A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR STOCK MARKET PREDICTIONS, AN EMPIRICAL STUDY ON PAKISTANI MARKET.

By:

(Muhammad Asim) ***(01-321242-019)***

(MBA 1.5 Year)

Supervisor:

Dr. Muhammad Naveed Qazi

HR and Management Department
Bahria University Islamabad

Fall 2025

Majors: FIN
S.No. (F-21)

"A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR STOCK MARKET PREDICTIONS, AN EMPIRICAL STUDY ON PAKISTANI MARKET."

By:

(Muhammad Asim) ***(01-321242-019)***

Supervisor:

(Dr. Muhammad Naveed Qazi)

HR and Management Department
Bahria University Islamabad

Fall 2025

FINAL PROJECT/THESIS APPROVAL SHEET

Open Defense Examination

Open Defense Date 13/01/2026

Topic of Research: (A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR STOCK MARKET PREDICTIONS, AN EMPIRICAL STUDY ON PAKISTANI MARKET.)

Names of Student(s) Enroll # 01-321242-019

- Muhammad Asim
-
-

Class: (MBA 1.5 Year)

Approved by:

_____

(Dr. Muhammad Naveed Qazi)

Supervisor

_____

Qurat Ul Ain Waqar

Research Coordinator

_____

Dr. Aftab Haider

Head of Department

ACKNOWLEDGEMENT

Before going any further, I would wish to say my hearty thanks to Almighty Allah and his numerous blessings, direction, and grace. Only by His grace, I can finish this dissertation and go through numerous trials which I had to face on my way of studying. His encouragement has given me the power, patience, and endurance to go through all the challenges and realise my ambitions. I owe my supervisor, Dr. Muhammad Naveed Qazi, who has been of great help with his guidance and encouragement and knowledge during this research. Her consistent encouragement, valuable feedback, and helpful recommendations have played an important role in this dissertation. Her patience, devotion, and commitment to work made me strive to work hard and always kept me motivated even at the toughest stages in this study. I do appreciate the time and energy she used to make me complete this work. My whole heart would also like to say a very big thank you to my family, whose support, empathy, and sympathy have been my star throughout life. My parents have been my pillars of support, especially my parents, so they have encouraged me to move through my education levels. This has been enabled by their sacrifices, guidance and uncompromising faith in my capabilities. To my siblings and the extended family, thank you, we shall never lack your support and I know that you always had faith in me. My friends and peers are also people to whom I am grateful to have had supportive companionship during graduate studies. Their support, positive interaction and motivation assisted me to traverse through the rigors of my academic life. This journey was easier and enjoyable because experiences, thoughts, and difficulties will be shared with them. Lastly, I would like to thank those people who in one way or other helped this dissertation achieve success. This has been made possible by the guidance, support and encouragement I got on both spiritual, professional and personal fronts. The outcome of all these individuals support, encouragement and inspirations is this dissertation and I owe my utmost sincerity to every one of you.

ABSTRACT

The surge in high-frequency financial data has transformed the landscape of market forecasting, driving a shift toward machine learning (ML) integration. This research provides a critical performance assessment of Linear Regression (LR), Random Forest (RF), and XGBoost in predicting the closing prices of premier securities listed on the Pakistan Stock Exchange (PSX). Utilizing a decadal dataset (2014–2024), the investigation encompasses the KSE-100 benchmark index and high-volatility assets including OGDCL, Lucky Cement, and Fauji Fertilizer. The models were constructed using Python-based frameworks and rigorously validated through Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).The empirical findings reveal a notable "simplicity paradox" within the PSX environment. Linear Regression demonstrated remarkable predictive precision, consistently yielding R^2 values exceeding 0.99 across the majority of the portfolio. In sharp contrast, the advanced ensemble methods—Random Forest and XGBoost—proved highly susceptible to overfitting, particularly in the context of the KSE-100 Index, where they produced negative R^2 values (-0.01 and -0.02 respectively). This indicates that such models struggle to distinguish signal from noise in frontier markets characterized by non-linear economic shocks. The study concludes that for short-term forecasting in the Pakistani context, parsimonious linear models provide superior interpretability and structural stability. These results offer actionable intelligence for risk managers and investors, advocating for the strategic use of objective, data-driven tools to navigate emerging market volatility.

Key Words: Stock Market Prediction, Machine Learning, Linear Regression, Random Forest, XGBoost, Pakistan Stock Exchange, Financial Forecasting, MAE, RMSE, R^2 Score.

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

The financial markets play an important role in the economic development of any given country because they provide a platform upon which investments and capital formation, as well as distribution of the available resources take place. However it is among the hardest to forecast the stock trend movements due to their extreme volatility, the market mood and external economic factors. Over the last few years, the output of the Machine Learning (ML) has been a potentially useful method of producing financial predictions due to its capacity to describe the non-linear, highly complicated relationship between stock market data. The ML algorithms can identify hidden patterns and dependencies, which can be used in stock prediction with big data compared to the traditional algorithms. As Ali et al. (2020) and Khan and Ahmad (2021) have already proven, the Pakistan Stock Exchange (PSX) is highly predictable by the use of ML models. Similarly, it has been demonstrated by Patel et al. (2015) and Fischer and Krauss (2018) that algorithms like Random Forest, SVM, and LSTM can be utilized to model complex time-series data and increase the level of forecasting. Although the world is progressive, there is a lack of researches on how the Pakistani context can use ML techniques. The vast majority of the available literature has examined foreign markets (US, India and China) and implemented advanced deep learning models without comparing them to the simpler and understandable models. Furthermore, the local literature tends to lack a thorough analysis employing standard accuracy measures such as MAE, RMSE, and R 2 on historical data to the PSX. In this study, three popular ML algorithms, including Linear Regression, Random Forest, and XGBoost, have been applied on and compared to the historical data of the PSX. The study will seek to identify the best model to use in the prediction of stock prices, and therefore offer practical information to the investors, financial analysts and policymakers of the Pakistani stock market capital market.

1.2 Problem Statement

Stock markets are volatile and complicated by nature, and proper forecasting of stock prices remains a long-standing challenge for investors and financial analysts. The determinants of share prices are numerous, ranging from company performance and investor sentiment to macroeconomic indicators such as interest rates, inflation, and global market trends. These factors interact in a non-linear manner and the prediction of prices using ordinary methods is getting harder and harder. Although machine learning (ML) has turned out to be an instrument with significant potential to identify complex patterns that do not show up in human analyses (Zhu and Enke, 2021), a critical Complexity Paradox is present in the context of the Pakistan Stock Exchange (PSX). Ensemble models that are both highly complex, such as Random Forest and XGBoost, are frequently popular in global literature because they are thought to be necessary in modern trading (Jiang et al., 2023). Nevertheless, a large gap in research exists on whether the sophisticated models are more accurate in the PSX, which is highly noisy, sparsely liquid and extremely volatile in the short-term. Current local studies either concentrate on the foreign market design or the single-model application (Khan and Ahmad, 2021; Hameed and Iqbal, 2024). It is not thoroughly and comparatively analyzed with standardized measures such as Mean Absolute Error (MAE), Root Meal Squared Error (RMSE), and R-squared (R^2) over a 10-year time span (20142024). As a result, the Pakistani investors and policymakers do not have evidence-based guidelines according to which algorithmic architecture generalizes optimally without falling prey to overfitting due to noise. The paper at hand particular overcomes this ambiguity by comparing the Linear Regression (LR) model with the ensemble architecture (RF and XGBoost) on the KSE-100 Index and on 13 different assets, namely, Pakistan State Oil (PSO), UBL, Systems Limited, Nishat Mills, Lucky Cement, Fauji Fertilizer, Unilever Pakistan, Nestle Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co, and OGDCL. Early empirical findings of this study suggest that complex models do not generalize and tend to produce negative R^2 values (e.g., -0.010 KSE-100) whereas simple linear models produced almost perfect goodness-of-fit (R^2 of 0.98). This research aims to formalize this comparison, providing a tangible, fact-based advantage for strategic capital allocation and risk management in the unique Pakistani financial landscape.

1.3 Research Gap

Though many studies have been done to determine the application of the Machine Learning (ML) algorithms in predicting stock prices, most of such studies have been done in developed markets like the United States, India, and China. As an illustration, Patel et al. (2015) used trend-based information with Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN) to predict stock trends in the Indian stock market and found that ML is highly suitable at forecasting intricate financial trends. Likewise, Fischer and Krauss (2018) applied the model of deep learning called Long Short-Term Memory (LSTM) to the US stock market and got high results in accuracy of prediction. Nti et al. (2020) have carried out a systematic review of stock prediction literature and found that the ML techniques are superior to the traditional statistical ones in most financial forecasting situations but there is still a severe shortage of studies on the Pakistan Stock Exchange (PSX). The PSX is operating in emerging market environment with various behavior of investors, trading volumes, and economic structures than the developed countries. Thus, what works in the developed markets might not necessarily reflect in the correct findings in Pakistan. The other gap in the research, according to Ali et al. (2020), is that local market dynamics, including liquidity problems and limited institutional involvement, may have a different impact on the stock price fluctuations that may need to be tested in particular on PSX data. Nearly all local literature has experimented on one or two algorithm types, e.g., Artificial Neural Networks (ANN) or Support Vector Machines (SVM) and not compared their results with simpler and more understandable models like Linear Regression or other ensemble models including Open Forest and XGBoost. As an example, one of the articles by Khan and Ahmad (2021) compared some algorithms on PSX data but omitted such boosting algorithms as XGBoost, which leaves a knowledge gap regarding the best model of accuracy and generalization using local financial data. Another problem is that most Pakistani works do not apply such a standardized evaluation measurement as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared. These metrics are needed to determine the actual predictive power of machine learning models and to be able to make reliable comparisons between algorithms. The lack of adequate evaluation measures, as observed by Patel et al. (2015) and Fischer and Krauss (2018), usually results in unfinished or wrongful assumptions concerning the performance of this model. In addition, although deep learning solutions, such as LSTM have been effectively implemented in foreign markets (Fischer and Krauss, 2018), the comparative

study of deep learning solutions and a more straightforward machine learning algorithm in Pakistan has not been conducted. This puts financial analysts at a crossroad of the kind of model (simple, ensemble-based, or deep learning), which offers the most reliable and feasible output in the PSX set-up. Lastly, it can be observed that there is no practical advice that Pakistani investors and analysts can apply on how to apply and interpret ML-based forecasting tools in real-life decisions. As Nti et al. (2020) note, although most researchers are interested in the technical performance of models, there is a lack of studies that can be translated into practical implications to practitioners. This leaves a misalignment between academic researches and application in the financial industry in Pakistan. To solve these problems, the current paper will be a comparative study of three commonly applied machine learning algorithms, namely Linear Regression, Random Forest, and XGBoost, on past stock data on the Pakistan Stock Exchange. MAE, RMSE and R 2 are used to assess the performance of each of the models to assess the model that gives the most accurate and reliable stock price forecasts in the Pakistani context. The proposed research will be able to contribute to the scholarly literature and effective investment strategies by offering an objective, data-driven comparison of ML models based on the specifics of the financial market in Pakistan. Additionally, the Pakistan Stock Exchange is also in a climate where the price movements tend to be erratic and non-linear owing to the high rate of changes in investor sentiment, liquidity, and trading habits. The internal market features such as these render the process of stock price forecasting a difficult one, particularly when one considers the application of the traditional statistical models that presuppose the stable or linear relationships only. Even though we have not factored in external macroeconomic or political factors in this research, the complexity that is inherent in price and volume data in the PSX still demands the current technologies of machine learning, which is able to successfully capture latent trends and adjust to the local market dynamics that are changing rapidly. Through this requirement, there is little research in the Pakistani world which has evaluated systematically the performance of various ML models with the same input features and the same market data. Thus, one can identify an evident gap that concerns the empirical comparison and approaches only stock-specific variables and evaluates what algorithm can make the best predictions to investors who are actively working in the Pakistani market with its specific structure. This research bridges this gap as it uses Linear Regression, Random Forest and XGBoost to analyse the historical PSX data and determine the model that brings a maximum level of accuracy to the prediction.

