting the market's directional movements.

Through the systematic application of these steps, our goal is to refine the model's predictions, optimize our investment decisions, and improve the overall performance of our trading strategy.

1. **Feature Engineering**

Feature engineering is the process of converting unprocessed data into a format that boosts the efficacy of machine learning models. This entails selecting, creating, or altering features to enhance the model's ability to recognize patterns, generate precise predictions, and achieve superior overall performance.

In our project, we decided to delete the symbols that don’t have data information in the period chosen to be analyzed. Then we decided to add the Technical indicators and the Macroeconomic factors that we introduced in the first chapter.

As previously mentioned, our dataset consists of time series data, and the integration of time-related features significantly enhances the model's ability to capture relevant market dynamics, leading to more insightful predictions of future returns. To leverage this, we chose to integrate Relative Daily Prices (RDP\_5, RDP\_10, RDP\_20) into our model, as they proficiently capture trends across diverse time frames. Positive trends detected in short-term RDPs (RDP\_5) suggest a possible momentum in stock returns for the following trading day. Additionally, favorable trends in intermediate (RDP\_10) and long-term (RDP\_20) relative prices may indicate sustained momentum, which could potentially impact stock returns over an extended period. This strategic integration of time-based characteristics, particularly RDPs, enhances the model with a nuanced comprehension of temporal patterns, thereby improving its predictive abilities.

The equation below demonstrates how to calculate the RDP fo every symbol:

**RDP\_x =[** **( P(t) - P(t-x) ) / P(t-x) ] ×100**

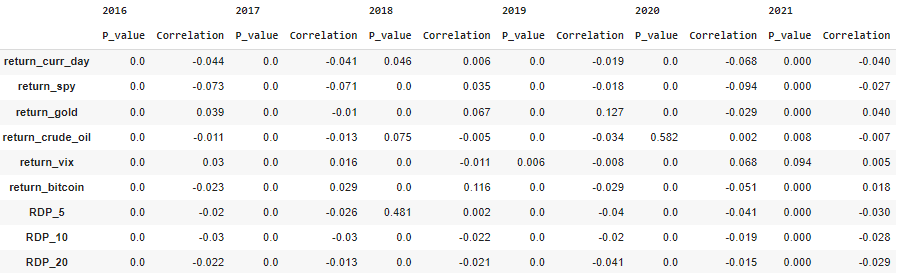
Where **RDP\_x** : The x-day relative difference in the percentage of the market

1. **Feature Selection**

As noted in the preceding chapter, our feature selection methodology analyzes the correlation and p-values of the selected features. Our predictive modeling has strategically incorporated both macroeconomic factors and technical indicators. on their considerable economic influence on stock behavior, as previously discussed.

Now, our objective is to empirically assess the statistical impact of the selected features on our predictions. We are following the specific feature selection methods which we have carefully chosen for this project. This entails a thorough examination of correlation and p-values to determine the statistical significance of our chosen features in the context of predictive modeling. We aim to achieve our goal by refining and optimizing the features that most effectively contribute to our predictive results.

We present the findings of the P-value and correlation analyses for each feature, integrating Macroeconomic Factors and Technical Indicators. These analyses were conducted annually throughout the training set.



Upon examination, it is clear that, in most instances, the p-values linked to our features surpass 0. This pattern demonstrates statistical significance and posits that these features are valuable for integration in our predictive models. Furthermore, after conducting a closer analysis of the correlation coefficients, there is a discernible connection between the 'return\_next\_day' and the features. Nevertheless, the observed correlations have mostly been low in magnitude and negative in value.

1. **Modeling: Results and validation**

For the modeling phase, we selected three types of machine learning algorithms: Linear Regression, Decision Tree, and Random Forest.

To identify the best-fitting models, we used cross-validation and grid search for both the Decision Tree and Random Forest algorithms. The resulting hyperparameters for each model are shown below:

* **For the Decision Tree:**

**max\_depth=4:** This parameter limits the maximum depth of the decision tree to 4 levels. Limiting the depth helps prevent overfitting by controlling the complexity of the tree.

