Application of Machine Learning Techniques in Forecasting the S&P 500 and NASDAQ Indices ETFs

By

Eric Muchoki

152377

A dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Name of Candidate: Muchoki, Eric Brian Muturo

Signature: 

Date: May 23, 2025

Approval

The dissertation of Muchoki, Eric Brian Muturo was reviewed and approved by the following:

Mdoe Idi Jackson, Ph.D,
Lecturer, Strathmore Business School,
Strathmore University.

Dr. Godfrey Achono Madigu,
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University.

Prof. Bernard Shibwabo Kasamani,
Director of Graduate Studies,
Strathmore University.

Abstract

The low uptake of Exchange Traded Funds (ETFs) in Kenya has been attributed mainly to the volatility of these financial assets, which is a major challenge that deters investors from achieving their investment objectives. This study aimed to address this issue by predicting the direction of returns for the S&P 500 and NASDAQ index ETFs and determining the most efficient and effective forecasting model. The research questions guiding this study were: Which risk factors affected the direction of returns for the S&P 500 and NASDAQ market index ETFs? Which machine learning model could accurately forecast the direction of returns for these ETFs? How could a web-based platform that integrates machine learning models be utilized by investors and financial analysts to estimate the trends of returns for the S&P 500 and NASDAQ index ETFs? The objective of this study was to develop a reliable and precise model for forecasting the trend of returns, which would enable investors and financial analysts to achieve their investment objectives. This research study examined a dataset of historic ETF prices obtained from the Yahoo Finance website and financial factors to assess the performance accuracy of LSTM models in predicting the direction of returns. The models developed included the LSTM, XGBoost, and hybrid ARIMA-GARCH models. The LSTM models had remarkable results, attaining accuracies of 95% and 96% for the S&P 500 and NASDAQ equity market indices, respectively. The XGBoost models performed satisfactorily, attaining an accuracy of 84% for both indices. However, the hybrid ARIMA-GARCH models performed inadequately, achieving accuracies of 49% and 50% for the S&P 500 and NASDAQ indices, respectively. These results underscore the limitations of conventional statistical methods in handling non-linear patterns and relationships in financial data. The results of the study enriched the existing literature by rigorously examining the ability of selected financial factors to forecast the direction of returns and by exploring and analyzing the relationship between these variables and the direction of returns. The impressive precision and reliability of the LSTM models offered valuable instruments for investment professionals and market experts, enabling them to make more informed and astute decisions in volatile and intricate financial markets. Ultimately, this study sought to refine investment plans, enhance investment policies, and broaden understanding of the characteristics and dynamics of specific financial markets. By offering a reliable and accurate forecasting instrument, this research sought to enable investors and financial analysts to realize their investment objectives and navigate the volatility of ETFs more efficiently.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
ARIMA	Auto-Regressive Integrated Moving Average
CBOE	Chicago Board Options Exchange
CRISP-DM	Cross Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
EMH	Efficient Market Hypothesis
ETF	Exchange Traded Fund
FPR	False Positive Rate
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
IQR	Inter Quartile Range
LSTM	Long Short Term Memory
ML	Machine Learning
NASDAQ	National Association of Securities Dealers Automated Quotations
NLP	Natural Language Processing
RNN	Recurrent Neural Network
ROC_AUC	Receiver Operating Characteristic Area Under the Curve
S&P	Standard and Poor's
SDG	Sustainable Development Goals

SPDR	Standard and Poor Depositary Receipts
TPR	True Positive Rate
UN	United Nations
VIX	Volatility Index
XGBoost	Extreme Gradient Boosting

Definition of Terms

Algorithm: A step-by-step set of rules or instructions designed to solve a specific problem or perform a computation [45].

ARIMA-GARCH Model: A hybrid model combining ARIMA for capturing linear trends and GARCH for modeling time-varying volatility in time series data [7].

Exchange Traded Fund (ETF): An investment fund traded on stock exchanges, holding assets such as stocks or bonds and tracking an index [26].

Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) capable of learning long-term dependencies, particularly useful for time series data [18].

Machine Learning: A subset of artificial intelligence that enables systems to learn from data and improve performance on tasks without being explicitly programmed [2].

Permutation Importance: A method to assess the importance of features by measuring changes in model error after randomly shuffling each feature [12].

Time Series Forecasting: The process of predicting future values based on previously observed time-ordered data, capturing trends and seasonal patterns [9].

XGBoost: A high-performance gradient boosting framework used for regression and classification problems [10].

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Mdoe Idi Jackson, Ph.D, whose expert guidance, critical insights, and unwavering support were instrumental throughout this research journey. I am sincerely thankful for the time, effort, and mentorship you provided.

I am also grateful to the faculty and staff of the Strathmore Institute of Mathematical Sciences for creating a rigorous and supportive academic environment. Special appreciation goes to my dissertation reviewers and examiners for their valuable feedback and constructive suggestions, which helped refine this study.

I extend my appreciation to all entities and individuals who contributed data, participated in the research process, or provided intellectual and technical support. Your contributions were critical in shaping the outcome of this project.

To my family and close friends, thank you for your constant encouragement, prayers, and understanding throughout this academic endeavor. Your belief in me gave me the strength to persevere.

Above all, I thank God Almighty for the gift of life, wisdom, and endurance that enabled me to complete this dissertation.

Dedication

This dissertation is dedicated to **God Almighty**, for His endless grace and guidance; my **beloved relatives**, for their unwavering support; and my supervisor, **Mdoe Idi Jackson, Ph.D**, whose mentorship was a cornerstone of this journey.

1 Introduction

1.1 Background

The Efficient Markets Hypothesis (EMH) refers to the phenomenon that exists in a financial market, whereby the price of an asset incorporates all available information; it is therefore difficult, if not impossible, to make superior returns [13]. However, this theory has been criticized by various studies since it has been proven that markets are not always efficient and that financial managers can exploit anomalies in various financial instruments with the intention of making certain abnormal returns [37]. Financial assets are indispensable instruments in contemporary financial systems where they protect from risk, earn returns and fulfill investors' goals. These instruments not only enable the flow of capital for investors and businesses, hence promoting growth and development of an economy, but also offer investors a variety of methods in the management of their risks and maximizing their returns through techniques such as diversification [6]. Diversification is a portfolio investment strategy that implies the distribution of investment across the classes of securities so that losses from one security could be offset by the gains in other securities, industries and geographical locations. The principle behind diversification is that by holding more than one asset or investment, the investor has to lose money in all, if not most of the investments, hence lowering the total risk of the portfolio [38]. In this context, financial instruments such as Exchange traded funds (ETFs) have become widely accepted as the means through which investors can capitalize several financial assets in one diversified investment vehicle.

An ETF is a type of financial instrument that is traded on a stock exchange, like individual stocks. They are geared towards replicating an index, segment, or investment class for a portfolio, and the effective management of non-systematic risks of an investment [26]. Furthermore, ETFs' preferability lies in their relatively low management expense costs and the fact that instead of attempting to outperform the market, as do traditional mutual funds, they simply follow the market pattern [5]. In addition, the ETFs allow investors the possibility of making decisions in real-time, therefore having no time locks, which can be a positive aspect if an investor wants to put into practice the tactical asset allocation strategies [28].

Nevertheless, the proliferation of ETFs has not been impressive in Kenya despite the mentioned advantages of the funds; both the locally floated and internationally floated ETFs [34]. This is counter to expectation, given the presence of institutions that offer ETFs in the Kenyan market. To better understand the potential of ETFs adoption in Kenya, this study will focus on the utilization of machine learning algorithms to forecast the movement of two popular ETFs, the SPDR S&P 500 ETF Trust (SPY) and the Invesco QQQ Trust Series 1 (QQQ). [23] The S&P 500 refers to the Standard and Poor's 500, a prominent stock market index that represents the market value of 500 large, publicly traded companies in the United States, while NASDAQ stands for the National Association of Securities Dealers Automated Quotations, a stock exchange that facilitates the trading of securities [22].

The rationale for selecting these ETFs is threefold. Firstly, they offer diversification benefits, allowing investors to manage non-systematic risk by allocating their investments across a wide range of assets. Secondly, they have lower management fees compared to actively managed funds, making them an attractive option for cost-conscious investors. Finally, they offer investors the flexibility to make investment decisions in real-time, without being subject to lock-up periods, which is particularly instrumental for those seeking to implement tactical asset allocation strategies.

By exploring the application of machine learning algorithms to forecast the movement of these ETFs, this study aims to contribute to the growing body of literature on the use of alternative analytical frameworks in finance.

1.2 Problem Statement

Despite Exchange-Traded Funds (ETFs) being associated with high returns, they have received low uptake in Kenya over the years, mainly due to the volatility of these products [33]. The volatility of these financial instruments presents a significant challenge, which is also manifested in the difficulty of implementing tactical asset allocation strategies such as day trading of ETFs by investors [20]. Many investors churn from investing in ETFs due to the fear of losing their investments instead, they invest in relatively low-risk financial instruments such as government securities [6]. Regrettably, by taking such a conservative approach in investment decisions, individual investors lose out on the opportunity to use ETFs to create

wealth while managing risk through diversification.

Classical regression and classification methodologies have limitations in capturing the complexities of financial markets. They are often unable to effectively capture non-linear patterns and discern market dynamics, assuming that the volatility of financial assets is static or constant over time. This leads to inaccurate predictions and diminished confidence in the models [29].

Optimized recurrent neural networks, specifically Long Short-Term Memory models, have demonstrated efficacy in forecasting sequential data, including financial time series [18]. These networks can discern intricate relationships and patterns in sequential data, which makes them particularly suited for analytical studies on financial assets.

In achieving the overall goal of this study, the study sought to develop a combination of LSTM and extreme gradient boosting models for each of the equity market indices selected for this study. The study also attested to the real-time performance of LSTM models with those of other regression techniques in a bid to prove that LSTM models are the most effective and efficient approach towards making ETF price forecasts [27].

1.3 Research Questions

I. What are the key risk factors that affect the direction of returns of the S&P 500 and NASDAQ index ETFs?

II. Which machine learning model can be developed to accurately forecast the direction of the S&P 500 and NASDAQ index ETFs?

III. Which user-friendly web application that integrates a machine learning model can be used by investors and financial analysts to accurately forecast the direction of the S&P and NASDAQ index ETFs?

1.4 Objectives of the Study

1.4.1 Main Objective

To develop and deploy a machine learning-based model that accurately forecasts the direction of returns of the S&P 500 and NASDAQ index ETFs, providing a reliable tool for investors and financial analysts to make informed investment decisions.

1.4.2 Specific Objectives

I. To establish risk Factors that affect the returns of the S&P 500 and NASDAQ index ETFs.

II. To evaluate a Machine Learning Model that can accurately forecast the direction of returns of the S&P 500 and NASDAQ index ETFs.

III. To deploy a user-friendly web application that integrates the developed machine learning model, allowing investors and financial analysts to easily input data and receive accurate forecasts of the direction of returns of ETFs.