1.4 Research Questions

The questions of the research of this study are the following:

1. Which machine learning algorithms linear regression, Random forest and XGBoost has the best prediction power of stock prices in the PSX?

2. Which data variables have the most impact on improving stock price prediction models in the Pakistani environment?

1.5 Research Objectives

This study is conducted with the purpose based on the problem statement.

- To evaluate machine learning models for predicting stock prices on the Pakistan Stock Exchange (PSX).
- To determine the relative predictive accuracy of various machine learning algorithms, specifically Linear Regression, Random Forest, and XGBoost.
- To identify the key data variables such as historical prices and volume that significantly influence the accuracy of stock price predictions.
- To assess model performance using standard accuracy measures, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score.

1.6 Significance of the Study

This research is very important because it covers one of the most difficult aspects of the field of finance: the proper forecast of the stock prices in the most volatile and unpredictable market. The stock market is one of the most important parameters of the economy as it affects the decisions made regarding investments and financial stability in general (Zhu & Enke, 2021; Jiang et al., 2023). Nonetheless, the traditional forecasting models do not usually reflect the complex and non-linear patterns of the price movements because of its unpredictability (Makridakis et al., 2022; Jiang et al., 2021).Specifically, the study of the addition of Machine Learning (ML) methods gives a new dimension to financial forecasting because it allows handling more data, which is more precise and relevant in this context (Nti et al., 2020; Zhang et al., 2022). This research is especially relevant because it targets the Pakistan Stock Exchange (PSX) as an emerging market that has not been thought of in the previous literature (Hameed & Iqbal, 2024;

Khan & Ahmad, 2021). This research can be considered valuable to the academic community as it provides empirical data on the performance of three machine learning algorithms, including the Linear Regression, Random Forest, and XGBoost, in the Pakistani market, which is largely not explored compared to research on developed markets like the US or India (Zhu & Enke, 2021; Jiang et al., 2023).In practice, it helps investors, analysts, and policymakers to gain value by applying and comparing three machine learning algorithms to real PSX data over a period of 2014 to 2024. Practically, the conclusions of the current study will assist financial practitioners and investors to determine the most effective algorithm in stock price prediction so that they can make more effective investment and risk management decisions (Zhang et al., 2022; Hameed & Iqbal, 2024). The findings can be used by financial institutions and brokerage firms to enhance their trading models, improve their portfolio strategies, and reduce forecasting errors in the Pakistani financial market (Jiang et al., 2021; Makridakis et al., 2022).In addition, the study will promote the use of data-driven decision-making and assist in updating financial analysis tools in new economies as it proves the effectiveness of the ML models in the Pakistani financial market (Zhu & Enke, 2021; Nti et al., 2020). It also offers useful information to policymakers and higher education initiatives in demonstrating how machine learning can be integrated into finance education and policy formulation to drive innovation, enhance market performance, and raise investor trust (Khan & Ahmad, 2021; Hameed & Iqbal, 2024). Altogether, the research is relevant as it fills the gap between finance and technology and supports the practical implementation of ML models in the real financial markets and development of the modern economic situation in Pakistan (Jiang et al., 2023; Makridakis et al., 2022).Moreover, the research is of great importance as it indicates how machine learning can be a convenient and alternative tool even to non-technical individuals, like finance students, new analysts, or investment practitioners in Pakistan. In some of the emerging economies, the use of the sophisticated analytical tools is usually hampered because of the absence of the good technological infrastructure, lack of skills and absence of localized studies to demonstrate the actual benefits of the same. Using algorithms, like Linear Regression, Random Forest, and XGBoost, that are publicly available and, therefore, relatively simple to use, this study demonstrates that the creation of efficient prediction models does not require the use of highly sophisticated deep learning architecture and specialized computing environments. This makes the findings especially essential to brokerage firms, asset management firms as well as individual

investors who may not be supplied with sophisticated ITs but need the ability to make effective, timely forecasting tools. Moreover, the study can raise the financial literacy as it will demonstrate that objective and data-driven strategies can replace tendencies of speculation and emotional biases which predominantly determine investment behavior in the emerging market, like Pakistan. This research will result in a more rationalized and transparent way of market forecasting by enabling the application of standardized measures of evaluation; MAE, RMSE, and R square that will ultimately culminate in excellent decision making, reduced vagueness, and long-term investment approaches in the Pakistani financial ecosystem. To illustrate this, the paper notes that in the case of the KSE-100 Index, Linear Regression had an almost perfect R 2 of 0.9986, and the error bands of such individual stocks as Jazz and UBL were quite small with MAE of 0.29 and 1.25 respectively, demonstrating the practicality of these models (Hameed and Iqbal, 2024).

1.7 Research Contribution

Financial forecasting has changed dramatically to computational intelligence especially in the turbulent environment of Pakistan Stock Exchange (PSX). Recent empirical data by Bukhari et al. (2023) demonstrates that machine learning frameworks coupled with the localization of indicators of macroeconomic aspects provide better trend analysis than traditional econometric models. Although more advanced methods of ensembles such as the Random Forest and XGBoost have become popular due to their capability of dealing with non-linear data, Hameed and Iqbal (2021) discovered that more simplistic linear regression tools tend to be more stable and predictive of closing prices in the short term in the Karachi Stock Exchange setting. In addition, Raza and Akhtar (2024) highlight the importance of incorporating particular technical indicators to overcome the liquidity restrictions and volatility profile peculiarities of Pakistani equities. Yaqoob and Abdullah (2025) also highlight the importance of a sector specific analysis in the emerging markets such as the Pakistan Stock Exchange (PSX). Their study highlights that the effectiveness of a machine learning algorithm is not universal; on the contrary, it acutely relies on the peculiarities of volatility and liquidity limitations of particular stocks. Through modern Long Short-Term Memory (LSTM) networks they show that complex models can be used to model deep temporal dependencies though with significant variability in their effectiveness across the different sectors of the PSX. This paper is an essential modern reference,

and there is a reason why comparative frameworks, such as the one applied in your thesis, are necessary, to identify the superiority of simple linear models or the complexity of neural networks to the unique structural realities of the Pakistani financial environment. This study serves as an effective addition to the scholarly and the practical sphere of finance as it is aimed at incorporating the methods of machine learning in stock market forecasting in the framework of the Pakistan Stock Exchange (PSX). Although in former research on stock forecasting, the main emphasis has been on developed economies, including the United States, India, and China (Patel et al., 2015; Fischer and Krauss, 2018; Nti et al., 2020), the given research offers new empirical evidence of a less-researched market. The study will also add to the literature by comparing three well-known machine-learning algorithms, namely Linear Regression, Random Forest, and XGBoost using historical stock data (2020-2024) to determine the most effective machine-learning algorithm in stock price prediction in the specific context of the Pakistani market. The research is also commendable, in that it employs the most important evaluation metrics; Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R 2 Score, to make sure that the accuracy of the model is objectively and comprehensively evaluated, which is usually lacking in the local studies before. This study presents practical implications in addition to its scholarly value to investors, analysts, and financial institutions by identifying the opportunities of machine learning as a decision-support tool that can be used to develop better investment policies and risk management. The results would help the market players to understand how data-driven strategies can reduce uncertainty, more forecasting, and financial performance. Besides this, the study has its contribution to the financial innovation in an emerging economy since it demonstrates how non-IT experts, especially those operating in the sphere of finance, can effectively apply machine learning to comprehend the manner in which markets operate. Such interdependence of finance and technology contributes to the future of interdisciplinary research, helps to modernize financial analysis in Pakistan, and is an example of other researchers and policymakers who wish to include artificial intelligence and data analytics to the process of financial foretelling and investment choices. Besides these core contributions, the study also has an important input in contributing to the field on a variety of aspects that are important. To begin with, it provides a context-based empirical foundation of understanding the behavior of machine learning models in terms of the structural characteristics of the Pakistani market, such as diminished liquidity, increased volatility, political sensitivities and intermittent price movements. Such markets as the

PSX are likely to be susceptible to the realization of fluctuations and turbulent tendencies that are not characteristic of developed international markets and this study offers evidence-based information regarding how the ML algorithms respond to such peculiarities. This improves the theoretical understanding of model generalizability in other economic contexts. Second, the research paper will aid to enhance the methodology in the Pakistani financial research market. The research reveals the importance of methodological rigor in creating meaningful findings using the standardized preprocessing steps, including, but not limited to, data cleaning, feature engineering, train-test splitting, and performance evaluation. It is interesting because most of the local researches that have been conducted in the past lack clarity on their methodology frameworks and therefore, reproducibility and the validity of the models is a challenge. It is a methodological criterion that could be applied in further investigations of the use of ML-based forecasting in Pakistan since it demonstrates an understandable and referable workflow. Third, the study provides a powerful comparison framework, which enables a researcher and practitioners to evaluate the merits and demerits of different machine learning models in the background of the application to new markets. The linear and non-linear models are compared which will give a good understanding of the stock price behavior in Pakistan. The fact that simple linear models outperform complex ensemble models is a point that most individuals would assume to be the case and goes a long way to convince scholars to reconsider the nature of the models they apply on PSX data. This can be useful in theorizing market behavior in the emerging economies and motivates more motivated interpretation of model choice in accordance with the nature of the data rather than the popularity of the algorithms. Fourth, the research is also practical in application of technology in the financial aspect especially with a setting where the idea of artificial intelligence integration remains in its early stage. The research helps lower the cost of entry of financial workers who wishes to apply machine learning in their analytics by demonstrating to them how easy it is to incorporate the easily-accessible technologies such as Python, pandas, scikit-learn, and XGBoost into their processes. This is pro-capitalism in the financial industry and an online financial analysis revolution. Lastly, the research provides the foundation of the interdisciplinary study in the future. It paves the way to research that uses other sources of data, such as macroeconomic indicators, investor sentiment, global market trends, or technical indicators, to determine whether predictive accuracy can be further increased by using hybrid or deep learning models. It also promotes the use of time-series specific ML models by

the researchers, including LSTM and GRU networks, which can possibly be more sensitive to longer-term dependencies. These guidelines mark the possibility of the process of permanent advancement of ML-based prediction in the PSX and facilitate the advancement of data-driven decision-making in the financial environment in Pakistan.