**max\_leaf\_nodes=100:** This sets the maximum number of terminal nodes (leaf nodes) in the tree to 100. This also helps to control the complexity of the tree.

**min\_samples\_leaf=50:** This parameter specifies the minimum number of samples that must be present in a leaf node. In this case, a leaf node must contain at least 50 samples. Setting a minimum number of samples helps to avoid very small leaf nodes that may capture noise in the data.

**random\_state=42:** This parameter ensures reproducibility by setting a fixed random seed (42 in this case). If the same seed is used, the results of the model will be consistent across different runs.

* **For The Random Forest:**

**max\_depth=20:** This parameter limits the maximum depth of each decision tree in the random forest to 20 levels. Controlling the depth helps manage the complexity of each tree and can be effective in preventing overfitting.

**max\_leaf\_nodes=100:** This sets the maximum number of terminal nodes (leaf nodes) in each tree to 100. Similar to controlling depth, limiting the number of leaf nodes helps manage the complexity of individual trees.

**min\_samples\_leaf=5000:** This parameter specifies the minimum number of samples that must be present in a leaf node. In this case, a leaf node must contain at least 5000 samples. This setting helps to create larger and more robust leaf nodes, which reduces the risk of overfitting.

**n\_estimators=4:** This parameter determines the number of decision trees in the random forest ensemble. In this case, there are 4 trees in the ensemble. A higher number of trees generally results in a more robust and stable model.

**random\_state=42:** This parameter ensures reproducibility by setting a fixed random seed (42 in this case). A consistent random seed means that the results of the model will be the same across different runs.

* 1. **General Results**

Here are the results of our three machine learning models: LR, DTR, and RFR. Below is a table detailing both the statistical evaluation and the investment evaluation for each model.

| **Statistical Evaluation** | | | |
| --- | --- | --- | --- |
| **Metrics** | **RFR** | **DTR** | **LR** |
| **R² Training** | 0.075 | 0.087 | 0.017 |
| **R² Test** | -0.07 | -0.084 | -0.038 |
| **R² Validation** | -0.05 | 0.00135 | -0.0144 |
| **EV** | -0.07 | -0.079 | -0.3 |
| **MSE** | 6 | 6.07 | 5.82 |
| **RMSE** | 2.45 | 2.46 | 2.41 |
| **MAE** | 1.77 | 1.74 | 1.76 |
| **Investment Evaluation** | | | |
| **Metrics** | **RFR** | **DTR** | **LR** |
| **Sharpe Ratio** | -0.0022 | -0.03 | -0.0055 |
| **PDA** | 74.33 | 74.674 | 74.3 |
| **Cumulative Sum of Return** | -2.535 | -2.8 | -4.123% |

The statistical evaluation of the three machine learning models RFR, DTR, and LR provides nuanced insights into their performance. When considering the training phase, DTR has the highest R² value of 8.7 %, indicating a relatively better ability to explain the variance in the dependent variable during model training. However, the R² values for the test and validation sets are negative for all models, indicating challenges in generalizing predictions to new and unseen data. Notably, DTR tends to outperform the other models in both the test and validation scenarios and we can see that it’s the only model having a positive R² for the test set with a low value of 0.135%.

When examining explained variance (EV), LR has the lowest value, indicating a potential limitation in capturing variability within the data. In terms of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), LR consistently shows the lowest values, suggesting superior accuracy in predicting the target variable. Meanwhile, the mean absolute error (MAE) is minimal for all models, with DTR slightly outperforming the others.

The comprehensive analysis highlights the strengths and weaknesses of the models in different aspects of their predictive performance. While DTR excels to some extent in training and generalization, LR shows promising results in terms of accuracy metrics. However, the negative R² values and EV underscore the challenges these models face in capturing the underlying patterns within the dataset.