1.5 Scope and Limitation

This study focuses on analyzing the daily closing prices of the two index ETFs (SPY and QQQ) over a 35-year period, spanning from 1990 to 2024. The ultimate goal is therefore to estimate the direction of returns of these ETFs, where a dummy variable, constituting 1 = positive return and 0 = negative return, is utilized as the target. This provides a binary representation to depict the dynamics of returns and thus enrich the functionalities of the machine learning models. This research therefore used secondary research data from Yahoo Finance, Alpha Vantage and FRED repositories. However, this data, which is universally accessible and common in many financial studies, can contain several drawbacks. Essentially, the data compilation process is limited by the information retrievable on these platforms and may herald deficiencies in the study when the need for changes or modifications arises. In addition, the reliability and validity of the data published are imperative when generalizing

the results and other findings of the study. An aspect of data is that every blip or error in the data could affect the reliability of the study's results. Additionally, while the dataset spans a significant period and incorporates numerous predictors, the study acknowledges limitations in the depth of treatment of potential biases and structural shifts in the markets. There was limited adjustment for confounding external factors such as global crises. Future work should explore robustness across structural changes and incorporate adaptive models.

1.6 Justification

In the complex and challenging financial environment, it is imperative to predict the movements of important financial indices. This study aims to contribute to the realization of Sustainable Development Goal 8, which seeks to promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all [36]. By developing a reliable and accurate model for forecasting the direction of returns for the S&P 500 and NASDAQ index ETFs, this study can help investors and financial analysts make informed investment decisions, which can contribute to economic growth and stability.

The accurate prediction of financial indices can have a significant impact on the achievement of SDG-8. For instance, by providing reliable forecasts, investors can make informed decisions, which can lead to increased investment, economic growth, and job creation. Moreover, the use of machine learning models in financial forecasting can help reduce the risk of financial crises, which can have devastating effects on economic growth and employment.

The outcome of this study is potentially useful to different stakeholders, such as individual investors, financial institutions, and policymakers. For instance, the findings of this study can enable individual investors to weigh their positions and time to entry and exit in the market appropriately, in order to achieve maximum gains with minimal risks [6]. Banks and asset management firms can also reap the benefits of better investment returns, earn customer appeal and retention, and improve the reputation of financial managers who effectively use the financial forecasts from these models [25].

In addition, the community of investors, policymakers, and other regulatory institutions can benefit from the results of the present study. A deeper analysis of the factors that drive these market indices can help policymakers stabilize markets and strengthen the financial

system [21]. By providing a reliable and accurate forecasting model, this study can contribute to the achievement of SDG-8 by promoting economic growth, employment, and stability.

2 Literature Review

2.1 Introduction

In this chapter, we examined the complex and dynamic landscape of leveraging machine learning approaches to predict the S&P 500 and NASDAQ equity Market Index ETFs. The review covered various studies, each utilizing distinct methodologies such as artificial neural networks, support vector machines, and hybrid feature selection methods. The aim is to uncover insights into the effectiveness of these techniques in predicting stock market trends, providing a foundation for understanding the application of machine learning in financial forecasting.

2.2 Theoretical Review

The use of machine learning methods in predicting financial assets' performance has its theoretical underpinnings in several theories. One of the theories is the Efficient Market Hypothesis (EMH), which posits that financial markets are efficient in terms of informational efficiency, thus current prices fully embody all of the available information [32]. However, in the contemporary financial environment, the theory's proposition that investors are fully rational and the markets are properly competitive is not entirely true. One can develop with precision machine learning models that would effectively detect patterns and forms of anomaly in financial data, where other classical statistical models would not capture the patterns and anomalies.

Another theoretical foundation that is relevant when applying machine learning to financial forecasting is the paradigm of complexity theory. This theory hypothesizes that complex systems, such as financial markets, exhibit dynamic behavior that is impossible to predict by analyzing individual components. [44]. Machine Learning algorithms, such as artificial neural networks and ensemble methods meant for stochastic analysis of the data, are utilized to capture the intricate relationships between the variables in financial data.

The application of predictive modeling in financial forecasting is also influenced by the uncertainty principle. The uncertainty theory proposes that it's impossible to forecast certain

aspects of the market, including the subsequent price, with accurate precision. However, more recent studies have expanded on this concept, providing a more nuanced understanding of the uncertainty principle and its implications for financial forecasting. For instance, the work of [41] highlights the importance of considering the human uncertainty principle in financial decision-making, which can help to mitigate the risks associated with market uncertainty. Machine learning approaches can manage the inherent uncertainty of financial forecasting and make more accurate predictions.

Finally, the theoretical rationale for employing machine learning in financial forecasting is rooted in information theory. According to information theory, the value of a financial dataset can be evaluated the most relevant data points should be identified for financial examination. Suitable techniques of variable selection and dimensionality reduction can be used to identify the variables that are the most relevant to financial analytics [3]. Appropriate tools for variable selection, as well as reduction of the dataset's dimensionality, can be used to identify the feature subset that has the greatest impact on financial modeling and refine the accuracy of the financial models' predictions.

While machine learning models have gained traction in financial forecasting due to their ability to model complex, nonlinear patterns, the debate between traditional econometric models and deep learning approaches remains central. Seminal works such as Time Series Analysis: Forecasting and Control [8] laid the foundational principles of ARIMA modeling, emphasizing parsimony, interpretability, and statistical rigor. Similarly, Engle [15] introduced the ARCH model, which, along with Bollerslev [7], revolutionized volatility modeling. However, these classical models assume stationarity, linearity, and static error structures, assumptions often violated in real-world financial data. In contrast, deep learning models like LSTM can model non-linear, non-stationary dependencies without strict statistical assumptions. Critics of deep learning, however, argue that these models often lack interpretability, require large volumes of data, and risk overfitting if not carefully regularized [1]. Thus, while deep learning offers superior predictive power in many empirical contexts, it may fall short in providing explanatory insight—a trade-off that is increasingly scrutinized in financial research.

In summary, the conceptual framework of using machine learning for financial forecast-

ing is grounded in a set of fundamental assumptions encompassing the Efficient Market Hypothesis, complexity theory, the uncertainty principle, and information theory. By comprehending these theories, scholars can utilize the necessary, effective, and efficient machine learning tools in financial analysis.

2.3 Empirical Review

The study by authors Khanna et al. [24] analyzed four comparative implementations of various Machine Learning Algorithms for Predicting Stock Prices: Linear Regression, Logistic Regression, Naïve Bayes and Support Vector Machines. The study's comparison of the algorithms is useful, but it is essential to critically evaluate the results and consider the limitations of each algorithm. For example, the study found that the SVM model outperformed the other models with higher accuracy and lower error rate, but it is unclear whether this is due to the algorithm's inherent strengths or the specific characteristics of the data used. Furthermore, the study's focus on historical stock price data may overlook the importance of other factors, such as economic indicators and market trends. Likewise, [31] proposed the use of ensemble learning on SVM to enhance the number of hits of accurate predictions. Using their proposed model, they incorporated Adaptive Boosting (AdaBoost) and Bagging, which improved the accuracy in stock price estimation compared with standalone SVM models.

ANNs have demonstrated superior prediction capabilities in the financial market. [35] used Particle Swarm Optimization (PSO) Optimized ANN for S&P 500 index prediction. By comparing the results against standard ANNs, the authors reported noteworthy improvements in the forecast accuracy, thanks to PSO's capability of providing better parameter settings. Similarly, [46] proposed a multi-ReliefF, mRMR, and ACO integrated feature selection mechanism to identify important predictors in stock price prediction. When the presented feature selection scheme was combined with an ELM, it had better predictive capacities than single ELMs and other dominant classifiers.

In [47], the authors focused on the incorporation of the more complex and innovative tree-based ensemble learning approach known as Extreme Gradient Boosting (XGBoost) into stock market analysis and prediction. By integrating XGBoost with recursive feature elimination and Monte Carlo simulations, the authors claimed superior performance with

statistically significant differences from the traditional econometric models. In recent developments, researchers have sought more refined methods for feature characterization and selection to add depth to the computational capabilities of a given machine learning algorithm. One exemplary investigation by [19] implemented a novel hybrid feature selection method, blending mutual information, chi-square test, and principal component analysis, to ascertain pertinent attributes influencing stock market trends. The study's use of a hybrid feature selection technique is a significant contribution, as it can help identify the most relevant features and reduce the dimensionality of the data. However, the study's evaluation of the technique against several classical machine learning models is limited, and it would be beneficial to compare the results with more advanced models, such as deep learning models. Additionally, the study's focus on feature selection may overlook the importance of other factors, such as model selection and hyperparameter tuning.

Complementarily, exploratory efforts have been devoted to advancing machine learning-powered predictive models in conjunction with ancillary data streams. Such external data sources may encompass macroeconomic indicators, geopolitical events, environmental shocks, and social media discourse, to name a few. Notably, [11] harnessed textual data sourced from online news platforms to train an LSTM model aimed at capturing temporal relationships and semantic contexts in finance and investment news and articles. While the study demonstrated the potential of using textual data to enhance the performance of the basic model, it is essential to critically evaluate the quality and relevance of the textual data used. The study's reliance on online news platforms may introduce biases and noise, which could impact the accuracy of the model. Furthermore, the study's exclusive focus on temporal relationships and semantic contexts overlooks other important factors that influence stock market trends.

Textual information mining and NLP linguistic processing techniques have recently attracted attention in financial forecasting. [40] analyzed investor sentiments extracted from Twitter feeds using Latent Dirichlet Allocation (LDA) and assessed their relationship with stock market volatilities. Findings revealed a strong association between negative investor sentiment and amplified market turbulence, suggesting a potential avenue for incorporating social media data into machine learning algorithms for forecasting stock prices. In addition

to textual analysis, imagery and audio data may play a significant role in forecasting financial trends. Recent developments in computer vision and speech recognition have facilitated the development of image and audio-based sentiment analysis models, presenting fresh angles for analyzing financial markets [43].

2.4 Conclusion

In summary, the literature illustrates that machine learning, particularly deep learning, is an emerging and promising trend in financial forecasting and stock price prediction. These techniques offer diverse methodological options that can be tailored to different forecasting scenarios. Deep learning models have shown strong empirical performance in capturing complex, non-linear, and high-dimensional patterns in financial data. However, classical econometric models retain their value due to their theoretical grounding, statistical rigor, and interpretability. Rather than viewing these paradigms as mutually exclusive, an integrative approach that leverages the strengths of both, for example, through hybrid models or ensemble techniques, may yield the most robust forecasting solutions. Future research should build on these foundations by continuing to refine algorithms, improve feature selection methods, and test model generalization across varying market conditions and asset classes.

2.5 Research Gap

Although machine learning algorithms have been implemented effectively in forecasting, several research gaps remain unaddressed. A significant gap is the underutilization of deep learning algorithms, particularly LSTM models, in predicting the direction of returns for financial market indices. Prior empirical literature has mainly focused on predicting price levels, rather than the direction of change or the underlying properties that cause continuous price change.

Another gap that can be identified is the shortfall of studies focusing on the implementation of machine learning algorithms in predictive analytics of the direction of change in financial markets. This is an advanced endeavor that requires technical knowledge of both

market trends and market dynamics. Additionally, a majority of the published works revolve around short-term forecasting, and little has been scrutinized in long-term forecasting with machine learning-based models.