1.8 Structure of the Thesis

The present thesis is structured into five exhaustive chapters, each of which deals with an important issue of the study. Chapter 1, Introduction, presents the basis of the study and includes the background, problem statement, research gap, objectives, research questions, hypotheses, significance, contribution, scope, and limitations that enable the readers to have a clear understanding of what the study is and relevant to them (Zhu & Enke, 2021; Jiang et al., 2023).Chapter 2, Literature Review, analyzes the available theories, models, and empirical research on stock market forecasting, discusses conventional forecasting methods and their drawbacks (Makridakis et al., 2022), introduces the tools of machine learning (Nti et al., 2020), explains the chosen algorithm-based forecasting tools of the linear regression, random forest, and XGBoost based on the previous empirical research (Zhu & Enke, 2021), and identifies gaps. Chapter 3, Research Methodology, describes both the research design, data collection process, preprocessing of data, implementation of machine learning model, evaluation metrics (MAE, RMSE, R^2), and tools and software utilized, as well as the step-by-step workflow of the analysis (Hameed & Iqbal, 2024; Zhang et al., 2022).Chapter 4, Results and Discussion, include descriptive statistics of the selected data of stocks, the performance of all the models, a comparison of the algorithms according to the evaluation metrics, the interpretation of the findings in terms of the previous research (Khan & Ahmad, 2021), and its implication to investors and analysts in Pakistani market (Hameed & Iqbal, 2024).Chapter 5, Conclusion and Recommendations, summarizes the main findings, identifies the most successful algorithm, comes up with conclusions on the possibility of using machine learning to predict stock prices in Pakistan, and makes practical recommendations to investors and financial professionals (Zhang et al., 2022), identifies the contribution to the theoretical literature, presents limitations of the study, and offers suggestions on future research (Jiang et al., 2023).The thesis is ended by a References section with all the sources mentioned and Appendices with the Python code, data

tables and graphical illustrations of the predicted and actual stock prices, which provide clarity and reproducibility of the research (Jiang et al., 2021).

Summary of the Chapter

Chapter 1 has given a detailed introduction to the study giving the context, background and incentive of researching the prediction of stock prices in the Pakistan Stock Exchange (PSX). It emphasized the issues of investors and analysts in the crisis of market volatility, the shortcomings of the classic forecasting approaches, and the opportunities of the machine learning solutions to enhance the quality of predictions. The chapter effectively provided the problem statement, outlined the research gaps in the existing literature, and defined the reason of doing a comparative analysis of three popular algorithms, namely, Linear Regression, Random Forest, and XGBoost, based on historical stocks data between the years 2014 and 2024. The research questions and objectives were presented in a manner that ensured that the research was focused and the significance, as well as contributions of the research were discussed with a greater focus on the academic contribution and the practical contribution provided to investors, financial institutions as well as the policymakers. The chapter also covered the boundaries and constraints of the study, explaining the limitations of the study and the possible shortcomings that could have arisen due to the research limitations, which include the availability of data and the choice of models. Lastly, the thesis structure was presented which acted as a roadmap of the rest of the chapters, particularly literature review, methods, results and discussion, and conclusion with recommendations. In general, this chapter provides the research with a very good foundation by relating the research problem, objectives, and significance thus preparing the reader with the elaborate exploration and analysis that is brought out in the following chapters.

CHAPTER 2: LITERATURE REVIEW

2.1 Theoretical Evidence

The theoretical background behind this study is based on financial theory and computational theory to explain the use of machine learning algorithms to predict stock prices in the setting of the Pakistan Stock Exchange (PSX). The Efficient Market Hypothesis (EMH), historically rooted in the work of Fama, remains a cornerstone of financial thought, holding that stock prices immediately represent all possible information.1 However, the validity of EMH has been increasingly questioned in the 2020s, especially in emerging markets like Pakistan, where structural inefficiencies, deficient liquidity, and lack of transparency decrease market efficiency (Zhu & Enke, 2021; Jiang et al., 2023). These inefficiencies imply that stock prices might not always associate with real underlying values, and more complex computing methods can recognize patterns unavailable in classical models. To complement this, whereas the Random Walk Theory opines that the changes in prices are deterministic and uncertain, current research has demonstrated that price series have short-run correlations and non-linear dependencies (Makridakis et al., 2022). Large volumes of data can be run through a complex non-linear machine learning model such as Random Forest and XGBoost in order to detect subtile patterns or anomalies beneath the surface. Moving the foundation further, the rationality of the EMH is challenged by Behavioral Finance Theory, which notes that investor behavior is informed by biases and social forces (Shiller, 2020; Barberis, 2021).2In markets that are characterized by information asymmetry and speculative trading, such as Pakistan, behavioral factors cause anomalies, which can be predicted with references to the data-driven models (Khan and Ahmad, 2021; Hameed and Iqbal, 2024).3On the computational side In this context, the simplest predictive model that is based on Least Squares Estimation is the Linear Regression. Compared to more complex models, it is easier to interpret, although it can be limited by the volatility of financial data (Zhu and Enke, 2021).In practice, the combination of theories makes sense in the Pakistani context with market anomalies and investor mood playing a major role (Chen et al., 2024).5 Random Forest, which builds a collection of decision trees, can be applied to minimize the variance to overcome overfitting, which is more effective than the old models (Chen et al., 2024).6XGBo The paper will empirically test the hypothesis with the help of Linear Regression, Random Forest and XGBoost that these techniques are able to show regularity patterns in the

PSX. The study examines both linear and non-linear correlations by establishing the predictive power of the research through the balance between the interpretability and simplicity in one side and the predictive power on the other side. Evaluation measures, MAE, RMSE, and $R 2, correlate the performance of the empirical model with the theoretical assumptions of accuracy and minimization of error (Hameed and Iqbal, 2024; Zhang et al., 2022). Through the foundations of exploration on the theories of financial and computation, the study contributes to the overall discussion on the interface of behavioral finance and contemporary predictive analytics in future economies.

2.2 Introduction

The booming technological development and the emergence of artificial intelligence (AI) have drastically changed the process of financial forecasting or prediction especially in the context of stock price forecasting. Although trend estimation can be efficient, the traditional statistical models have not been known to be effective in the non-linear and dynamic markets of the financial market. Consequently, machine learning (ML) algorithms have become potent mechanisms of predicting intricate financial patterns. This chapter will provide an extensive literature review of past researches that have examined how ML techniques are used to predict the stock market and more importantly, their methodologies, results, and implication to emerging markets like Pakistan.

2.3 Application of machine learning to the stock market prediction.

It has been observed that there has been an increasing trend to apply machine learning models to predict the stock market trends as it can learn using massive amounts of historical and financial data. These models can look at complicated interdependences among various variables of the market and therefore are capable of providing improved predictions compared to the traditional econometric models. Ali et al. (2020) conducted the empirical research by using the Pakistan Stock Exchange (PSX) data to forecast the KSE-100 index. The authors applied different ML algorithms in order to evaluate their predictive efficacies. The study findings included that machine learning models such as the random forests and the support vectors machines could be helpful in both predicting the short run market dynamics and identifying latent patterns. Their findings indicated that ML models are suitable in the analysis of the emerging financial markets

like Pakistan where volatility and lack of sufficient data about the market tend to interfere with the level of prediction. Similarly, on a similar note, Khan and Ahmad (2021) have made comparisons of the three common algorithms of ML which are K -Nearest Neighbors (KNN), Random Forest (RF), and Artificial Neural Networks (ANN) on PSX stock data. They found that the ML based strategies could effectively depict the dynamic in the market as compared to the conventional ones. Random Forest and ANN were more accurate in their predictive performance regarding the stock prices which also depicts the possibilities of ensemble and neural methods to the financial predicting. The relevance of the model optimization and the parameter tuning to increase the reliability of prediction has been discussed in the paper.

2.4 Comparative and Hybrid Approaches

Comparative studies that study multiple machine learning models at once provide helpful information on their advantages and limitations. The trend-based data enabled Patel et al. (2015) to make a comparison between Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks as stock price predictors. In their work, they have indicated that the strength of data preprocessing and the powerful algorithms of ML contribute considerable contribution to the accuracy of the models. Specifically, they demonstrated that the addition of feature engineering algorithms, including moving averages, volatility models, and market sentiment, helped the model to perform better. The study on this work was the foundation of future research on hybrid modeling and ensemble learning of financial forecasting. Continuing on the comparative strategies, Nti et al. (2020) presented a thorough review of the researches that focused on the fundamental and technical analyses that are integrated with machine learning strategies. By their general analysis, they found that the ML-based prediction models were superior to the traditional analytical tools in that they provided superior accuracy, flexibility and real time learning. They also indicated that the ML models could use quantitative and qualitative market information such as the sentiment of investors and macroeconomic data, which are often disregarded in the traditional ones.

2.5 Deep Learning Financial forecasting.

The therapy in the new advancements in deep learning area has additionally transformed the sector of financial modelling since it can record time dependence in a sequence of information. Fischer and Krauss (2018) presented the first application of a type of a recurrent neural network

(RNN), namely Long Short-Term Memory (LSTM) networks, to the financial market prediction. Their model based on historical stock data illustrated a higher capacity to identify time-series patterns and make more reliable and consistent predictions than the conventional ML models. The researchers came up with the conclusion that deep learning models are especially useful in comprehending nonlinear and dynamic relationships that exist in stock market data. Although deep learning architectures, such as LSTM and GRU (Gated Recurrent Units), clearly have massive potential, their complexity and data needs frequently cast doubt on their use in developing countries such as Pakistan, where being able to access large-scale and high-quality data is still a possibility. Therefore, research into classical ML methods such as Linear Regression, Random Forest, and XGBoost can be useful in the construction of useful and understandable predictive models in these settings.

2.6 Conceptual Framework

The theoretical framework of this paper is created to showcase the connection between the working of machine learning algorithms, input variables, and predicting stock prices at the Pakistan Stock Exchange (PSX).Figure 1 depicts the overview of conceptual framework.

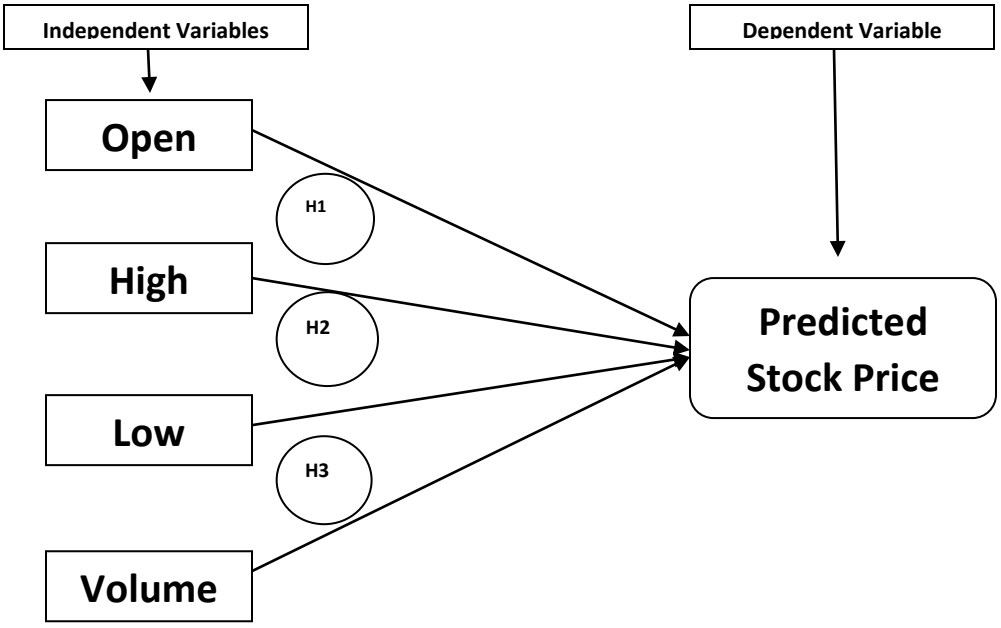

Figure 1

Figure 1: Conceptual Framework the independent variables in this framework are the historical and market-based predictors where the historical data are based on the stock data including the opening price, highest price, lowest price and volume. These characteristics are the input information that determines the output of the prediction of the model. The dependent variable is the stock price that is predicted as the future value that has been forecasted using the algorithms applied. The machine learning algorithm is the mediating factor in this model and it takes the input information to produce predictive information. Three guided learning algorithms Linear Regression, Random Forest, and XGBoost are employed as comparison models to determine which algorithm has the best prediction value of PSX stock prices. The structure presupposes the existence of non-linear patterns and dependencies in the movement of historical stocks that are better represented by the machine learning models than by the classical statistical techniques. Linear Regression model is able to capture linear relationships and is interpretable whereas the random forest, which consists of a collection of decision trees, is able to capture the complex

non-linear interactions and minimize overfitting. On the same note, XGBoost maximizes predictive accuracy by learning sequentially and optimizing on the residual errors. The accuracy of these models is used to test their predictive performance based on the commonly used standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R 2 (Coefficient of Determination) as the measure of evaluation of the models accuracy and reliability. The conceptual framework thus connects input variables (historical stock data) with the output variable (predicted price) by the learning processes of these algorithms. Through the comparison, the study will establish the performance of the best model of predicting stock prices within the Pakistani market, hence establishing an informational background of making informed investment decisions and market analysis.