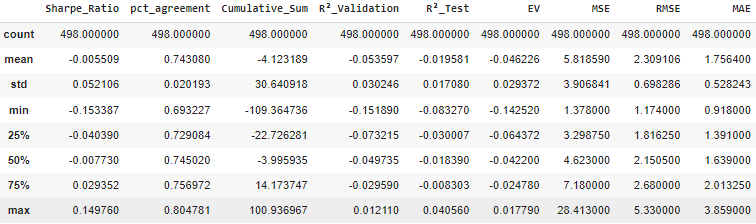
The results obtained did not meet our expectations, revealing poor predictive capabilities across various statistical and investment evaluation metrics. The models struggled to provide accurate predictions when considering all 500 symbols together. However, it is important to recognize the complexity of predicting a diverse set of symbols, each belonging to different price groups-high, medium, and low. With this in mind, a more granular analysis is needed. By breaking down our predictions and evaluating the investment evaluation metrics for each symbol individually, we aim to gain deeper insights into the specific performance of the models across different price categories.

* 1. **Results per symbol**

We will now examine the results of isolating the test set for each symbol individually and calculating the relevant metrics. This method allows for detailed analysis of patterns, potential strengths, and weaknesses on a symbol-by-symbol basis. This granularity is critical to refining our models, fostering a deeper understanding of their performance nuances, and ultimately improving their overall predictive effectiveness.

These are the results when we divided our test set and validation set for each symbols and we’ve got different results

* For the Linear Regression Model:



We can see here that this results can give us a new highlight about our model’s predictions :

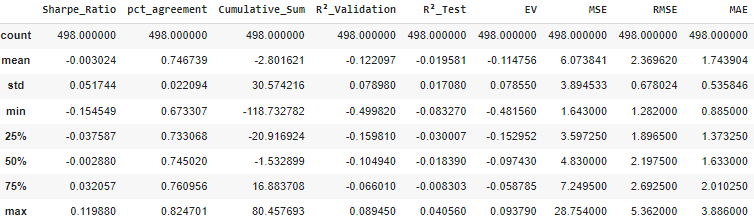
* **Sharpe Ratio:** The minimum Sharpe Ratio of -0.15 reflects a scenario where the model's performance is suboptimal, indicating that in some cases the return per unit of risk is below average. On the other hand, the maximum Sharpe ratio of almost 0.15 suggests that there are instances where the model achieves an acceptable level of excess return relative to risk. The wide range between these extremes underscores the variability in the model's ability to balance return and risk.
* **Percentage Agreement:** The minimum and maximum values for percentage agreement help gauge the consistency of the model's predictions. A minimum value of approximately 69.3% indicates instances where the model's predictions are less consistent with expected outcomes. Conversely, a maximum value of approximately 80.5% indicates instances where the model has a higher level of prediction accuracy.
* **Cumulative Sum:** The minimum Cumulative Sum of approximately -109.36 indicates a significant decrease in cumulative returns, corresponding to periods of pronounced negative performance. The maximum Cumulative Sum of approximately 100.94 indicates instances of substantial positive returns. This wide range highlights the model's susceptibility to significant fluctuations in cumulative performance.
* **R² (validation and test):** The range between the minimum and maximum R² values illustrates the variation in the explanatory power of the model. Notably, the best R² values for the validation and test sets are 1.2% and 4%, respectively. These values represent good results for financial data and for predicting stock market behavior. In stock market prediction scenarios, it is unusual to expect significantly higher R² values. This reinforces the importance of achieving positive R² values, even at relatively modest percentages, as it suggests some level of explanatory power in the model's predictions given the inherent complexity and volatility of stock market data.
* **Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)**: As mentioned before, lower values of these metrics indicate a better performance of our model. In this context, we observe that they reach their minimum values of 3.9 for MSE, 0.7 for RMSE, and 0.52 for MAE. These results are commendable and indicate that the model has effective predictive accuracy and minimized errors in its predictions.

In reviewing these results, there is a notable distinction between evaluating the model's performance for each symbol individually and evaluating the performance of all symbols collectively. By examining symbols individually, we gain valuable insights into the model's effectiveness in both statistical and investment evaluation.

It is crucial to note that the analysis pertains to a total of 498 symbols, with those having missing values omitted. Our model shows significant predictive capabilities for some symbols, accurately capturing results and patterns. Nonetheless, it's crucial to note that the model's performance varies, and it doesn't display the same level of accuracy for all symbols.