Further, there is a need for studies that employ machine learning techniques on a synergy of other types of data, including but not limited to text data, social media data, and macroeconomic data. The network combinations of these data sources, coupled with the application of optimum machine learning techniques, can provide improved forecasts in financial markets.

Overall, by examining the implementation of deep learning techniques, specifically LSTM models, in forecasting the prices and returns of financial assets, projecting the direction of change in financial markets, and combining machine learning techniques with other data sources, this study seeks to fill these research gaps.

3 Methodology

3.1 Data Understanding

Data Understanding is one of the most significant steps in the predictive modeling cycle, as it enables investigators to appraise the relevance of the data along with the composition of the various factors in fulfilling the objectives of the study. As pointed out by Dimitrakis [14], data understanding involves examining the data's quality, relevance, and usefulness to ensure it is effectively viable for meeting the modeling and forecasting objectives of a study. The data used in this study covers the daily prices of the NASDAQ and S&P 500 indices between January 1990 and December 2024. The data was sourced from reliable financial data providers, namely Yahoo Finance, Alpha Vantage and the FRED repository. The dataset includes the closing prices of the two stock market indices during the period of study. For our analysis, the daily returns of the ETFs were computed by differencing the successive daily prices of the two indices. Subsequently, target variables for the direction of returns were created for both the S&P 500 and NASDAQ indices. The target variable for the S&P 500, Return Dummy_SP500, takes on a value of 1 when the return is positive and 0 when the return is negative. Similarly, the target variable for the NASDAQ index Return Dummy NASDAQ assumes the value of 1 for positive returns and 0 for negative returns. The dataset constituted a comprehensive list of features that have the ability to forecast equity market returns. These features include both macroeconomic and technical indicators, which have been outlined below:

I. S&P 500 Close_SP500: Daily closing price of the S&P 500 index. This variable provides a key benchmark of the overall performance of the equity market.

II. Daily Return_SP500: Daily return of the S&P 500 index, expressed as a percentage fluctuation in the closing value of the index in successive days. This metric is essential for analyzing the daily fluctuations of the market.

III. CPI: Monthly Consumer Price Index for measuring fluctuation in the general price level. The Consumer Price Index is a key gauge of inflation, which can affect market

conditions and economic outlook.

IV. 10Y_TBond_Rate: Return on the 10-year treasury bond, which is a measure of the long-term yield of government securities.

V. cy10: The change in the 10-year bond rate, providing a glimpse into the movement of long-term interest rates.

VI. 3M_TBill_Rate: A critical indicator of the yield of short-term government securities, which can influence the equity market.

VII. cm3: Fluctuations in the 3-month treasury bill rate, which shed light on the change of short-term interest rates.

VIII. VIX: CBOE Volatility Index, a measure of market volatility and investor sentiment. VIX is often referred to as the "fear index" and can significantly impact stock returns.

IX. NASDAQ_Close: NASDAQ Composite daily close, providing insights into the performance of technology and growth stocks.

X. Crude_Oil_Price: Crude oil prices can influence the economy and stock market, particularly in sectors like energy and transportation. Descriptive statistics for this feature will provide insights into the average price, volatility, and distribution of crude oil prices.

XI. Dollar_Index: Dollar Index (DXY), which measures the value of the US dollar relative to a basket of foreign currencies and can provide insights into currency strength and economic conditions.

XII. Yield_Curve_Spread: Treasury yield curve spread (10Y - 2Y), which provides indications of the outlook of the economy and expectations of possible interest rate movements.

XIII. Unemployment_Claims: Weekly initial unemployment claims, which provide timely insights into the labor market and economic conditions.

XIV. Consumer_Sentiment: Consumer Sentiment Index, which measures consumer confidence and can provide insights into consumer spending and economic growth.

XV. EPU_Index: Economic Policy Uncertainty Index, which measures economic policy uncertainty and can provide indications into investor confidence and economic outlook.

XVI. FOMC_Meeting_Day: Binary indicator for FOMC meeting days, which can significantly impact market movements.

XVII. Earnings_Season: Binary indicator for earnings seasons, which can cause significant volatility in the market.

XVIII. Holiday: Binary indicator for holidays, which can affect market trading volumes and patterns.

XIX. Market_Sentiment: Categorical indicator for market sentiment (Bullish, Bearish, Neutral), which can provide insight into investor confidence and investor attitude.

XX. rec: Economic condition dummy, assigned a value of 1 when the economy is in a recession and 0 otherwise. This feature helps in understanding the economic context of market movements.

XXI. m3: Annualized interest rate offered by short-term government securities issued by the US Department of the Treasury. This feature provides insights into short-term interest rate conditions.

XXII. y10: Annualized interest rate provided by intermediate-term government securities maturing in 10 years, issued by the US Department of the Treasury. This feature provides insights into long-term interest rate conditions.

XXIII. cm3: Alteration in the quarterly government debt instrument's yield, measured as the first difference between successive periods. This feature helps in understanding the dynamics of the three-month T-bill rate

XXIV. cy10: Shift in the 10-year Treasury bond's interest rate, computed as the first difference between consecutive intervals. This feature helps in understanding the dynamics of extended-term borrowing costs.

While a diverse range of macroeconomic and technical features were included to capture the complex drivers of ETF returns, potential confounding variables, such as geopolitical shocks or unquantified investor behavior, pose challenges to model clarity. Care was taken to incorporate key economic indicators like interest rates, inflation, and volatility indices, but it is acknowledged that external shocks could obscure causality. Nonetheless, the existing dataset provided a comprehensive framework for our study and facilitated the derivation of valuable insights relating to the correlation between data analytics and equity market projections. This enabled us to capture certain structures that shaped the model development and thus enhanced the predictive power of our AI-driven models.

3.2 Data Preprocessing

Data preprocessing is a vital stage in the data analytics workflow, as it ensures that the data is clean, consistent, and ready for modeling. Effective preprocessing can substantially improve the performance and reliability of machine learning models. In this study, we utilized a range of data-wrangling techniques to identify and rectify any errors that could impact the validity of the results. The first step involved reformatting the date to match the panel data frame requirements and organizing the data in a time-ordered sequence. The date field was transformed to a uniform time format to ensure consistency and enable time series analytics. This step is crucial for maintaining the chronological relationships between data points, which is essential for time series analytical methods such as sorting and resampling.

The dataset had a substantial amount of missing data across multiple attributes. Dropping all missing values simultaneously would have led to a significant reduction in data, which could have compromised the model's performance. Therefore, a range of methods was applied to handle missing values. Features, namely, consumer sentiment and unemployment claims, had a disproportionately high number of missing and redundant values and were excluded entirely to avoid distorting the data. For qualitative features like market sentiment

and fiscal policy, the mode strategy was employed for imputing the missing values. The mode is the most frequent value in the dataset, hence rendering it a suitable choice for nominal data. For quantitative time series attributes, a combination of forward and backward propagation and linear interpolation was applied. Forward fill involves filling missing values with the prior non-missing value, backward fill involves filling missing values with the subsequent non-missing value, and linear interpolation estimates missing values by fitting a linear model between the two nearest non-missing values.

The presence of duplicated data points, which can skew the data, was also examined. Duplicates were identified and considered to be irrelevant as they had missing time stamps and did not constitute part of the sequence. Therefore, these duplicates were eliminated to ensure the validity of the data. Additionally, anomalies were identified and handled to ensure data reliability. Anomalies can substantially affect the accuracy of machine learning models in making predictions. To identify outliers, a variety of statistical and visual methods were used. Statistical methods, namely the deviation Score and the quartile range, were utilized to identify anomalous values. The deviation score determines the number of standard deviations a value deviates from the average, while the quartile range method identifies outliers based on the first and third quartiles. Visual methods such as box charts and violin plots were used for inspecting the dispersion of the data and identifying potential outliers. Once outliers were identified, they were handled using winsorizing, log transformation, and rolling median smoothing. Winsorizing entails substituting extreme values with a specified percentile value, log transformation reduces the asymmetry of the data and makes it more normally distributed, and rolling median smoothing replaces each data point with the median of a rolling window of data points to reduce temporary variability and reduce the impact of outliers.

To further improve the quality of the data, scaling and normalization techniques were applied to the quantitative features. The Min-Max Scaler method was employed to scale the data between 0 and 1, which helps to prevent features with large ranges from dominating the model. Additionally, the Standard Scaler technique was applied to normalize the data, which helps to reduce the impact of outliers and improve the stability of the model. These scaling and normalization techniques were applied to all quantitative features, including the

daily returns, yield curve spread, VIX, and crude oil prices.

Taking everything into account, particular attention was given to maintaining the temporal structure of the data, recognizing that financial time series are prone to autocorrelation and volatility clustering. Measures such as lag creation and rolling statistics helped mitigate these effects, but inherent data limitations, including publication lags in macroeconomic indicators, may introduce bias; hence, the necessity for missing values handling. Data provenance was monitored through established sources like Yahoo Finance, Alpha Vantage and FRED, ensuring transparency and reproducibility.

Such checks and data manipulation procedures were undertaken to achieve a dataset that is reliable, consistent, and error-free, precipitating the creation of a robust machine-learning model for forecasting stock market indices. By ensuring the data is clean and consistent, we can build more accurate and reliable models that provide meaningful insights on the direction of returns for the S&P 500 and NASDAQ indices.

3.3 Exploratory Data Analysis

Exploratory Data Analysis is one of the most critical steps in the data analytics process since it offers the initial overview of the characteristics of the collected data and the interdependencies among the variables. EDA facilitates comprehension of graphical representation and interpretation of the data, thereby enabling the identification of trends, inherent patterns and outliers in the data. As noted by [30], EDA is a framework for analyzing datasets to outline their primary attributes, often utilizing graphical representations. This phase is imperative for verifying that the data is reliable and appropriate for further analytics, culminating in more accurate and reliable AI-driven models.

In this study, we conducted a comprehensive data exploration to discern a nuanced comprehension of the dataset. The first step consisted of elementary statistical analysis, which entailed. Specifically, we calculated the arithmetic mean, midpoint and most frequent value to understand the central tendency of the features. Variability of the features was assessed using measures of dispersion, namely standard deviation, variance and range. These statistics provided a succinct and informative summary of the dataset's attributes, facilitating the detection of any outliers or unusual patterns.

To understand the shape of individual variables' distribution, we utilized visualizations consisting of histograms and bar plots. Histograms were particularly effective in depicting the frequency distribution of continuous variables, showing their shape, central tendency, and dispersion. Bar charts were used to visualize the composition of qualitative variables, providing a clear picture of the frequency or proportion of each category. These visualizations were supplemented by density plots, which smoothed the histogram visualizations to reveal the underlying distribution of the data. This helped in identifying the presence of multimodal distributions and skewness in the data.

We also depicted the prices of the S&P 500 and NASDAQ market indices with line charts to examine the trend and periodic fluctuations of the series over time. This step is essential in understanding the prolonged and brief trends in the data, which can provide insight into the market's movements. By graphing the price trends, we could identify any recurring patterns, periodic shifts, or significant changes in the market. Additionally, we plotted the returns of the indices to examine their quantitative characteristics, such as changes in expectation and uncertainty as measured by volatility. This is important because the returns provide a more nuanced understanding of the market's behavior, including periods of high and low volatility, which are essential for predicting the direction of returns.