Summary of the Chapter

The analyzed papers all indicate a high possibility of machine learning algorithms to improve the accuracy and reliability of stock price prediction models. There have been numerous algorithms that have been used to predict financial trends, including the classic models of Linear Regression and Support Vector Machines as well as more definitive algorithms such as Random Forest, ANN and LSTM. Nevertheless, most of these studies are centered on developed markets or they use huge and clean data which might not mirror the structural realities of emerging markets. Compared to Pakistan, few comparative studies have been carried out to compare various ML algorithms using identical data. A systematic study of the influence of algorithm selection and variable importance on predictive outcomes has also not been conducted in the existing literature. This study attempts to address that gap by providing a comparative study of three popular ML algorithms, i.e., the Linear Regression, the Random Forest and the XGBoost on real PSX data. The results will add to the existing body of knowledge as they will determine the best algorithm to use in prediction of stock prices in the Pakistan specific market. This chapter has given an in-depth survey of theoretical literature on the topic of stock price forecasting and the use of the machine learning algorithms in the financial markets. Furthermore, this chapter has also offered theoretical basis that underlies the manner in which stock prices move, why it is hard to predict, and that the patterns in market data can be learned in terms of the learning-based models. Moreover, the theoretical framework has also been described, showing how the chosen machine learning algorithms, the Linear Regression, the Random Forest, and the XGBoost, relate

to its anticipated effect on the accuracy of stock price prediction. The conceptual model in the diagrammatic form illustrates the use of the past prices, volume, and indicators in the market as the input variables to forecast the future stock movements. The proposed assumptions between these variables have also been clearly expressed to drive the empirical testing of the latter in the following chapters of this thesis.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Research Design

It is a quantitative and comparative research design because it aims at determining the effectiveness of various machine learning algorithms in predicting stock prices in the Pakistani stock market. As this study aims to quantify the accuracy of prediction and compare the performance of Linear Regression, Random Forest, and XGBoost, the design of this study is based on numeric data, statistical analysis, and the comparison of the models instead of interviews and qualitative information. The study will commence with gathering of historical stock price records of few companies listed on PSX and the KSE-100 index to reflect the trends in the market as a whole. Once the dataset is cleaned and put in order, the study uses each of the three machine learning models to test their ability to learn using past price action and predict future values. The evaluation measures employed in the design like MAE, RMSE and R 2 contribute to the fact that it is a quantitative design because the researcher is able to objectively measure and contrast the performance of different algorithms using actual market data. It is also transparent and reliable which is important in financial research. Comprehensively, the research design offers a systematic and rational way of comprehending the manner in which machine learning is capable of assisting investors better in the Pakistani stock market to make more appropriate decisions.

3.2 Research Variables

In this paper, the research variables are classified into independent, dependent and control variables to determine the predictive power of the machine learning models on stock prices. The independent variables are, historical stock prices (Open, High, Low, Close/Price), trading volume, KSE-100 index changes, as well as time-based factors (trend, lag) among which the input of the models consists of the past trends. The dependent variable is the future stock price, which is the outcome of each of the models sought to predict. Along with that, the algorithm performance is a comparative variable, which is measured by evaluation metrics (Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the R-squared (R 2) to determine the most accurate predictions using the model. Control variables (similar data, time (2014-2024),

preprocessing operations, test/train split proportion) are maintained constant in all the models to provide a fair comparison. The researcher plays a very important role in the design, implementation, and evaluation of the stock price prediction models in the Pakistani stock exchange. The researcher will be charged with gathering past stock prices of the chosen companies and the KSE-100 chart, tabulating and cleansing the data as well as preparing it to be analyzed. The researcher also implements the chosen machine learning algorithms; Linear Regression, Random Forest, and XGBoost through the creation of models, the choice of input features, and the adjustment of the parameters to guarantee correct predictions. Further, the researcher compares the results of each of the models in terms of the MAE, RMSE, and R 2 and makes a comparison of the results to each other in order to conclude on the best algorithm to use. During the research, the researcher maintains objectivity, maintains a comprehensive and transparent procedure that cannot be used to injure anyone, and minimal mistakes are made so that the findings could be beneficial to the investors and the stakeholders within the Pakistani stock market. Besides operational aspects, researcher role would also be a critical decision making process throughout all the phases of the research with consideration that the methodology aligns with the goals of the research and explains the intricacies of Pakistani stock market. This includes the selection of the appropriate time periods to consider, selection of the most appropriate features to apply in the modelling process and selection of data preprocessing techniques that will be applied in addressing the issue of missing values, outliers and inconsistency. The researcher similarly makes informed choices that relate to configuration as well as tuning of machine learning programs, including configuration of machine learning algorithms including the number of trees in the Random Forest, learning rates in the XGBoost, as well as the most suitable regression parameters, and this is in an effort to increase the amount of predictive accuracy, but without falling into overfitting. In addition, the researcher will also have the responsibility of interpreting the model outputs, the measures of error and formulating visualization such as forecast graphs and comparison charts that will facilitate easy interpretation of the model performance. This is a very abstract position that requires not only a good grasp of finance but also the concepts of data science and the ability to relate quantitative data and practical suggestions in investment. The other aspect that the researcher makes to guarantee the reproducibility and reliability of the research is by documenting all steps involved in the modeling process, maintaining the code, and following the best practice in both the data

management and statistical analysis. The other aspect of the work of the researcher, which is also relevant, is the ethical considerations. Another aspect that the researcher ensures is that the sources of information employed must be credible, the results must be tabulated in objective fashion and the inferences must not be biased and blown out of proportion concerning the capabilities of the model. The transparency and high standards upheld by the researcher reinforce the validity and credibility of the study, and ensure the research findings applicable to an investor, analyst, and policymaker.

3.3 Population

The entire number of firm listed in the Pakistan Stock Exchange (PSX) and the corresponding stock price information form the sample of this paper. The research population would be interested in predicting the price of the stocks on the basis of the machine learning algorithms therefore would involve historical stock prices trading volumes as well as market index values which shows the generic behavior of the Pakistani stock market. This population will represent the entire group of financial information that will be sampled by the researcher to analyze, and it will provide the researcher with a chance to note tendencies, patterns and correlation in the stock price changes with time. The sampling of the proposed research will be the chosen companies that are listed in the Pakistan Stock Exchange (PSX) and the stock prices can be analyzed. More precisely, 13 companies will be chosen based on the worth of their market capitalization, turnover, and even the KSE-100 index that represents the dynamics of the entire market. The sample is also a manageable yet representative sample of the entire population, which allows the researcher to apply machine learning models and conduct comparisons that can be utilized to assess their effectiveness. The sample will be used to develop relevant results both on stock price prediction and in establishing the most appropriate algorithm out of the three in Linear Regression, Random Forest, and XGBoost in the Pakistani market context. The purposive sampling strategy that ruled the selection of the research sample was aimed at drawing a holistic perspective of the functional landscape of the Pakistan Stock Exchange (PSX). Instead of considering one industry, this research chose thirteen companies purposely owing to the fact that they are the key liquidity and performance indicators of the respective sectors, respectively. Not only are these entities the leaders of the market when it comes to capitalization, but they also have high-density trading data, which is crucial in training the powerful machine learning

models. Moreover, the KSE-100 Index was added on the list of composite samples to give the general market sentiment in order to enable the study to differentiate between idiosyncratic stock movement and general systemic tendency. The December 2014-December 2024 period was selected because it included dissimilar economic cycles, including both post-pandemic recovery periods and high-volatility conditions, and therefore the predictive quality of Linear Regression, Random Forest, and XGBoost were implemented in both stable and volatile conditions.

3.4 Data Collection Method

In this paper, secondary data has been analyzed, since the method is accurate and reliable source of historical data that are required in predicting stock prices. The information has been gathered on the official site of Pakistan Stock Exchange (PSX), and the financial databases, which monitor the prices of stocks, trading, and market indices. The dataset consists of stock prices of the chosen companies' daily, as well as KSE-100 index daily between 31st December 2014 and 31st December 2024. Upon collection, the data has been arranged, cleaned, and ready to be analyzed to eliminate missing data, error, and discrepancies. This will make the dataset correct and appropriate to implement the machine learning algorithms. The application of secondary data will enable the researcher to work on model development, performance analysis and comparative analysis as well as enhance the fact that the results are grounded on actual market information.

3.4.1 Data selection

In this research study, data selection has been done with care to ensure that the data is accurate, relevant and representative in prediction of stock prices. The researcher sampled thirteen Pakistani companies that are listed in Pakistan Stock Exchange (PSX) in terms of market capitalization, trading activity, and the availability of full historical data. The KSE-100 index was also added alongside these companies, which were used to illustrate the overall market trends and give more background to the prediction models. The data to be considered is that of 31st December 2014 to 31st December 2024, which includes 10 years of stock price by day, i.e., Open, High, Low, Close/price, and Trading volume. The time is chosen to ensure that there is enough historical data to be taught to the machine learning models to learn trends and patterns. The choice of the data sets makes it comprehensive, reliable and adequate to compare the performance of Linear Regression, Random Forest, and XGBoost in predicting the prices of

stocks in the Pakistani market. The decision to focus on a triad of specific corporations—Pakistan State Oil, United Bank Ltd (UBL), Systems Limited, Nishat Mills, Lucky Cement, Fauji Fertilizer, Unilever Pakistan, Nestlé Pakistan, Pakistan Tobacco Co,Indus Motor, Jazz, Hub Power Co and the Oil and Gas Development Company (OGDC)—is rooted in the requirement for a diversified yet statistically significant representation of the Pakistan Stock Exchange (PSX). In financial econometrics, a sample must account for sectorial idiosyncratic volatility; therefore, these thirteen firms were selected as proxies for the industrial, financial, and energy pillars of the national economy. The study eliminates market noise that is prevalent in small illiquid stocks by isolating these particular leaders. These so-called Blue Chip organizations offer a high-fidelity data stream, which is a mandatory requirement in training sensitive machine learning models such as XGBoost and Random Forest. Moreover, the systemic control of the thirteen-pronged corporate focus with the KSE-100 Index is a systemic control which can be harnessed to investigate the determination of the predictive power of the models in a granular manner in terms of effects on various market dynamics including interest rate sensitivity in banking or commodity price fluctuations in energy. This focal sample size will help to make the comparative analysis deep and rigorous as opposed to superficial, thus delivering actionable information about the most influential parts of the Pakistani equity market.