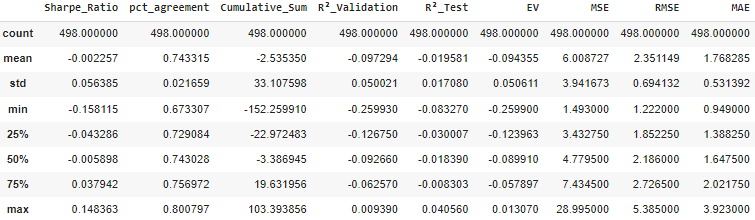
In this context, the individual symbol analysis not only reveals commendable investment metrics, indicating that the model helps investors make informed decisions, achieving approximately 80% accuracy in trade selection. In addition, statistical metrics such as MAE, RMSE, and MSE, although lower, support the model's predictive power, indicating that the overall predictions have a high degree of accuracy. This dual perspective underscores the comprehensive nature of the model's evaluation and highlights its potential to provide investors with reliable insights for effective decision making.

* The results of the Decision Tree Model:



Here we see that the Sharpe Ratio reaches a maximum value that is lower than that obtained with the linear regression model. However, we find a higher percentage of agreement, indicating the effectiveness of the model in capturing the direction of returns. In particular, there is a more significant improvement in the maximum value of R² for validation. The rest of the statistical metrics show consistency, with results comparable to those obtained in the Linear Regression model.

* The result of the Random Forest Model:



* 1. **Results without 2020:**

In the previous section, there was a noticeable instability in the closing prices and returns of symbols in the chaotic year of 2020. Specifically, there was a significant surge in closing prices accompanied by notable fluctuations in returns. This increased variability occurred at the same time as the global COVID-19 pandemic, which impacted not only public health but also reshaped investment landscapes, especially within the stock market. The aftermath saw the emergence of new trends and changes in investment habits.

To investigate the impact of these evolving trends on our stock predictions, we excluded data from the year 2020 from our training set. This exclusion enables us to assess whether the unique characteristics of that year introduce biases or effects in our predictive models.

| **Statistical Evaluation** | | | |
| --- | --- | --- | --- |
| **Metrics** | **RFR** | **DTR** | **LR** |
| **R² Training** | 0.055 | 0.028 | 0.002 |
| **R² Test** | -0.029 | -0.0178 | -0.005 |
| **R² Validation** | -0.053 | 0.00383 | 0.00157 |
| **EV** | -0.027 | -0.0145 | -0.0049 |
| **MSE** | 5.77 | 5.7 | 5.63 |
| **RMSE** | 2.4 | 2.39 | 2.37 |
| **MAE** | 1.73 | 1.72 | 1.71 |
| **Investment Evaluation** | | | |
| **Metrics** | **RFR** | **DTR** | **LR** |
| **Sharpe Ratio** | -0.004 | -0.009 | -0.006 |
| **PDA** | 74.45% | 74.54% | 74.67% |
| **Cumulative Sum of Return** | -3.53% | -5.76 | -4.47 % |

* For the Statistical evaluation Metrics:
  + **R² (Training, Test, Validation):**  RFR has the highest R² in training, followed by DTR and LR, suggesting a better ability to explain variance during the training phase. However, all models have negative R² values for the test and validation sets
  + **Explained Variance (EV):** LR has the lowest EV, suggesting potential limitations in capturing variability within the data.
  + **MSE, RMSE,and MAE:** LR consistently shows the lowest MSE and RMSE values, indicating superior accuracy. DTR slightly outperforms the other models in MAE.
* For the Investment Evaluation Metrics:
  + **Sharpe Ratio**:All models have negative Sharpe Ratios, with RFR having the least negative value.
  + **Percentage of Directional Agreement (PDA)**: PDA is similar for all models, with DTR having a slightly higher value.

- The negative R² values and difficulties in generalization are consistent with the previous interpretation.

- LR has consistently lower MSE and RMSE, consistent with the previous analysis.

- The absence of positive R² values for DTR in the "Results without 2020" table contrasts with its positive R² value in the previous interpretation. This may be due to the exclusion of the 2020 data.

* + **Cumulative Sum of Returns**: LR outperforms in Cumulative Sum of Return, having the least negative value.
  1. **Conclusion**

**Conclusion & Perspectives**