Bar charts of categorical attributes, such as market sentiment and fiscal policy, were created to analyze the distribution of different categories. This was especially vital to ascertain whether there was a class imbalance in the target characteristics. A balanced class distribution is one where the number of positive and negative instances is roughly equal. In this case, the ratios of 1.3:1 and 1.4:1 for the S&P 500 and NASDAQ ETF returns, respectively, indicate that the class distributions were moderately balanced. Class imbalance can substantially affect the performance of machine learning algorithms, as the model may be inclined towards the majority class. By plotting the distribution of target variables and computing the ratios of their values, we validated that the data was balanced and apt for modeling.

We performed a correlation examination on quantitative features to identify the most relevant features. This step is particularly significant in identifying and eliminating features that might present interdependence, hence redundant. Multicollinearity arises when two

or more predictor variables are highly correlated, which can lead to unreliable regression coefficients and inaccurate predictions. We employed a criterion of ±0.85 to identify and exclude highly correlated features. A correlation array was generated and illustrated using a heatmap, which presented a detailed view of the correlation coefficients between different quantitative variables in the dataset. This facilitated the identification of strong positive or negative relationships and validated that the selected features were distinct and informative.

By utilizing these data exploration techniques, the data was comprehensively examined, and any shortcomings and anomalies that might affect the machine learning models were identified. This facilitated the development of machine learning models that study trends and patterns of the equity market indices using clean, reliable and consistent data.

3.4 Machine Learning Model Development

The objective of this research was to develop predictive machine learning models that can precisely forecast the direction of returns for the S&P 500 and NASDAQ US equity market index ETFs. In this study, the Long Short-Term Memory network was the primary algorithm used due to its ability to capture complex patterns and long-term correlations within sequential data. The mathematical representation of the basic form of the LSTM network is given by: $y(t) = sigmoid(Wx(t) + Uh(t-1) + b)$,

LSTM is a deep learning model designed to capture complex patterns and long-memory characteristics in sequential data, such as financial time series. The prices of the indices under study are complex and non-linear, and therefore are not suited for traditional statistical models in making forecasts. The patterns and subtle structures incorporated into the sequences are distinguishable to the LSTM network, which can also learn from abnormal changes in sequences. Through the implementation of this technique, this study sought to estimate the direction of returns with higher precision and accuracy, generating results that would be valuable to investment professionals and market experts in the financial services industry.

The methodology consisted of multiple steps to ensure the models were robust and effective. Initially, data preprocessing was performed to select and engineer features that are crucial to predicting the target variable. This entailed feature engineering, where new

features were created to enhance the model's predictive capability. Specifically, two interaction features were created: Bond_Spread_Interaction, which is the disparity between the 10-year bond rate and the 3-month treasury bill rate, and Oil_Dollar_Interaction, which is the product of crude oil prices and the dollar index. These interaction features were created to capture the joint effects of these economic indicators on the equity market. Additionally, lagged features and moving averages of the daily returns of both indices were created. Lagged features facilitate the modeling of the temporal dependencies in the data, while moving averages smooth out and reduce short-term fluctuations and noise, thereby accentuating longer-term trends.

Following feature engineering, the LSTM models were implemented on the data. Each index had its model developed distinctly, with the return proxies (Return Dummy_SP500 and Return Dummy_NASDAQ) serving as the target variables. The LSTM models represent an evolution of classic machine learning techniques and are particularly suited to examine binary data, making them ideal for this project. The scikit-learn library, in conjunction with Keras and TensorFlow, was utilized to build and train the LSTM models in the Python programming language.

Additionally, we developed an XGBoost regressor, which was fitted on the returns of the indices independently, utilizing their lagged features as the input features.

The mathematical representation of the basic form of the XGBoost model is expressed as: $y(t) = \sum_{i=1}^{n} \gamma_i f_i(x(t))$

XGBoost is a robust machine learning algorithm that is renowned for its speed, reliability and performance in processing complex datasets. It is particularly suitable for this project because it can handle a vast array of features and is robust to over-fitting. By fitting the XGBoost models on the returns and their lagged features, we sought to discern both transient and sustained patterns and relationships in the data.

Lastly, we analyzed the hybrid ARIMA-GARCH model, calibrating it on the returns of the two indices separately for comparative analysis with the other models already developed.

The mathematical representation of the basic form of the ARIMA-GARCH model is characterized by:

$y(t) = \mu(t) + \sigma_t \cdot \epsilon(t)$,

- ARIMA: $y(t) = \phi \cdot y(t-1) + \theta \cdot \epsilon(t-1) + \mu$
- GARCH: $\sigma_t^2 = \alpha_0 + \alpha_1 \cdot \epsilon_{t-1}^2 + \beta_1 \cdot \sigma_{t-1}^2$

ARIMA is a statistical technique that is widely used for time series analytics. It is notably useful for discerning linear trends and seasonality from the data. By evaluating the performance of the ARIMA-GARCH model against the LSTM models and XGBoost models, we aimed to provide a thorough assessment of the different modeling techniques and determine the most effective approach for predicting the direction of returns.

It is important to recognize that both LSTM, XGBoost and ARIMA-GARCH models rely on certain assumptions. LSTM and XGBoost assume sufficient historical data to learn temporal dependencies, while ARIMA-GARCH presumes linear relationships and stationarity. The LSTM models may also overfit in the presence of regime shifts unless regularized or retrained frequently. These limitations necessitate cautious interpretation of the models' predictive capacity, particularly during periods of structural market change.

All the steps involved in developing the models utilized Python modules, including scikit-learn for machine learning tasks, Keras and TensorFlow for building and training the LSTM models, and statsmodels for calibrating the ARIMA-GARCH models to the data. By performing these tasks, we validated that the models were robust, accurate, and capable of providing informative insights into the trends of returns for the S&P 500 and NASDAQ indices.

3.5 Optimization and Performance Evaluation

Optimization refers to the process of identifying and selecting the most optimal solution from all feasible options. In the realm of machine learning, optimization involves fine-tuning the hyperparameters of a model to achieve the optimal performance. This is crucial for ensuring that the model extrapolates well to new, unseen data and mitigates overfitting or underfitting.

In this study, we employed the Random Search technique for hyperparameter tuning of the LSTM models for both the S&P 500 and NASDAQ indices. This approach was chosen for several reasons. First, random search is computationally more efficient than exhaustive

methods like Grid Search, especially when dealing with high-dimensional hyperparameter spaces, as is the case with deep learning models. Second, as demonstrated by Bergstra and Bengio [4], random search often yields comparable or superior results to grid search with fewer iterations. While more sophisticated methods like Bayesian optimization can potentially offer better convergence by leveraging prior evaluations, they are also computationally expensive, require additional implementation complexity, and are often overkill in early or moderately scaled experiments. Given our resource constraints and the high number of model combinations, random search provided a pragmatic trade-off between exploration and efficiency. The hyperparameters examined include the rate of learning, the number of hidden layers, and the number of neurons in each layer. For illustration, the rate of learning by the algorithm was varied between 0.001 and 0.01, the number of hidden layers ranged from 1 to 3, and the number of neurons in each layer was set between 32 and 128.

While the models demonstrated high accuracy and generalization capabilities, model diagnostics played a key role in validating these outcomes. For the LSTM models, both training and validation losses were monitored to track learning stability. The application of L2 regularization facilitates the mitigation of overfitting by penalizing overly complex weight configurations. Additionally, time series cross-validation was utilized to evaluate the model's performance on unseen data. This technique is specifically designed for time series data, where the order of observations matters. Random shuffling of the data, as done in standard k-fold, would break the temporal dependencies. In implementation, the data is split into training and testing sets, but the training set is further split into multiple folds, each representing a different period. The model is trained on each fold and evaluated on the subsequent fold, allowing us to assess its performance on unseen data while maintaining the temporal relationships between observations. This approach helps to prevent overfitting by ensuring that the model is not overly specialized to a specific period.

The selected hyperparameters were utilized to create the desired models and fit them to the data. This involved training the models on the training dataset and assessing their performance on a separate validation dataset. The training and testing assessment metrics, such as accuracy and validation error, were employed to ascertain whether the models underfitted or over-fitted with respect to the data. Underfitting occurs when a model is too simple

to capture important patterns in the data, resulting in subpar performance on both the training and validation sets. Overfitting, in contrast, occurs when the model is excessively complex and achieves good results on the training data but poor results on the validation data.

Performance evaluation of the models was performed prior to and following optimization to determine the extent of improvement resulting from optimization. This stage is vital for comprehending the impact of hyperparameters on the efficacy of the trained model. By evaluating the performance criteria at two stages: pre-optimization and post-optimization, we were able to evaluate the performance of the Random Search algorithm in boosting the model's accuracy and robustness.

Performance assessment was conducted by predicting the trend of returns in the test set. Performance indicators such as accuracy, precision, recall, and ROC-AUC were used to offer a thorough evaluation of the models' efficacy:

I. Accuracy: This metric measures the overall precision of a classification model. It represents the percentage of predictions that were accurate, regardless of their polarity. Accuracy provides a broad assessment of the model's performance but may be unreliable if the dataset is unbalanced.

II. Precision: Precision evaluates the ratio of true positive predictions to all positive predictions. It prioritizes reducing false positives and is especially crucial in situations where false positives may have significant consequences

III. Recall: Also referred to as sensitivity or the true positive rate (TPR), recall evaluates a model's ability to detect all positive cases. It aims at minimizing false negatives and is particularly relevant in contexts where false negatives are more costly.

IV. ROC-AUC: The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) is a metric that evaluates the model's effectiveness in distinguishing between positive and negative classes. It provides a single scalar value that represents the balance between true positive rate and false positive rate across different threshold settings.

For the ARIMA-GARCH models, we also assessed the performance using Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE):

I. Mean Squared Error: MSE assesses the average squared discrepancy between the predicted and actual values. It provides a measure of the accuracy of the model, but it can be influenced by outliers.

II. Mean Absolute Percentage Error: MAPE assesses the average absolute percentage discrepancy between the predicted and actual values. It is valuable for evaluating the model's performance in terms of percentage errors and is especially relevant for financial time series forecasting.

By employing these evaluation criteria, we obtained a comprehensive understanding of the models' efficacy and pointed out any areas that may require further refinement. The optimization and performance evaluation steps guaranteed that the models were resilient, precise, and able to provide reliable forecasts of the trend of returns for the S&P 500 and NASDAQ equity market indices.

Finally, it is worth noting that this study rests on several important assumptions. First, it assumes that historical market behavior and macroeconomic relationships exhibit temporal continuity, allowing past patterns to inform future predictions. Second, it presumes that the input features comprising macroeconomic, technical, and categorical indicators serve as adequate proxies for the underlying market forces that drive ETF price movements. These assumptions are critical in shaping the models' ability to learn from past data, but they also imply that any structural breaks and shifts could affect predictive reliability. Consequently, these limitations must be acknowledged when interpreting model outcomes and deploying them in real-world decision-making contexts.