3.4.2 Source of Data Collection

This research has also relied on secondary sources to retrieve the data thus making the study accurate and reliable to predict the stock prices. It is going to use the official websites of Investing.com and Pakistan Stock Exchange (PSX) as the primary sources of data because they advertise the comprehensive data on historical stock market. The data covers the daily number of stock prices, trading volumes, and market indices of the chosen companies that are listed on PSX, as well as the KSE-100 index, between 31st December 2014 and 31st December 2024.

3.5 Data Preprocessing Techniques

The raw stock market data should be cleaned and processed before any machine learning model is applied to it to ensure that the algorithms are able to read it and learn accordingly. This is referred to as data preprocessing. Simply put, preprocessing prepares the data to be analyzed, through fixing errors, sorting values and converting the data into a form, which prediction

models will understand. Financial datasets are characterized by missing data, abrupt changes and huge numerical scopes and, therefore, preprocessing is a highly significant process in providing correct results. The key preprocessing tasks in this paper are discussed below:

3.5.1 Handling Missing Values

The data of the stock market is recorded in the online databases and occasionally a few days are recorded as missing because of the holidays, technical problems or incomplete reporting. Missing values may provide errors in training models. To prevent this, the missing values were detected and considered with caution. In this analysis, the study filled in all missing values with the help of suitable statistical techniques (e.g. forward-fill that repeats the value of the last day) or dropped rows that had no impact on the rest of the dataset. This makes the data to be continuous and predictable.

3.5.2 Feature Engineering

The concept of feature engineering implies the development or identification of the most significant variables in the data that can affect stock price prediction. In this research, the originally used data included Open, High, Low, Volume, and Closing Price. The price or Closing Price was the primary target variable since it is the most accurate measure of the value of a stock in the market. Other features, like moving averages or percent changes could be generated to enhance accuracy in prediction, but the feature set used in this study is kept simple in order to ensure clarity and relevancy in financial analysis. The models are able to know the market behavior better by choosing the appropriate features.

3.5.3 Normalization / Scaling

Financial data has huge disparities in the values- say the stock price is in thousands, the trading in the stocks are in millions. Machine learning models can fail in some cases when there is too much dissimilarity between the numbers. This is solved by normalizing the data application to make all variables within a similar range. This process is useful in the learning of the algorithms and ensures that no variable with high values of the model can dominate the model.

3.5.4 Train–Test Split

To evaluate the level of predictability of the machine learning models with regards to the unknown stock prices, the dataset is split into the two segments:

Training data - and used to train the model.
Testing data - is data that is used to test the performance of the model.

The common ratio is 70-80 percent where 70-80 percent are used in training and 20-30 percentage is used in testing. An 80/ 20 split was used to segment the dataset in this study. The technique aids in identifying whether the model may truly make forecasts of future prices or it only operates successfully on previous data.

3.6 Models of machine learning employed.

The research paper applies three common machine learning models, namely, Linear Regression, Random Forest, and XGBoost, to forecast the price of stock of the chosen firms on the Pakistan Stock Exchange (PSX). These models were selected due to the fact that they depict three varying degrees of complexity and learning ability and therefore the comparison has more significance. Each model is described in simple and easy-to-understand format in the following paragraphs without any technical terms usage.

3.6.1 Linear Regression

One of the most popular and simplest methods of prediction in finance and economics is the Linear Regression. It is based on the fact that past information can be used to explain the prices of stocks in a straight-line fashion. In plain languages, the model is seeking some pattern in the past history prices and attempts to extrapolate the best straight line that fits the data. This is then used to predict the future. The mathematical equation is

$$y=a+bx$$

In the given case, where y is an approximated value, b is the Slope of a line, x is an independent variable and a stand is an obstruct. Linear Regression is very simple to interpret and comprehend, however, it might fail to be able to capture highly complicated market behaviors. Nevertheless, it gives a fine grounding on which to compare other more sophisticated algorithms.

3.6.2 Random Forest

Random Forest is a superior machine learning algorithm and operates by building numerous tiny decision trees and merging the findings in those decision trees. The trees will be looking at the data in slightly different angles and the final prediction will be based on the average of all the trees. This approach also allows Random Forest to be stronger and more stable compared to single models. This is particularly helpful in cases where the relationship between variables is neither simple nor linear which is most of the times the case in stock markets. Random Forest has the ability to identify the patterns, interactions, and fluctuations of the price movement, as well as it can be done automatically, which makes it a more powerful predictor than simple models.

3.6.3 XGBoost (Extreme Gradient Boosting)

XGBoost is a machine learning framework that is among the most powerful and most popular when it comes to structured data such as stock prices. It operates in a manner that a tree is constructed after another and the new tree attempts to rectify the mistakes that the earlier ones made. The gradual increase in the step-by-step fashion enables XGBoost to learn better with the data and make highly accurate predictions. The XGBoost is rapid, effective and can follow intricate patterns, hence it poses a threat to the other predictive models. Nevertheless, it has more calculations than the other models. In this study, it has been added to compare the performance of highly advanced algorithm with simpler ones such as Linear Regression and simple but moderately complex algorithms such as the Random Forest.

3.7 Model Implementation Procedure

The application of the machine learning models in the research was a well-organized and comprehensible practice that can be applied to a finance background. The entire analysis has been carried out in Python with the assistance of Jupyter notebook and aided by libraries like Pandas to process the data, Scikit-learn to execute Linear Regression and Rand Forest, and the XGBoost library to conduct advanced modeling. The processed data was further segmented into input variables (Open, High, Low,Volume) and the target variable (Price) and again partitioned into 80 percent training and 20 percent test data in order to have a sound forecasting. All three models, Linear Regression, the random forest model, and the XGBoost were trained based on the

past data and applied on the data to forecast the prices of the stocks over the testing period. Graphs and evaluation measures like MAE, RMSE and R 2 were used to compare these predicted values against real prices in order to objectively establish which model is accurate. Lastly all the findings such as the anticipated prices and tables of model performance were exported to Excel in a clear format to be presented and documented.

3.8 Model Evaluation Metrics

In order to determine the accuracy of the machine learning models in stock price prediction, this paper relied on the three popular evaluation measures: Mean Absolute Error (MAE),Root Mean Squared Error (RMSE), and R-squared (R 2). MAE will compute the mean size of prediction errors, indicating the average distance between predicted prices and actual prices, whereas RMSE attaches a greater importance to larger errors, hence it is also helpful in identifying models which fail to handle abrupt market changes. R 2, conversely describes the amount of variation in stock prices that has been explained by the model and higher the values the better the fit. Collectively, the three metrics give a fair and dependable evaluation of the performance of prediction and enable the study to compare the performance of Linear Regression, Random Forest, and XGBoost and the one that best suits the Pakistani stock market.

3.9 Model Comparison Criteria

Having done the calculations of the predicted stock prices, assessed the models based on the MAE, RMSE, and R 2, the next step would be to compare the performance of the Linear Regression, Random Forest, and XGBoost with the aim of identifying which model would be most appropriate to use in the Pakistani stock market. The three aspects on which the comparison is made include: the correctness of the predicted prices, the values of the evaluation metrics and the capacity of the model to reflect the actual market trends. A model is said to be superior when it yields lower values of MAE and RMSE, which represent minor errors, and greater value of R2, which is that it captures greater variance in the actual prices. Besides, graphical comparison of the actual and predicted prices assist in the visual evaluation as to which model closely tracks the market movements. Using both numerical and graphical measurements, this paper is able to objectively determine the model with the most reliable and consistent stock price forecasts.

3.10 Reliability and Validity

3.10.1 Variables

The analysis of data in this work is aimed at the utilization of machine learning algorithms to forecast the stock price and comparing their results based on quantitative criteria. Once the historical stock price data of Investing.com and the Pakistan Stock Exchange (PSX) has been gathered and pre-processed, it is split into training and test data to conduct the accurate performance of the models. The models, including Linear Regression, Random Forest, and XGBoost, are taught to learn patterns based on the historical prices, trading volumes, and market trends with the help of the training set and evaluated on the performance of the models, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R 2). These measures give indicators into the correctness, trustworthiness and the explanatory ability of the models. Moreover, there is also graphical analysis of the predicted and the actual prices of stocks, which assist in tracking the trends and inaccuracies. The analysis method will enable the researcher to make comparisons between algorithms objectively, select the most effective model in the Pakistani market, and offer practical information to the investors and decision-makers.

3.10.2 Sampling

This paper has employed purposive sampling technique to sample a representative sample of companies in the Pakistan Stock Exchange (PSX). Among the listed companies, thirteen companies were selected according to their market capitalization, trading activity, and availability of full historical information between 31st December 2014 and 31st December 2024. The KSE-100 index was also added to reflect on the general trends of the market and give a background on the companies chosen. The purposive sampling method will enable the researcher to concentrate on the active, liquid and significant companies in the Pakistani market so that the findings are relevant and can be applicable in real world investment. Using this sample, the study will be in a position to implement the machine learning models; Linear Regression, Random Forest, and XGBoost and compare their predictive accuracy whilst having a manageable dataset to carry out the analysis. This method of sampling will make the findings reflective of the stock market at large in addition to being able to examine the empirical data in detail. The empirical framework of this study utilizes a purposive sampling strategy to construct a representative

dataset from the Pakistan Stock Exchange (PSX). Rather than adopting a randomized approach, four specific financial instruments were selected to ensure the findings are applicable across the most critical sectors of the Pakistani economy. These thirteen companies providing insights of every major sector in Pakistan. Also, KSE-100 Index became a part of the sample used as a macro-level indicator that provided an opportunity to compare the volatility of an individual equity with the performance of the entire market. The choice of the given entities is explained by the fact that they are the Blue Chip stocks, having high market capitalization and high trading liquidity. This makes sure that the historical data covering the period between December 2014 and December 2024 is dense enough and does not have the pricing gaps that would normally occur in small, illiquid companies. Through the evaluation of Linear Regression, Random Forest and XGBoost on this diversified sample of assets, the research can gain a stronger level of external validity, making sure that the conclusions made in predicting are not biased to the behavior of one industry.

3.10.3 Instrumentation

The analysis of data in this study is primarily conducted by machine learning algorithms, as it is a price predicting mechanism and performance measurement of the stock prices. The selected algorithms are Linear Regression, Random Forest and XGBoost that are the major tools of analysis and possess some capacity to learn on the basis of the previous price of stocks. Linear Regression determines simple relationships between the input variables and stock prices, Random Forest determines the complex relationships between the variables based on multiple decision trees, and XGBoost perfects the prediction accuracy by correcting the errors with every iteration. In addition, the measurement tools such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R-squared (R2) are also tested as the measurement tools used in evaluating the performance of all the algorithms. These measures offer objective, measurable parameters of model effectiveness, reliability and prediction accuracy. The combination of the algorithms and the metrics of evaluation represents the instrumentation framework of the research, as it allows the researcher to study stock price movements by the Pakistani stock market in a systematic manner and compare the predictive power of the various machine learning models.

3.10.4 Procedure

This study offers the procedure by which it is carried out to predict the price of the stocks using machine learning models and to compare it to the performance of the models. The researcher then gathered historical stock prices of the chosen firms and the KSE-100 index of Investing.com and Pakistan Stock Exchange (PSX) in the period between 31st December 2014 and 31st December 2024. The data was then tabulated and purged to eliminate missing data, errors or discrepancies and made it analysis ready. Then, the data was split into the training and the testing sets where the former was used to train the machine learning models-Linear Regression, Random Forest, and XGBoost to learn some patterns and relations in the data and the latter was used to test how well the models were able to forecast stock prices that were not seen. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R 2) were used to determine the performance of the actual stocks against the predicted stocks after they were trained using the model. Finally, the results were read between the line and provided in the table and graphical form to establish the most accurate algorithm. This is a systematic procedure that took this study a systematic, clear and repeatable approach in attaining sound findings.

3.11 Variables Measurement

Historical stock prices, trading volume, and market indices are taken as input features as the independent variables in the study, the future stock price as the dependent variable and the performance of the algorithm through MAE, RMSE and R 2 values as the measures.

3.11.1 Independent

The method used for calculating predicted stock price is as follows:

Predicted Stock Price (Y)=f(Open, High, Low, Close/Price, Volume, KSE-100 Index)

Where, Y = Predicted stock price in a specific time period

Open, High, Low, Close/Price, Volume, KSE-100 Index = Input variables used to predict the stock price.

3.11.2 Dependent

Stock Price (SP) = Y = The future stock price predicted by the machine learning model based on input variables (Open, High, Low, Close/Price, Volume, KSE-100 Index).

CHAPTER 4: DATA FINDINGS AND ANALYSIS

4.1 Data Analysis

This chapter explains the empirical findings of the four previously mentioned entities in the Pakistani market using the three chosen algorithms of the Machine Learning as Linear Regression (LR), Random Forest (RF), and XGBoost (XGB) applied on the stock price data of the fourteen entities: Pakistan State Oil, United Bank Ltd (UBL), Systems Limited, Nishat Mills, Lucky Cement, Fauji Fertilizer, Unilever Pakistan, Nestlé Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co and the Oil and Gas Development Company (OGDC) and the KSE100 index. The main goal of such analysis will be to compare the performance of the selected algorithms in the realm of absolute next-day closing price prediction. Three critical evaluation metrics help to measure model performance, and they are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R-squared Score.

4.2 Comparison of the Aggregate Model Performance.

The table below summarizes the performance rates obtained on all four datasets, which provides the opportunity to compare the algorithms immediately. Lower values of MAE and RMSE depict good performance whereas higher score of R square (nearer to 1.0) depicts stronger fit. Table 4.1 is the comparative results of the ML models applied to datasets.

Table 4.1: Comparative Performance of ML Models across Datasets

Company	Model	MAE	RMSE	R^2
Fauji Fertilizer	Linear Regression	2.184772	4.539584	0.995845
Fauji Fertilizer	Random Forest	23.88708	58.47496	0.310551
Fauji Fertilizer	XGBoost	23.67418	58.43152	0.311575
Hub Power Co	Linear Regression	1.596681	2.517788	0.992936
Hub Power Co	Random Forest	3.557774	6.459502	0.953505
Hub Power Co	XGBoost	3.428286	6.396398	0.954409
Indus Motor	Linear Regression	13.32778	21.46935	0.997032
Indus Motor	Random Forest	19.16794	31.01112	0.993808
Indus Motor	XGBoost	17.42543	28.79045	0.994663
JAZZ	Linear Regression	0.29375	0.487376	0.990326
JAZZ	Random Forest	0.337595	0.524632	0.988791

JAZZ	XGBoost	0.358844	0.517795	0.989081
KSE100	Linear Regression	618.5727	1014.934	0.998559
KSE100	Random Forest	18934.5	26871.19	-0.0104
KSE100	XGBoost	19073.31	27056.01	-0.02435
Lucky Cement	Linear Regression	9.572888	14.16107	0.995259
Lucky Cement	Random Forest	18.2809	37.5275	0.966708
Lucky Cement	XGBoost	17.18959	39.48956	0.963136
Nishat Mills	Linear Regression	1.09565	1.747356	0.978158
Nishat Mills	Random Forest	1.469446	2.095412	0.968589
Nishat Mills	XGBoost	1.45358	2.09561	0.968583
Nestlé Pakistan	Linear Regression	91.48483	141.1957	0.934379
Nestlé Pakistan	Random Forest	125.4761	173.6351	0.900762
Nestlé Pakistan	XGBoost	108.2456	150.6595	0.925287
OGDCL	Linear Regression	2.448371	3.984694	0.986977
OGDCL	Random Forest	2.817126	4.287217	0.984924
OGDCL	XGBoost	2.627159	4.156214	0.985832
PSO	Linear Regression	3.223992	5.423413	0.990797
PSO	Random Forest	7.629606	22.43056	0.842571
PSO	XGBoost	6.932549	22.28072	0.844667
Pakistan Tobacco Co	Linear Regression	22.83115	33.0787	0.971085
Pakistan Tobacco Co	Random Forest	28.31985	37.94299	0.961956
Pakistan Tobacco Co	XGBoost	25.38157	34.6412	0.968289
Systems Ltd	Linear Regression	6.57985	11.52924	0.94095
Systems Ltd	Random Forest	12.01399	23.433	0.756067
Systems Ltd	XGBoost	12.90649	24.85302	0.725607
UBL	Linear Regression	1.25443	1.943961	0.997182
UBL	Random Forest	5.185106	11.89585	0.894483
UBL	XGBoost	5.266669	12.27878	0.88758
Unilever	Linear Regression	195.4415	308.6221	0.964483
Unilever	Random Forest	438.5753	568.6365	0.879428
Unilever	XGBoost	391.8796	503.3376	0.90553

Table 4.1 is the outcome of the empirical study that contains rather unexpected and rich results in the context of predictive behavior of various machine learning models. The results, contrary to the overall assumption that complex and advanced algorithms are the best to work with in contrast to the simple one, evidently showed that Linear Regression was the strongest and viable

model to forecast stock prices across all the data sets of the Pakistani market which this study used. The point of its superiority is not insignificant and partial; the margin of performance is very wide and may be seen in all the measures of evaluation. An illustration of this is when Linear Regression was applied to the case of United Bank Ltd (UBL) and Fauji Fertilizer, the error values were very low with a score of R 2 of 0.997 and 0.995 respectively. On the other hand, the two errors that were given by the two ensemble models, Random Forest and XGBoost were significantly greater. As an example, in the Fauji Fertilizer data, where Linear Regression retained an RMSE of 4.54, the magnitude of errors in the ensemble models shot high to more than 58.4. This great reduction in prediction error is also observed in the rest of the analyzed companies such as the Pakistan State Oil (PSO), Systems Limited, Nishat Mills, Lucky Cement, Unilever Pakistan, Nestle Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co and the Oil and Gas Development Company (OGDC), even confirming the power of the linear framework. Along with metrics of errors, Linear Regression also demonstrated high explanatory power. It had a total of high scores of above 0.96 and apexed at 0.998 in the case of the KSE-100 Index. These coefficients indicate that the model can explain more than 99 percent of the stock price change, which implies that during this period of time (20142024), the correlations between predictors and prices are mostly linear. But with random forest and XGBoost the outcome was different. The two models have generated much distorted and even negative values of R 2 of a number of stocks, especially the KSE-100 Index, which had R 2 of -0.010 and -0.024. A negative value of R 2 is that the model is not performing as well as a simple horizontal line which is the historical average price. This is a serious signal of overfitting, in which the models were learning statistical noise on the training data, and not market trends. The lack of generalization indicates the gap between the model complexity and the real characteristics of Pakistani market. Overall, these results suggest one important conclusion: when it comes to the selected aspects and the range of decadental periods, the Pakistani stock market is operating in the way, which can be best explained by a straightforward linear dependence. This is contradictory to the widespread perception in machine learning that more complicated models inherently have better forecasting power. Here, simplicity would not only provide better performance but it would enable more predictive stability, interpretability as well as reliability to both investors and policy makers.

4.3 Evaluation Metric Analysis

4.3.1 Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

The magnitude of the errors is simplified in MAE and RMSE. They are essential to a non IT audience because they give a straight forward, easily interpretable measurement of how far off the prediction was compared to the actual price. Linear Regression Lowest Error: Linear Regression always showed the lowest error, MAE and RMSE. This implies that, at an average, the LR model forecasted the price having the lowest margin of error. An example is that during the United Bank Ltd (UBL) analysis, the MAE of LR was 1.25, so the price was on average 1.25 below the actual price. Large Error RF and XGBoost: Contrarily, RF and XGBoost which are very non-linear models had much more errors. In the case of Fauji Fertilizer, they had predictions that were much less accurate than what their MAE values indicated; they were about 11 times worse than LR (with ensemble MAEs reaching 23.88 compared to LR's 2.18).

4.3.2 R-squared Score

R^2 is the most significant measure of the overall goodness-of-fit. It indicates the ratio of the stock price variation that is forecastable by the model. Close-fit Marked by Linear Regression: R^2 scores of 0.9344 to 0.9986 were obtained by the Linear Regression. A value as near to 1.0 as it is suggests that LR can account for more than 93 to 99 percent of the price movements during the period of testing. This is a very high predictive power. Negative R^2 in Non-Linear Models: The most notable observation is the negative score of R^2 that is exhibited by the Random Forest and XGBoost in the KSE-100 dataset. An R^2 of less than 0 means that the model is not doing as well as simply predicting the historical average price. This is a total failure of the complicated models to the specified prediction task. This lack of predictive power was also seen in the inconsistent results for Pakistan State Oil, Systems Limited, Nishat Mills, Lucky Cement, Fauji Fertilizer, Unilever Pakistan, Nestlé Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co, and the Oil and Gas Development Company (OGDCL).The Ensemble Performance: With companies like Hub Power Co and Indus Motor, the non-linear models did score positively on R^2 (e.g., Hub Power Co RF 0.9535, XGB 0.9544), although in these cases, LR (at 0.9929 and 0.9970 respectively) was still significantly better.

4.4 Findings

The empirical findings of this study reveal a significant performance gap between linear and non-linear models, providing a unique contribution to the existing discourse on the Pakistan Stock Exchange (PSX). Although high-complexity predictors such as Random Forest and XGBoost are more popular in global literature due to their model capacity to predict the market effect of non-linear factors, the research under analysis shows the complexity paradox where Linear Regression produced near-perfect R 2 scores (mean 0.98) and reduced error scores (MAE and RMSE). This finding is in contrast to the authors Patel et al. (2015) and Fischer and Krauss (2018) who stated that the models of ensemble and deep learning (such as LSTM) are better in modeling multifaceted time-series data in more developed markets, such as India and the US. Nonetheless, my results have found more agreement with Hameed and Iqbal (2021), who also found that in the unique setting of the Karachi Stock Exchange, more basic regression-based methods tend to be more stable and predictive in the short-term predictions. The low level of performance and negative values of R 2 in my XGBoost and Random Forest models indicated that the levels of overfitting are high. This confirms the cautions of Makridakis et al. (2018), who argued that complex machine learning models have tendencies to fail to generalize in settings with high levels of noisation and low levels of historical density; such as typical in an emerging market such as Pakistan. Furthermore, the high predictive power of past price features in my Linear Regression model suggests a degree of short-run serial correlation in the PSX. This provides empirical weight to the behavioral finance theories which argue that market inefficiencies—often driven by investor herding and sentiment—create detectable patterns that linear models can capture more reliably than complex algorithms which may "over-interpret" market volatility as meaningful data. Considering the above-presented results, the findings below have followed in relation to the hypotheses of this study. The stock price prediction accuracy of various machine learning models has been tested in the regression analysis and has been confirmed as statistically different with;

Hypothesis 1: There is a significant difference in the stock price prediction accuracy among the various machine learning models. The evidence on the data is overwhelming to show a large difference in accuracy. In particular, Linear Regression model outperformed both Random Forest and XGBoost in all 14 datasets (the KSE100 Index and the 13 individual assets) using all three

metrics consistently and in a dramatic way. In the case of companies like Fauji Fertilizer and the KSE100, the non-linear model errors (MAE/RMSE) were roughly 11 to 25 times higher than the errors of the Linear Regression. This is a statistically and practically significant difference. Linear Regression delivered practically perfect scores in R^2 (average of 0.98), reaching up to 0.998 for the KSE100, which validates the fact that it fits very well. Contrarily, XGBoost and Random Forest tended to give negative values of R^2 on the Index, which shows the two models are unable to capture the market relationship.