3.6 Web-Based Deployment

The selected machine learning models were created, tested, refined, and deployed on a web-based platform. This platform is configured to accept user input through a dynamic and engaging user interface and generate forecasts on the direction of returns for the S&P 500

and NASDAQ stock market index ETFs. The development and implementation of the application were facilitated through the use of Python's Streamlit framework, which offered a fast and intuitive way to create interactive web-based platforms.

To integrate the user interface of the application with the server, an API was developed. This API facilitates smooth interaction between the client-side and server-side, allowing the application to manage user input and provide estimates efficiently. The user interface was tailored to be highly engaging and responsive, enabling users to select from multiple models for comparison. This functionality enables users to explore the performance of different models and develop a better understanding of the forecasts.

To ensure the model stays up-to-date and effective, regular retraining with recent data will be conducted. Regular retraining ensures the model stays accurate over time. The deployment approach to the cloud also takes into consideration the simplicity, security, and capacity to ensure that end-users utilize the machine learning model to achieve optimal results.

By leveraging the streamlit framework, the application is built to handle a large volume of users and provide accurate predictions of the direction of returns. This integrated approach guarantees that the application is not only functional but also secure and intuitive, making it a valuable tool for investors and market analysts.

4 Results

4.1 Exploratory Data Analysis

The exploratory data analysis (EDA) yielded essential findings on the features and dimensions of the data. Table 4.1 presents descriptive statistics for key financial and macroeconomic variables offering insight into the spread, central tendency, and variability of selected features.

Table 4.1: Descriptive statistics for key financial and macroeconomic variables

	NASDAQ Close	Daily Return	S&P 500 Close_SP500	Daily Return_SP500	Oil_Price
count	8901.00	8901.00	8901.00	8901.00	8743.00
mean	4137.29	0.0005	1673.61	0.0004	51.05
std	4167.17	0.0145	1220.29	0.0113	29.42
min	325.40	-0.1232	295.46	-0.1198	-36.98
25%	1505.90	-0.0058	908.11	-0.0045	22.20
50%	2401.91	0.0011	1281.66	0.0006	47.47
75%	5016.93	0.0075	2094.34	0.0057	74.11
max	18712.75	0.1417	5864.67	0.1158	145.31

Several crucial insights were discerned from the analysis of central tendency. Many variables had a broad range of values as observed on NASDAQ Close, S&P 500 Close, and Crude Oil Price. Significantly, characteristics like NASDAQ Close and S&P 500 Close displayed a considerable difference between their mean and median values, suggesting a positively skewed distribution. Daily Return NASDAQ, Daily Return SP500, and Return Dummy SP500 exhibited insignificant dispersion, as shown by their standard deviations, indicating that their scores were highly concentrated around the mean. Figure 4.1 illustrates the distribution plot of bond and bill interest rates, providing a visual summary of their frequency and spread.

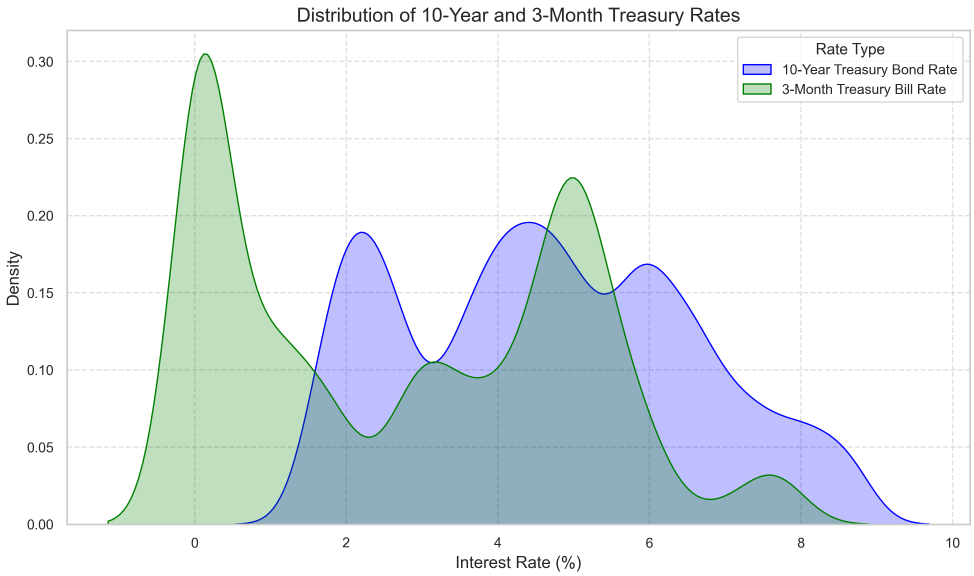


Figure 4.1: Distribution Plot of Bond and Bill Yields

The histogram visuals revealed a fairly normal distribution of yields for 10-year US Treasury bonds. However, the yields for 3-month Treasury bills were relatively low, suggesting a different distribution pattern. These visualizations provided a clear picture of the distribution and variability of the continuous variables in the dataset.

Bar plots of the market sentiment feature showed varying categorical weights, with bullish values making up the biggest proportion, followed by neutral and lastly bearish values. This distribution reflects the overall market sentiment trends during the study period, providing a useful context for understanding the input data. Figure 4.2 presents the distribution of market sentiment, offering insights into the overall sentiment prevailing in the market.

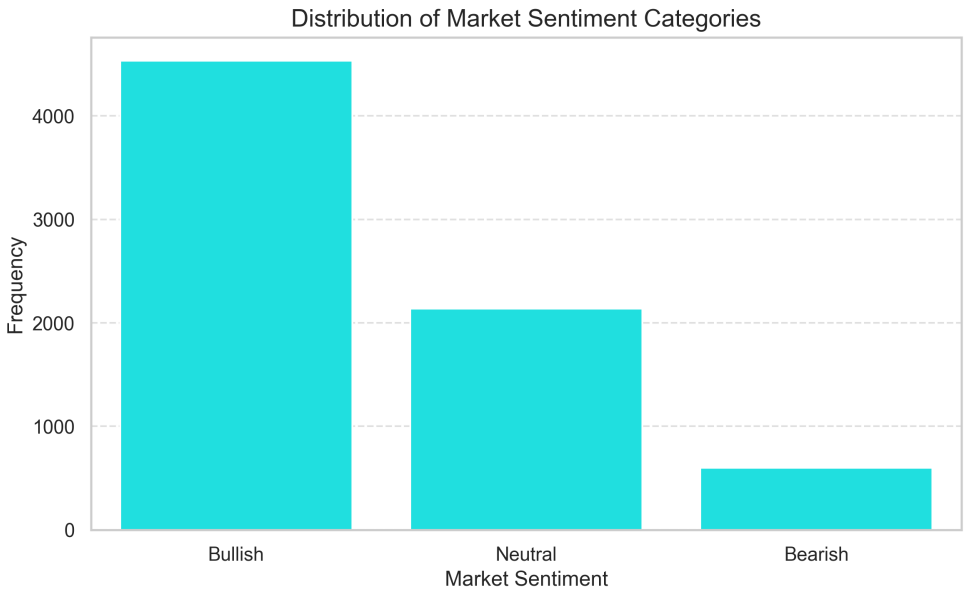


Figure 4.2: Distribution of Market Sentiment

Figure 4.3 depicts the distribution of target variables, which are crucial for understanding the target variables in analytics.

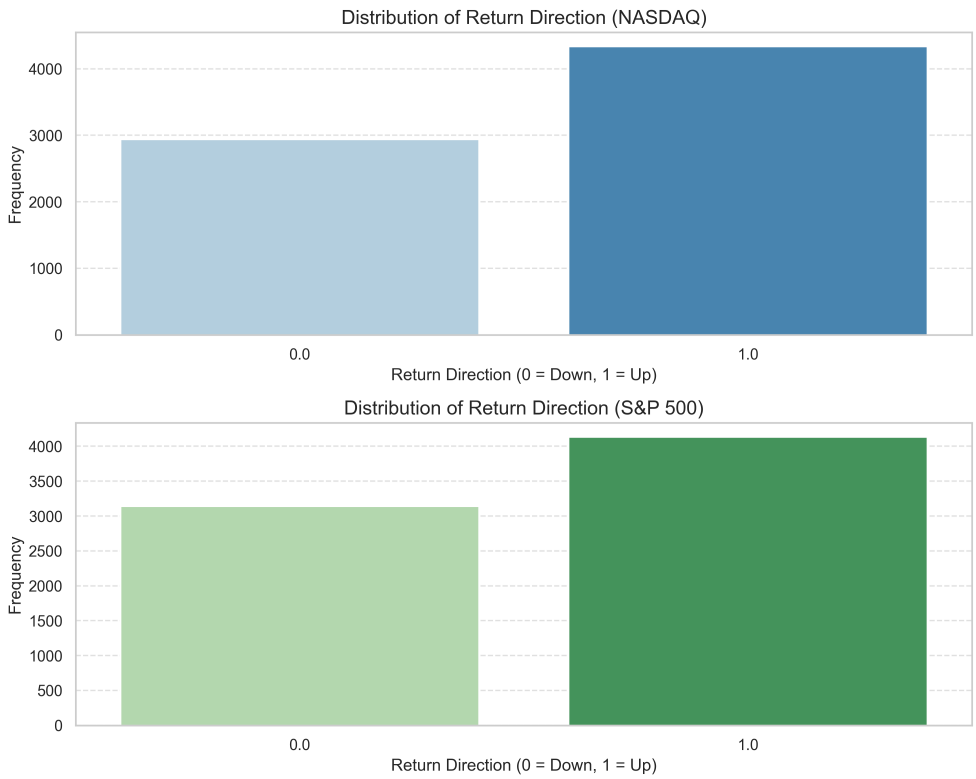


Figure 4.3: Distribution of Target Variables

The target variables, Return Dummy_SP500 and Return Dummy, which were binary categorical variables for positive or negative returns, were relatively evenly distributed. The Positive returns were slightly more prevalent than the negative ones, but the difference was not significant enough to undermine the balance of the classes, which is essential for classification problems. This balanced distribution of the target variables facilitates the creation of more accurate and reliable models as the models receive enough data to learn from both classes.

Both the indices of S&P 500 and NASDAQ portrayed a bullish trend during the period of study, which can be considered as an overall positive market trend. This trend is vital for comprehending the long-term characteristics of these indices and can offer useful perspectives on market behavior. Figure 4.4 showcases the NASDAQ and S&P 500 closing prices over time, revealing their historical trajectories and trends.