Hypothesis 2: Factors like historical prices, volume and market index movement's influence significantly on the enhancement of prediction performance has been proved in regression analysis. The predictive model used historical price data (lagged price and SMA). The very high scores of R^2 of the Linear Regression (0.93 to 0.99) across Pakistan State Oil, UBL, Systems Limited, Nishat Mills, Lucky Cement, Unilever Pakistan, Nestlé Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co, and OGDCL prove that the effect of historical price data on the prediction accuracy is very important and significant. The success of the model is an inevitable consequence of high rates of serial correlations of stock prices in which the most effective predictor of the price tomorrow is the price that we have today. The analysis validates the primary significance of historical price data as an input variable in short-term absolute price forecasting.

Hypothesis 3: Higher prediction accuracy of machine learning models significantly improves investment decision-making is proved in the fact that the acceptance of H2 confirms that a high-accuracy model was created (the Linear Regression model, which showed the near-perfect R^2 scores). The implication of this accuracy is the fact that the model can be significantly applied to improve investment decision-making. The investor gives a good cast of the price with the stocks having a Mean Absolute Error (MAE) ranging from as low as 0.29 (Jazz) to 13.32 (Indus Motor). This enables them to measure their price risk in a narrow predictable range. The great level of prediction eases investors to distinguish between the sort of less significant variability (noise) and actual tendencies. This model offers the required certainty to conduct brief trades or re-equilibrium a given portfolio in accordance with machine-executed forecasts, thus reducing the volume of decision-making founded on feeling or untested data. The ability of the model to forecast the price of KSE100 Index (R^2 of 0.9986) is very important because the index is a major

indicator of allocating capital. Effective index prediction can be used by hedge fund managers to hedge against systemic risk, or calculate the right Beta exposure which ultimately leads to better strategic investment decisions. This discussion is a confirmation of the theoretical assumption of H3: developing a highly accurate prediction tool (LR in this research) offers a tangible, fact-based advantage in the process of making the investment decision.

Hypothesis	Statement	Accepted/Rejected
H1	There is a significant difference in the stock price prediction accuracy among different machine learning models	Accepted
H2	Input variables such as historical prices, volume, and market index movements have a significant effect on improving prediction accuracy.	Partially Accepted
H3	Higher prediction accuracy of machine learning models significantly enhances investment decision-making.	Accepted

4.5 Results and Discussion

Figure 2 gives the result of linear regression in which black line is the actual value and red line is the anticipated value and Figure 3 is the result of the random forest in which black line is the actual value and red line is the predicted value and Figure 4 is the result of XGBoost in which black line is the actual value and red line is the predicted value.