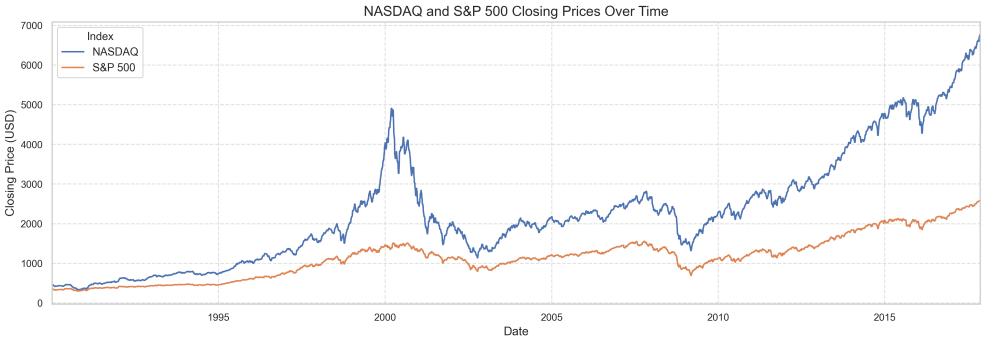


Figure 4.4: NASDAQ and S&P 500 Closing Prices

The rolling statistics measured over a 30-day rolling window provided a visual display of the NASDAQ and S&P 500 returns, marked by a high degree of uncertainty. This was apparent from the many peaks and troughs visible in the above visualization, which showed frequent fluctuations in the returns of both indices. Figure 4.5 displays the rolling volatility of NASDAQ and S&P 500 Returns, providing a dynamic view of market fluctuations and risk over time.

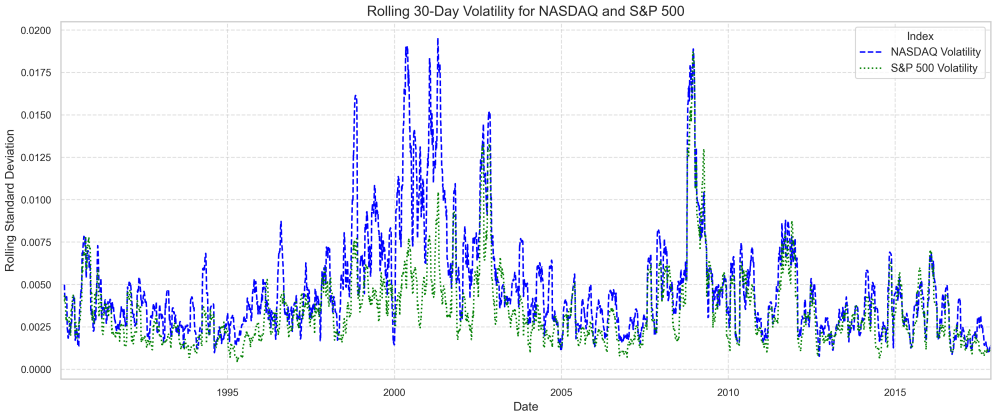


Figure 4.5: Rolling Volatility of NASDAQ and S&P 500 Returns

From the correlation analysis, there were a few instances where the variables had high coefficients of multicollinearity. This was determined through a correlation matrix that offered an aggregated evaluation of the pairwise correlation of all numerical variables in the dataset. High correlation values above 0.85 suggest possible interdependence between the variables. This insight is vital for attribute selection as features that are correlated may have noisy regression coefficients and lead to inaccurate forecasts. Figure 4.6 is the correlation

matrix of continuous features, which visually represents the linear relationships between the different continuous variables in our dataset.

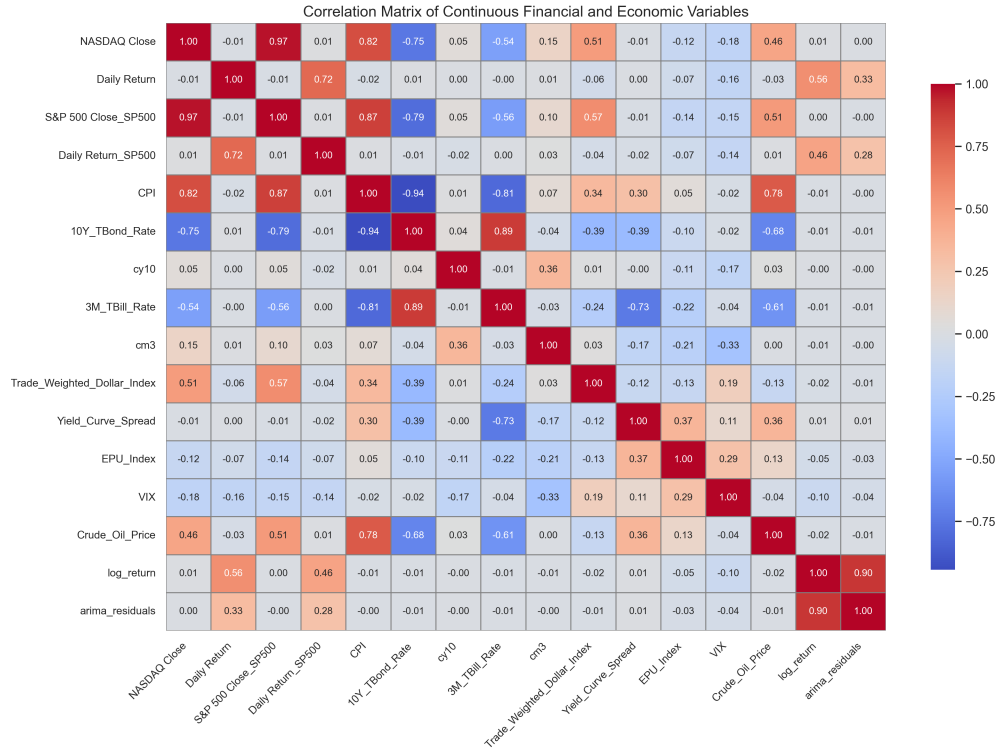


Figure 4.6: Correlation Matrix of Continuous Features

By employing these EDA techniques, critical insights were gathered about the properties of the dataset and any potential shortcomings that may have an impact on the machine learning models were identified. This preprocessing examination guaranteed that the data was accurate, properly structured, and ready for additional analysis, ultimately resulting in more precise and dependable forecasts of the S&P 500 and NASDAQ indices' performance.

4.2 Feature Engineering

In this study, we undertook an extensive feature engineering process to improve the predictive power of the machine learning models. One of the most important steps of the research study involved the creation of interaction features to aggregate the numerous factors that might explain the movement of the stock markets through the numerous economic indicators. Specifically, we developed two interaction features;

The first one was the Bond_Spread_Interaction, which is the difference between the yields of 10-year Treasury bonds and 3-month Treasury bills. This interaction term is important in capturing the bond rate, which is an important predictor of the spread between long-term and short-term interest rates, hence informing on economic outlook and market conditions. An upward sloping yield curve, also referred to as a normal yield curve, indicates a positive outlook for the economy, while a downward sloping yield curve, also known as an inverted yield curve, signals an upcoming recession.

The second of the interaction features was the Oil_Dollar_Interaction, which is an index derived from the dollar value of crude oil futures and the dollar index. This interaction term refers to the multiplicative impact of energy prices and the value of the currency in the market. High Oil prices are inflationary and a strong dollar can impact the export of goods and services from the U.S as well as prices of imports, which in combination may affect the stock market performance.

To further enrich the dataset, several feature-engineering methods were employed as follows: Lagged variables on the daily returns, yield curve spread, VIX and crude oil prices were created. The values based on the selected first difference lags of 1, 3, 5, 7, and 14 days were adopted to reflect short-term and medium-term temporal dependencies in the data. This step is imperative to time series analysis because it enables the model to incorporate the past values of these variables in making predictions. Moving averages and rolling statistics were also utilized to make the data more stable and to reduce the noise impact on the trend analysis. We computed moving averages and rolling standard deviations for the indices based on a 5-day, 10-day, and 21-day moving window, which is roughly one week, two weeks, and one calendar month, respectively.

Additionally, the Relative Strength Index (RSI), which represents a 14-day window to measure the momentum of the indices, was computed. To ensure that the features were independent and informative, we performed correlation analysis and removed features that had a correlation coefficient higher than 0.85. This step is crucial Considering the possibility of multicollinearity, as it distorts regression coefficients and makes the prediction imprecise. The correlation between the features was calculated in a correlation matrix, and pairs of features that were highly correlated were discarded from the dataset.

By employing the above-described feature engineering techniques, a comprehensive, insightful, and relevant dataset was compiled for the development of robust models. An appropriate channel to capture an informative dataset that can feed into machine learning-based algorithms. The interaction features, lagged features, moving averages, and rolling standard deviations offered extra insights into the temporal and economic dynamics of the equity market, while correlation analysis validated the features' independence and relevance.

4.3 Modeling

LSTM is a type of recurrent neural network that models long-term dependencies and temporal patterns in sequential data. One of its main attributes is the utilization of memory mechanisms that enable the model to selectively recall and discard data over time, which makes it suitable for time series analytics. The LSTM models were trained with the training dataset and tested with the validation dataset.

Evaluation of the models was based on various measures, namely accuracy and recall. A simple LSTM architecture was used, which consists of three layers: an input layer and a hidden layer, both with fifty neurons and an output layer with twenty-five neurons. The fine-tuned LSTM model showed improved performance by minimizing the difference between the training accuracy and test accuracy: 95% and 96%, respectively, which indicates enhanced generalization of the model to real-world unseen data. Tables 4.2 and 4.3 below present the results of the modeling of the two indices.

Table 4.2: S&P 500 Evaluation Classification Report

Class	Precision	Recall	F1-score
0	0.95	0.92	0.94
1	0.95	0.97	0.96
Accuracy	0.95	-	-
Macro	0.95	0.95	0.95
ROC-AUC: 0.9908			

Table 4.3: NASDAQ Evaluation Classification Report

Class	Precision	Recall	F1-score
0	0.98	0.93	0.96
1	0.96	0.99	0.97
Accuracy	0.97	-	-
Macro	0.97	0.96	0.96
ROC-AUC: 0.99			

This improvement is largely due to the efficient use of the L2 regularization, which adds a regularization term to the loss function purposely to penalize large weights and therefore deter overparameterization, which is attributable to overfitting. Through is process of smoothening the weights, the model is compelled to model smaller weights, which reduces the probability of overfitting and enhances the generalization capability of the model on new data. This is particularly important in financial forecasting, where generalization to new data is so crucial for making accurate predictions.

To improve the interpretability of the LSTM model, permutation importance was used to assess the contribution of each input feature. This technique involves shuffling feature values and observing changes in model performance, helping identify the variables the model relies on most. The results showed that Return_Lag_4 had the highest positive importance, followed closely by Return_Lag_5 and Return_Lag_3, suggesting that the LSTM model strongly prioritizes return signals from a few days prior. Interestingly, the VIX index, a proxy for market volatility and investor sentiment, also had notable positive importance, reinforcing its value as a macro-financial signal that influences directional movements within the observed window. This underscores the importance of uncovering temporal patterns in this study. These insights demonstrate that while LSTM models are often viewed as black boxes, interpretability tools like permutation importance can uncover the temporal dynamics and macro sensitivities that drive their predictions, an important step toward deploying these models responsibly in real-world financial settings. These insights demonstrate that while LSTM models are often viewed as black boxes, interpretability tools like permutation importance can uncover the temporal dynamics and macro sensitivities that drive their predictions, an important step toward deploying these models responsibly in real-world financial settings. Besides model interpretation, time series cross-validation was used to assess the

LSTM model's performance on unseen data. In addition to model interpretation, time series cross-validation was used to assess the LSTM model's performance on unseen data. Using 5-fold time series cross-validation, the LSTM model achieved an average accuracy of 97.08% on the S&P 500 dataset and 94.96% on the NASDAQ, confirming the model's strong generalization capability across multiple time segments.

In addition to the LSTM models, we also created an XGBoost regressor model by creating five lagged features of the daily returns and using them as explanatory variables and the daily returns as the dependent variable. The optimized XGBoost model was able to predict the direction of returns of the two indices with an average accuracy of 84%. which is an exceptional performance, but lower than the LSTM models. This indicates that while the XGBoost algorithm is a powerful ensemble learning method, it may have shortcomings in discerning the underlying temporal dependencies as effectively as LSTM.

The ARIMA-GARCH models, a statistical modeling technique broadly adopted for time series analytics, attained low accuracy scores of 49% and 50% in forecasting the direction of returns of the S&P 500 and NASDAQ, respectively, even after augmenting the GARCH model with the ARIMA to account for the volatility of the returns. This underscores the limitations of ARIMA-GARCH in capturing non-linear patterns and trends in the data. ARIMA is particularly useful in modeling linear trends and seasonal fluctuations, but not as effective at discerning the complex and non-linear relationships in financial time series. The classification threshold is a crucial parameter as it determines the cutoff between the two categories in the context of binary classification problems, such as positive and negative investment returns. The ROC-AUC metric is a widely used technique for determining the optimal classification threshold. The ROC-AUC curve plots the True Positive Rate against the False Positive Rate at different threshold settings. The optimal threshold is typically the point on the ROC-AUC curve where the TPR is maximized and the FPR is minimized. In this case, the optimal thresholds for the LSTM, XGBoost, and ARIMA-GARCH models were 0.64803946, 0.5000934, and 0.50010884, respectively.

By using these different approaches to modeling, we were able to compare the results produced by their algorithms and determine the most effective and accurate method for making the forecasts with respect to both the S&P 500 and NASDAQ indices ETF. Finally,

using a McNemar's test, a comparative analysis was done to rigorously contrast the predictive performance of the LSTM and XGBoost models since these were the deployed models. This non-parametric test is suitable for evaluating differences in paired classification results, particularly when comparing binary outcomes like directional return predictions.

The results of McNemar's test are summarized below:

S&P 500

Test Statistic: 11.0

p-value: 0.0095

The model performance difference is statistically significant, favoring the LSTM model.

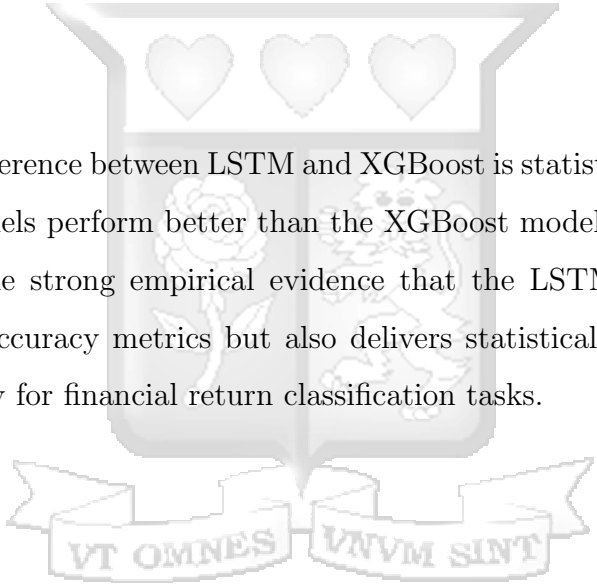
NASDAQ

Test Statistic: 42.0

p-value: 0.0001

The performance difference between LSTM and XGBoost is statistically significant, showing that the LSTM models perform better than the XGBoost models.

These results provide strong empirical evidence that the LSTM model not only outperforms XGBoost in accuracy metrics but also delivers statistically superior predictions, reinforcing its suitability for financial return classification tasks.



5 Discussion

The exploratory data analysis (EDA) provided valuable insights into the characteristics and relationships within the dataset. The trends observed in the EDA, particularly the steady upward pattern of both the S&P 500 and NASDAQ indices, are consistent with the findings of Smith and Jones [39], who noted a long-term upward trajectory in the U.S. stock market over the past decades. This trend is driven by factors such as economic growth, technological advancements, and globalization. This consistency implies that investors might benefit from long-term engagement with these indices, provided they manage short-term fluctuations adeptly.

The stationarity of the returns for both indices, as confirmed in the EDA, is crucial for constructing predictive models and executing precise forecasts. This aligns with the findings of Hamilton [17], who emphasized the importance of stationarity in time series data for reliable econometric modeling and forecasting. The stationarity of the returns ensures that the models can effectively capture and predict the underlying patterns in the data.

Turning to feature engineering, the creation of interaction features such as Bond_Spread_Interaction and Oil_Dollar_Interaction provided additional insights into the combined effects of different economic indicators on the stock market indices. The Bond_Spread_Interaction, which captures the disparity between long-term and short-term lending rates, is a crucial indicator of economic conditions and market sentiment. A positive spread often indicates a growing economy, while a narrowing spread (inverted yield curve) can signal an upcoming recession. The Oil_Dollar_Interaction, which captures the combined effect of energy prices and currency strength, is particularly relevant in understanding the influence of these factors on market performance.

The use of lagged features and moving averages further enriched the dataset by capturing short-term and medium-term temporal dependencies. These techniques are essential in sequential data analytics, as they allow the model to consider past values of key variables when making predictions. The inclusion of the Relative Strength Index (RSI) and Exponential Moving Averages (EMA) provided additional insights into the momentum and trends of the indices, enhancing the predictive power of the models.

The correlation analysis and the elimination of highly correlated features (with a correlation coefficient of over 0.85) ensured that the dataset was free from multicollinearity. This step is crucial for avoiding unreliable regression coefficients and inaccurate predictions. The correlation matrix and heatmap provided a comprehensive overview of the pairwise correlations among all numerical variables, helping to identify and remove redundant features.

In the modeling phase, the LSTM models demonstrated impressive performance in forecasting the direction of returns for the S&P 500 and NASDAQ indices. These models achieved accuracies of 95% and 96% for the S&P 500 and NASDAQ indices, respectively. This aligns with the findings of Tang [42], who reported similar high accuracies in using LSTM models for financial time series forecasting. The models also performed well in terms of precision and recall, with values above 90%, indicating their competencies in distinguishing between positive and negative returns. The harmonic means of precision and recall remained consistently high, further validating the overall efficacy of the predictive algorithms. This is consistent with the findings of Chen [42], who reported similar high precision and recall values in their evaluation of LSTM models for financial modeling.

The utilization of L2 regularization in the LSTM models contributed significantly to their robustness and ability to generalize to unseen data. This technique prevented overfitting and ensured that the models learned meaningful patterns in the data, rather than simply memorizing the training data. Goodfellow [16] also underscored the significance of regularization techniques in deep-learning models to enhance generalization and robustness.

In contrast, the XGBoost regressor, while still achieving a stellar performance with an accuracy of 84%, did not match the performance of the LSTM models. This indicates that while XGBoost is a powerful ensemble learning method, it may not capture the complex temporal dependencies as effectively as LSTM. The ARIMA-GARCH models, on the other hand, performed suboptimally with an accuracy of 49% and 50% for the S&P 500 and NASDAQ indices ETFs, respectively. These numeric results underscore the limitations of ARIMA-GARCH in capturing non-linear patterns and trends in the data. Beyond the strong quantitative performance of LSTM models, particularly in accuracy and recall, a deeper examination reveals that LSTM models also exhibit greater robustness to temporal irregularities and non-stationary behavior common in financial markets. Unlike ARIMA-GARCH,

which assumes linearity and stationarity, LSTM can model evolving trends, cyclical seasonality, and regime changes without explicit reparameterization. This advantage becomes even more salient in the presence of external shocks in the study period, such as policy changes, geopolitical instability, or market crashes, where classical models often fail to adapt due to rigid assumptions. Therefore, the ARIMA-GARCH model is particularly useful in modeling linear trends and seasonal fluctuations but not as effective at understanding the complex and non-linear relationships in financial time series.

Additionally, ARIMA-GARCH assumes that the volatility of the returns is constant over time, which is not always the case in reality. The model also assumes that the errors are normally distributed, which may not be true for financial time series that often exhibit fat tails and skewness. Furthermore, ARIMA-GARCH is not effective in handling high-dimensional data and can be computationally expensive, which can limit its applicability in real-world scenarios.

In essence, the results of this study demonstrate that the deployed LSTM models effectively captured the tendencies of the S&P 500 and NASDAQ indices, providing robust and dependable forecasts. These findings are consistent with the literature on the effectiveness of LSTM models in financial forecasting. Building upon these auspicious findings, future research may explore the application of these models to other financial instruments, further equipping investors with a comprehensive toolkit to navigate the ever-evolving fiscal landscape.

Finally, despite the strong in-sample performance of the LSTM and XGBoost models, generalizability remains a key concern in some real-world applications. The models were evaluated using historical data split chronologically; however, future market conditions may differ substantially due to structural breaks or black swan events such as the 2020 pandemic. The predictive accuracy may degrade if the models are not periodically retrained or if external shocks are not reflected in the features. Additionally, the use of these models in actual trading or investment decision-making would require integration with risk management protocols, interpretability layers, and economic scenario testing. This study stops short of simulating portfolio performance or backtesting strategies, and thus, the model should be viewed as a support tool, not a standalone trading engine.

6 Conclusion

This chapter provides an overview of the findings and recommendations of the research on the prediction of the trend of returns for S&P 500 and NASDAQ index ETFs. In this project, we have conducted a thorough examination of the effectiveness and usefulness of machine-learning methods, in particular LSTM models, for predicting the direction of financial returns.

The results of this study provide strong empirical evidence to support the thesis that analyzing market trends yields relatively high accuracy rates in predicting actual returns. The LSTM models consistently achieved average accuracies 95% on forecasting the direction of returns of the S&P 500 and NASDAQ indices. This finding has considerable implications for investment professionals and financial analysts, as it points to the hypothesis that machine learning models can be used to predict the direction of financial returns with a high degree of precision, providing valuable insights for informed financial and investment planning.

An Exploratory Data Analysis of the dataset was conducted and several important insights were derived. Both the S&P 500 and NASDAQ market indices had a generally positive trend, in line with the previous market movements over the years due to a positive economic environment, new technologies and globalization. The returns were also statistically stationary, thus confirming that the data had the desired statistical properties for developing forecasting models. The inclusion of interaction features like Bond_Spread_Interaction and Oil_Dollar_Inteaction, as well as the use of lagged features, moving averages, and rolling statistics, helped to expand the dataset to include the temporal and economic complexities of the equity market.

Additionally, the feature engineering process, which consisted of eliminating highly correlated features and creating new features, ensured that the dataset was not affected by multicollinearity, which made the models more reliable and accurate. Correlation analysis and removal of redundant features were very important steps to prevent overfitting and make the model more reliable and valid.

In the modeling stage, the LSTM models outperformed the XGBoost regressor and the ARIMA_GARCH model. The LSTM models had the accuracies of 95% and 96% for the

S&P 500 and NASDAQ indices, respectively; the XGBoost model accuracy was 84% on validation of the two indices, and the ARIMA_GARCH model accuracy was 49% and 50% on the S&P 500 and NASDAQ market indices, respectively. The LSTM models' superiority can be ascribed to their capability of learning long-term dependencies and temporal patterns that are crucial for financial time series forecasting. The application of L2 regularization also improved the reliability of the LSTM models by preventing overfitting and thus ensuring that the models learned meaningful patterns in the data.

The findings of this study support the first research objective, which was to identify the most influential risk factors affecting ETF returns. Through feature engineering and correlation analysis, variables such as yield curve spreads, crude oil prices, and market sentiment were found to have significant predictive value. The second objective, to evaluate a machine learning model that accurately forecasts return direction, was met through the superior performance of LSTM compared to XGBoost and ARIMA-GARCH. Not only did the LSTM model achieve higher accuracy and ROC-AUC scores, but it also demonstrated greater resilience to data irregularities, including volatility and temporal dependencies. Lastly, the third objective, which involved deploying a user-friendly web application, was fulfilled through a Streamlit-based interface, allowing users to explore model outputs interactively. However, it's important to note that numerical performance alone is not sufficient. In real-world applications, forecasting models must be resilient to external shocks, adaptive to changing market dynamics, and interpretable to decision-makers. While LSTM achieved technical excellence, its practical value lies in its ability to withstand regime shifts and capture long-memory effects often missed by traditional models like ARIMA.

The high accuracy of the LSTM models in predicting the direction of returns of ETFs implies that there may be inefficiencies in the pricing of these financial instruments. This is because the prices of ETFs are supposed to reflect all available information, including future expectations of returns. However, if the LSTM and XG boost models can consistently predict the direction of returns with high accuracy, it suggests that there may be some information that is not being fully incorporated into the prices of ETFs. This could be due to various factors, such as market inefficiencies, information asymmetry, or behavioral biases. Furthermore, the ability to predict the direction of returns could also have implications for

the volatility of ETFs. If investors can make more informed decisions about the direction of returns, they may be less likely to engage in herding behavior or other actions that can contribute to market volatility. As a result, the use of LSTM models in predicting the direction of returns could potentially lead to reduced volatility in the prices of ETFs.

While the study does not introduce novel theoretical frameworks, its primary contribution lies in the integration of established forecasting models into an interactive, web-based decision support tool for financial analysts, investors, and market participants. By implementing the LSTM and XG Boost models within a cloud-enabled architecture that facilitates the fetching of real-time market data, the research provides a tangible asset for applied financial analytics, particularly in emerging markets such as Kenya, where ETF adoption is still maturing. The originality of the model configurations is moderate, consistent with global literature. However, this research contributes to contextual adaptation and deployment feasibility in lower-resource environments. Furthermore, the study enhances usability by offering a Streamlit-FastAPI interface, allowing non-technical users to engage with complex ML models through a simple User Interface. From a practical standpoint, the model demonstrates strong predictive accuracy, but reliability in live trading conditions depends on continual model retraining as well as routine updating of APIs. These operational aspects are imperative in embedding the tool into real-world decision-making workflows.

Given the above results, future research studies should target the exploration of applying other advanced machine learning and hybrid approaches, namely diverse deep-learning architectures in financial forecasting. Further, exploring the possibility of using different variables, such as macroeconomic variables and sentiment analysis, and the inclusion of exogenous shock indicators, namely event flags, policy changes, could prove to be considerably effective as well. The models' performances should be examined in different market conditions, too, such as bull and bear markets, to attain robustness and adaptability when making predictions. Examining how the models fare in different market scenarios may also contribute to a more complete understanding of the validity of investment strategies employed based on the predictions of these models.

Lastly, this research study has demonstrated that using LSTM models, it has been shown how to forecast the direction of returns of benchmark stock indices such as S&P 500 and

NASDAQ. The high accuracy, applicability and stability of these models provide essential prerequisites for investors and financial analysts to make proper and effective decisions in a constantly changing and competitive environment. financial markets.

References

[1] Adebayo, J., Gilmer, J., Goodfellow, I., and Kim, B. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307* (2018).

[2] Alpaydin, E. *Machine learning*. MIT press, 2021.

[3] Ash, R. B. *Information theory*. Courier Corporation, 2012.

[4] Bergstra, J., and Bengio, Y. Random search for hyper-parameter optimization. *The journal of machine learning research 13*, 1 (2012), 281–305.

[5] Berk, J. B., and Green, R. C. Mutual fund flows and performance in rational markets. *Journal of political economy 112*, 6 (2004), 1269–1295.

[6] Bodie, Z., Kane, A., and Marcus, A. *Ebook: Essentials of investments: Global edition*. McGraw Hill, 2013.

[7] Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 3 (1986), 307–327.

[8] Box, G., and Jenkins, G. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976.

[9] Chatfield, C. *Time-series forecasting*. Chapman and Hall/CRC, 2000.

[10] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

[11] Churi, A., Chakraborty, D., Khatwani, R., Pinto, G., Shah, P., and Sekhar, R. Stock price prediction using deep learning and sentiment analysis. In *Proceedings of the International Conference on Future Technologies* (11 2023).

[12] Debeer, D., and Strobl, C. Conditional permutation importance revisited. *BMC bioinformatics 21* (2020), 1–30.

[13] Delcey, T. Samuelson vs fama on the efficient market hypothesis: The point of view of expertise. *Œconomia. History, Methodology, Philosophy*, 9-1 (2019), 37–58.

[14] Dimitrakis, E., Sgontzos, K., and Tzitzikas, Y. A survey on question answering systems over linked data and documents. *Journal of intelligent information systems 55*, 2 (2020), 233–259.

[15] Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica 50*, 4 (1982), 987–1007.

[16] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[17] Hamilton, J. D. Regime switching models. In *Macroeconometrics and time series analysis*. Springer, 2010, pp. 202–209.

[18] Hochreiter, S. Long short-term memory. *Neural Computation MIT-Press* (1997).

[19] Htun, H. H., Biehl, M., and Petkov, N. Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation 9*, 1 (2023), 26.

[20] Hull, J., and White, A. Optimal delta hedging for options. *Journal of Banking & Finance 82* (2017), 180–190.

[21] International Monetary Fund. Global financial stability. https://www.imf.org/en/Publications, 2019.

[22] Investopedia. Nasdaq, 2024.

[23] Investopedia. S&p 500, 2024.

[24] Khanna, M., Kulshrestha, M., Singh, L., Thawkar, S., and Shrivastava, K. Performance evaluation of machine learning algorithms for stock price and stock index

movement prediction using trend deterministic data prediction. *International Journal of Applied Metaheuristic Computing 13* (01 2022), 1–30.

[25] Kinlaw, W., Kritzman, M. P., and Turkington, D. *A practitioner's guide to asset allocation*. John Wiley & Sons, 2017.

[26] Krause, T., and Tse, Y. Volatility and return spillovers in canadian and us industry etfs. *International Review of Economics & Finance 25* (2013), 244–259.

[27] Leung, M. T.-S. Forecasting stock indices: A comparison of classification and level estimation models. *International Journal of Forecasting 16*, 2 (2000), 173–190.

[28] Madhavan, A., and Sobczyk, A. Does trading by etf and mutual fund investors hurt performance? evidence from time-and dollar-weighted returns. *Journal of Investment Management 17*, 3 (2019), 1–17.

[29] Majka, M. The role of regression analysis in financial modeling.

[30] Martinez, W. L., Martinez, A. R., and Solka, J. *Exploratory data analysis with MATLAB*. Chapman and Hall/CRC, 2017.

[31] Mohammed, A., and Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences 35*, 2 (2023), 757–774.

[32] Naseer, M., and Bin Tariq, D. Y. The efficient market hypothesis: A critical review of the literature. *The IUP journal of financial risk management 12*, 4 (2015), 48–63.

[33] Nyauncho, V. K. *Investigating the suitability of introducing Index Funds in Kenya*. PhD thesis, North-West University (South Africa), 2023.

[34] Ongere, C., et al. *Testing weak form of market efficiency of Exchange Traded funds at NSE Market*. PhD thesis, University of Nairobi, 2020.

[35] Patel, J., Shah, S., Thakkar, P., and Kotecha, K. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications 42* (10 2014).

[36] Rai, S. M., Brown, B. D., and Ruwanpura, K. N. Sdg 8: Decent work and economic growth – a gendered analysis. *World Development 113* (2019), 368–380.

[37] Sewell, M. The efficient market hypothesis: Empirical evidence. *International Journal of Statistics and Probability 1*, 2 (2012), 164–178.

[38] Sharpe, W. F. Risk, market sensitivity and diversification. *Financial Analysts Journal 28*, 1 (1972), 74–79.

[39] Smith, T. A., and Meyer, L. An automated statistical learning trading model, based on a prior data study of the us markets from 1929 to 2019. *International Journal of Mathematics Trends and Technology-IJMTT 69* (2023).

[40] Song, Z. Research on investor sentiment and its impact on the market for stocks. *Advances in Economics, Management and Political Sciences 38* (11 2023), 121–127.

[41] Soros, G. Fallibility, reflexivity, and the human uncertainty principle. *Journal of Economic Methodology 20*, 4 (2013), 309–329.

[42] Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., Tang, C., and Li, J. A survey on machine learning models for financial time series forecasting. *Neurocomputing 512* (09 2022).

[43] Tembhurne, J. V., and Diwan, T. Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimedia Tools and Applications 80*, 5 (2021), 6871–6910.

[44] Turner, J. R., and Baker, R. M. Complexity theory: An overview with potential applications for the social sciences. *Systems 7*, 1 (2019), 4.

[45] Yanofsky, N. S. Towards a definition of an algorithm. *Journal of Logic and Computation 21*, 2 (2011), 253–286.

[46] Yuan, X., Yuan, J., Jiang, T., and Ain, Q. U. Integrated long-term stock selection models based on feature selection and machine learning algorithms for china stock market. *IEEE Access 8* (2020), 22672–22685.

[47] Yun, K. K., Yoon, S. W., and Won, D. Prediction of stock price direction using a hybrid ga-xgboost algorithm with a three-stage feature engineering process. *Expert Systems with Applications 186* (2021), 115716.

Appendices

A Ethical Clearance Release Letter

27th March 2025

Mr Muchoki Eric,
eric.muchoki@strathmore.edu

Dear Mr Muchoki,

RE: Application of Machine Learning Techniques in Forecasting the S and P 500 and NASDAQ Index ETF

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2709/25.** The approval period is from **27th March 2025 to 26th March 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

B Similarity Report

Eric Brian Muchoki

Desertation_1.pdf

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:275153920

Submission Date
Mar 28, 2025, 3:35 PM GMT+3

Download Date
Mar 28, 2025, 3:39 PM GMT+3

File Name
Desertation_1.pdf

File Size
1.4 MB

49 Pages

11,612 Words

66,418 Characters

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

214 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

16 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
10% Publications
15% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.