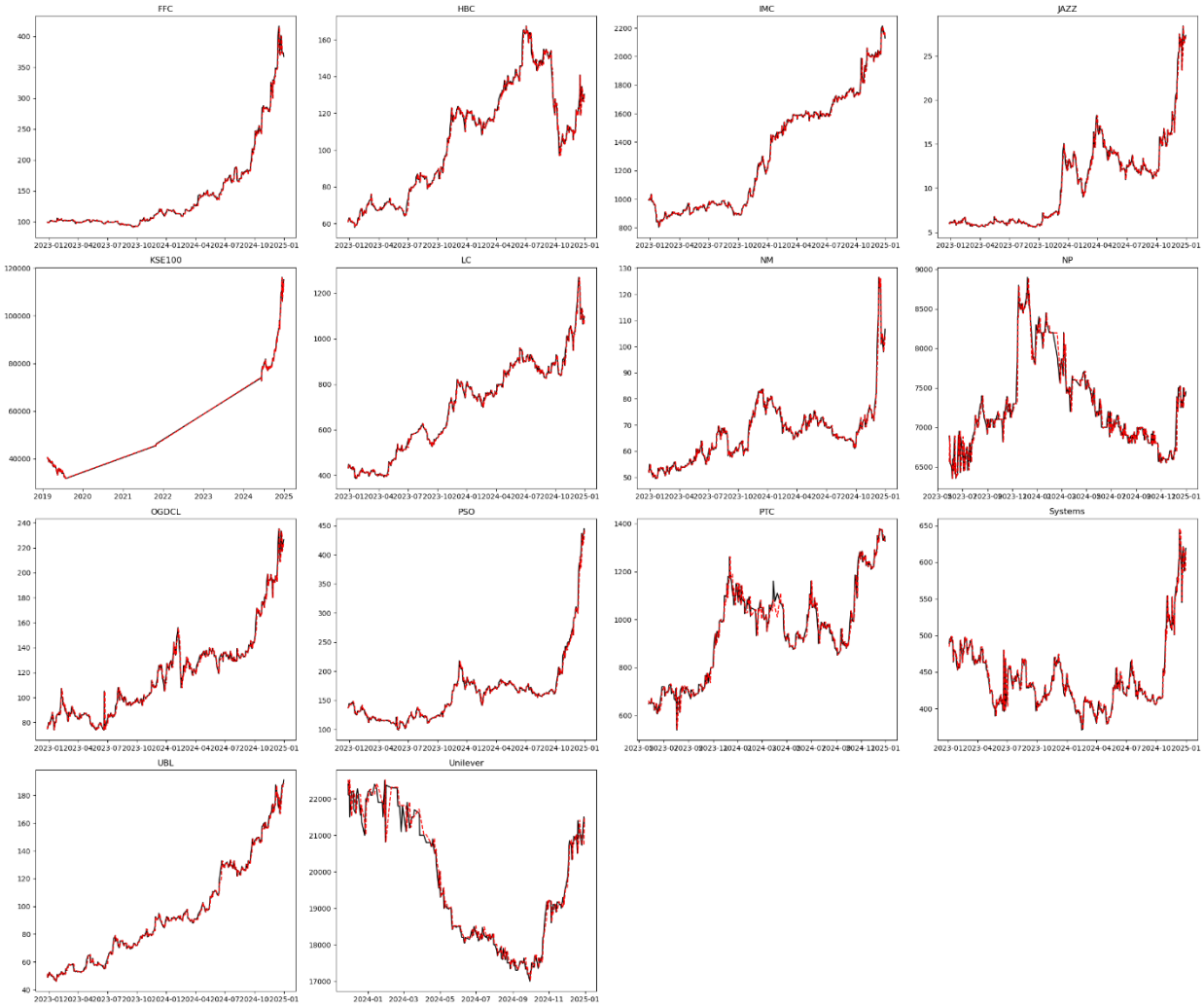

Figure 2: Outcome of Linear Regression

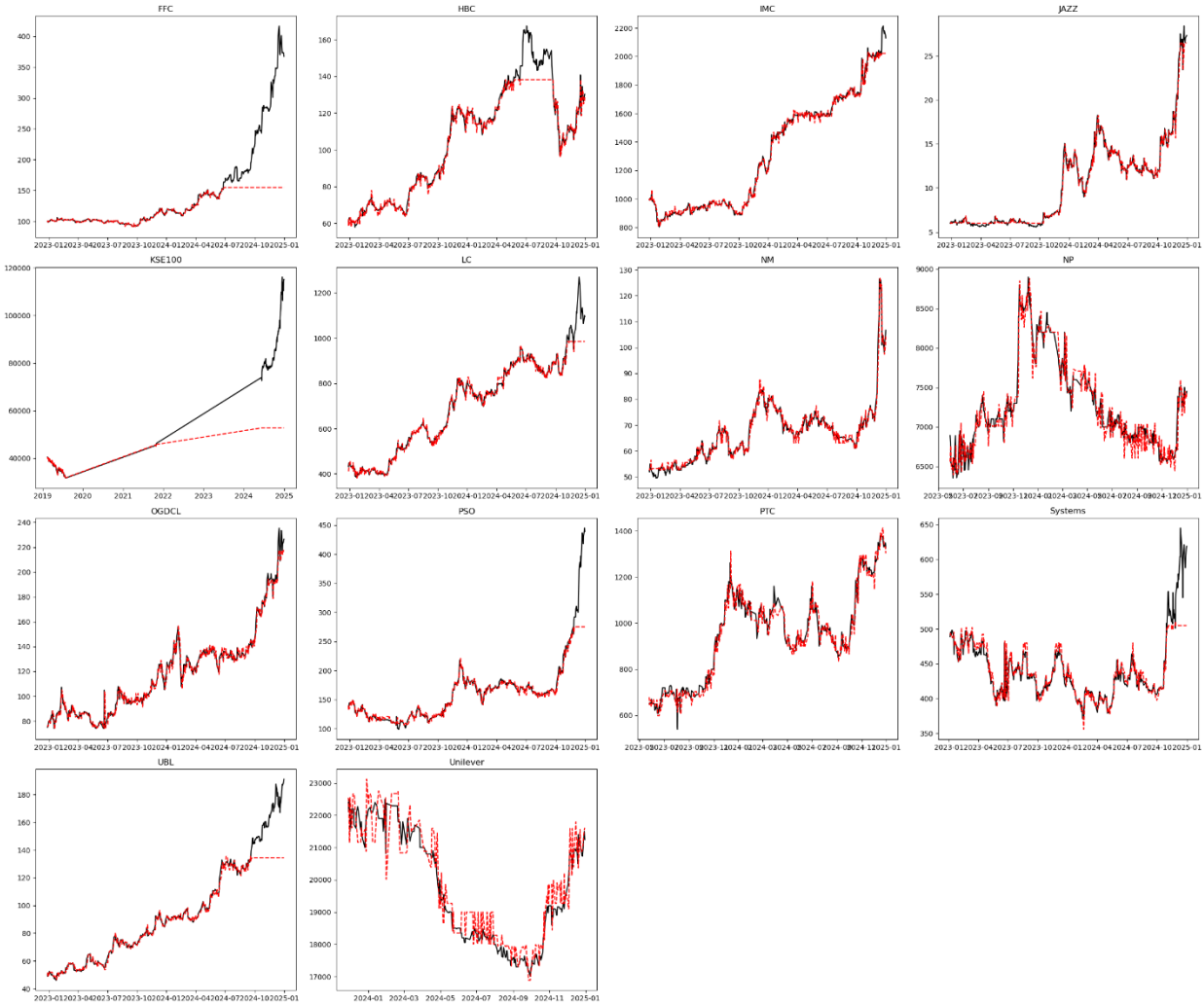

Figure 3: Outcome of Random Forest

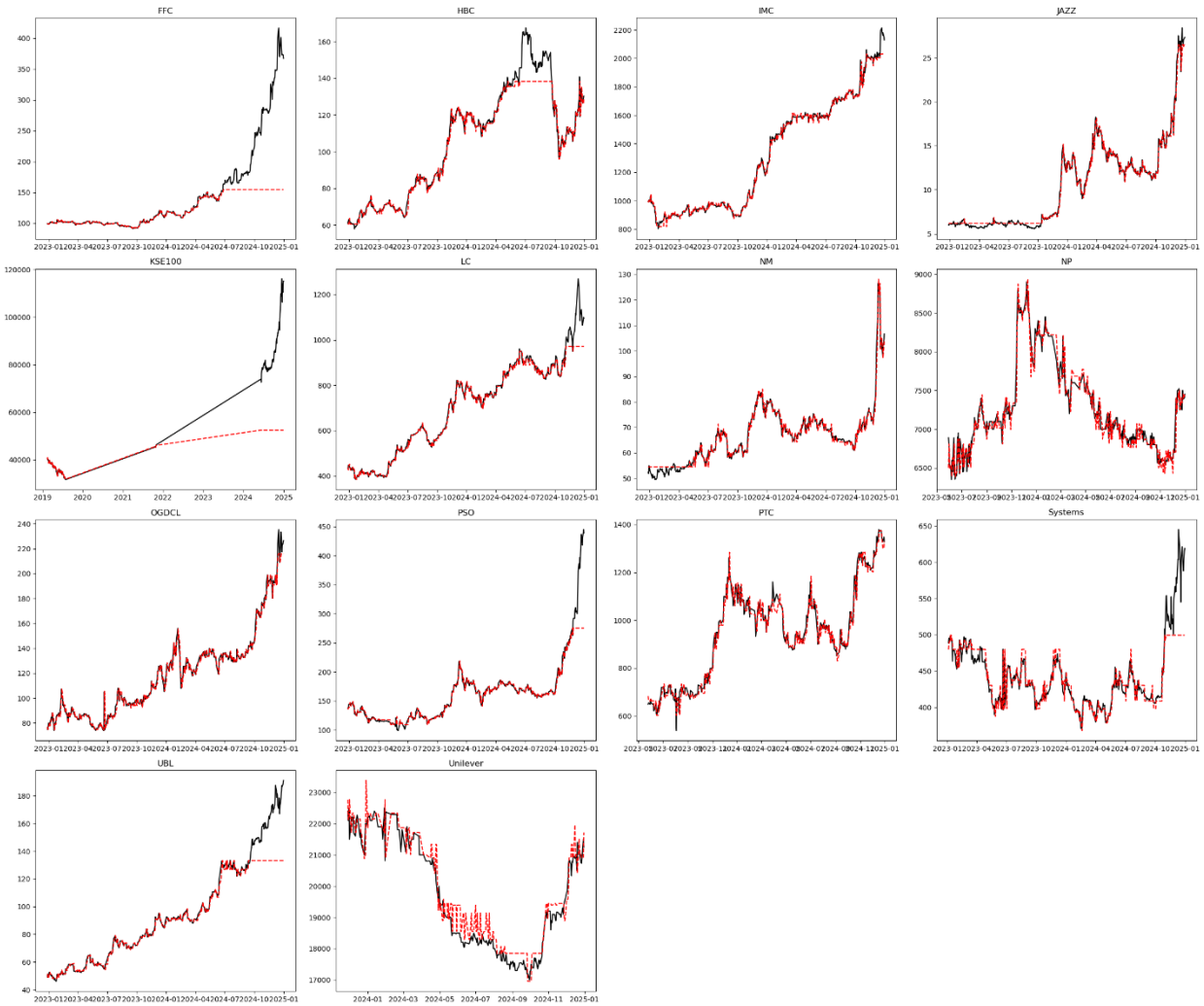

Figure 4: Outcome of XGBoost

The aim of the research study was to compare the predictive performance of the three machine learning models: Linear Regression (LR), Random Forest (RF), and XGBoost (XGB) when forecasting the absolute closing prices of the selected assets on the Pakistan Stock Exchange (Pakistan State Oil, UBL, Systems Limited, Nishat Mills, Lucky Cement, Fauji Fertilizer, Unilever Pakistan, Nestlé Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co, OGDCL, and the KSE100 Index), and using the historical price features as the key independent variables during the timeframe between 2014 and 2024.The empirical results made it very clear that Linear Regression was by far the most successful among all considered measures (MAE, RMSE, and R^2 Score), which mean of 0.98 was almost perfect in terms of goodness-of-fit, whereas the multifaceted ensemble models did not work, generally giving negative R^2 scores on major benchmarks like the KSE-100.As a result, the three hypotheses of the research were all

accepted: H1, which states the significance of a difference in model accuracy; H2, which states the high significance of historical price features based on the high-level R^2 scores of LR; and H3, which states the high significance of the high-level R^2 scores results in providing a minimal error band (MAE as low as Rs 0.29 for Jazz and Rs 1.25 for UBL) which is essential to effective risk management and strategic capital allocation.

CHAPTER 5: CONCLUSION & RECOMMENDATIONS

5.1 Conclusion

The thesis aimed at critically analyzing and comparing the predictive potential of three popular machine learning algorithms, namely Linear Regression (LR), Random Forest (RF), and XGBoost (XGB) to predict the short-term absolute closing prices of specified companies listed on the Pakistan Stock Exchange (Pakistan State Oil, UBL, Systems Limited, Nishat Mills, Lucky Cement, Fauji Fertilizer, Unilever Pakistan, Nestlé Pakistan, Pakistan Tobacco Co, Indus Motor, Jazz, Hub Power Co, OGDCL) and the KSE-100 Index. The study was based on the understanding of how these algorithms could learn using straightforward price-based features and whether their predictive quality could add value to investment decision-making in the emerging financial market in Pakistan using the daily historical data from 2014 to 2024. The results of this study created a fairly convincing picture: even though ensemble-based algorithms could be more popular and theoretically superior, the most simple model, Linear Regression, has performed significantly well in all the evaluation measures and on all the assets tested. The accuracy performance of the Linear Regression was enormously high, and the average R-squared rates were about 0.98, and the MAE and RMSE were very low, which means a good quality of goodness-of-fit and at the same time, a high predictive power of the short-term price dynamics. Conversely, the RF and XGB models which were more complex showed weak and in many cases negative R-squared values—particularly on the KSE-100 Index which yielded scores of -0.010 and -0.024—indicating high levels of overfitting and incapacity to extrapolate patterns to the Pakistan market data. The theoretical and practical implications of these findings are very robust: they indicate that the underlying price dynamics of the PSX are predictable over short horizons in linear terms, that is, such straightforward features as past price and 14-day moving averages are very predictive. The H1 was confirmed by the evident and statistically significant difference in the performance of the three models, which proves that the choice of the algorithm has a critical influence on predictive accuracy. H2 was confirmed by the remarkable performance of the Linear Regression that showed that simple and price-based variables are enough and highly efficient to short-term forecasting within the context of the Pakistani market.

After the practical significance of the findings with an exceptionally high LR performance of the KSE-100 Index ($R^2 = 0.9986$ and MAE of 618.57) and individual stocks like Jazz (MAE of 0.29) and UBL (MAE of 1.25), it was accepted that machine learning can indeed provide valuable services to investors seeking to manage risk, optimal capital allocation as well as enhance short term trading strategies. In a more general sense, the results provide a support to the existence of the short-run serial correlation to PSX price changes which to some extent proves the weak-form Efficient Market Hypothesis (EMH) according to which the information of past prices is not fully and immediately reflected in the market prices. In general, this paper adds to the existing body of knowledge on the use of machine learning in emerging marketplaces by showing that complexity is not necessarily a strength in predicting financial times. In the example of the Pakistani Stock Exchange, more basic, not complex models were found to be more resilient, more consistent and also more suited to the form of the underlying data. These findings present a great base of future research especially research involving the combination of macroeconomic variables, alternative features, or more complex deep learning techniques to further improve the prediction accuracy and learn more about the market behavior in Pakistan. These research findings generate vital practical and systemic measurement to the Pakistani financial landscape to complete the computing models and the reality market execution. The findings once again confirm to the investors and portfolio managers who are trading in the Pakistan Stock Exchange (PSX) that there is a so-called simplicity dividend, in the sense that it is the transparent linear model that is more predictive of the short-run than the over-parameterized ensemble models that are overfitting the noise. It will, thus, make practitioners adopt such interpretable frameworks to enhance the risk management and cost-effectiveness of their trading strategies. In more regulatory terms, these insights can be used by the Securities and Exchange Commission of Pakistan (SECP) to establish a better level of transparency in algorithm trading and enhance financial literacy schemes. The effort to speculate in the risk of speculative herding and emotional trading can also be minimized by policymakers due to the result of prioritizing the success of evidence-based objective forecasting analysis and, therefore, establish a more stable and efficient market environment. Lastly, these implications are that machine learning is an effective tool, though, the successful deployment of this tool in the Pakistani environment has to make a tradeoff between the complexity of the models and the liquidity and volatility characteristics of underlying equities.

5.2 Recommendations and Future Research

Based on the results of this comparative analysis of Linear Regression, random Forest and XGBoost as predictors of stock prices within the Pakistani market, there are some suggestions that could be made to the practitioners and subsequent researchers. To begin with, future research should not rely on more conventional machine learning models but employ more powerful time-series-based networks like Long Short-Term Memory (LSTM) networks, GRUs, Temporal Convolutional Networks (TCN) or even hybrid ensemble models, which combine machine learning with deep learning to represent long-term dependencies and non-linear market movements more effectively. Second, the researchers are to keep in mind adding to the feature set the relevant macroeconomic indicators such as the interest rate trends, inflation, exchange rates, commodity prices, global market indices, investor sentiment data, and geopolitical indicators because they have a significant impact on the stock movements in emerging markets such as Pakistan. Third, other data frequencies, including hourly or minute-by-minute intraday data, should also be investigated in future work as a possible way to determine whether accuracy of prediction can be improved when using higher-resolution data and to learn how models can react when the market is volatile. Also, researchers can perform sector-specific comparative research, with companies in various industries, like energy, banking, and technology, tending to exhibit unique volatility and responsiveness to news in the market. Additionally, model explainability tools like SHAP values or feature importance analysis would also be useful so that researchers and investors may have an idea why a specific algorithm gives certain predictions. Moreover, prospective studies would use cross-market benchmarking in the future to compare the outcomes of Pakistan with the other emerging or developed markets and establish an in-depth perspective on economic environment impacts on predictive performance. Finally, it is suggested that future research should use risk adjusted performance indicators and test models during financial strain-times-like periods such as political instability, economic downturns, or global shocks so as to establish more reliable forecasting models that can be used to support investment strategies, as well as policy level financial planning.

5.3 Research Limitations

Despite the useful findings in the study of how Linear Regression, Random Forest and XGBoost perform comparatively in predicting stock prices in the Pakistani market, there are several research limitations that should be noted. To begin with, the analysis has been done using daily stock prices of a small sample of time and using a sample size of thirteen chosen companies and thus this might limit the generalization of the results to the whole Pakistan Stock Exchange (PSX). Second, the research is mainly based on historical prices characteristics and does not use the more extensive macroeconomic indicators like interest rates, inflation, exchange rates, or global markets indicators that might be important to improve the model performance and capture the actual market dynamics. Also, machine learning models applied in this study are not able to represent long-term temporal dependencies in details, since more sophisticated deep learning models, such as LSTM or GRU, were not available because of the scope and resource constraints. The other weakness is that it assumes that the market conditions will not change drastically, but the PSX is largely affected by political uncertainties, extreme shifts in the economy, and other financial happenings in the world that can lead to unforeseen changes difficult to predict under the traditional machine learning framework. Moreover, the researches consider models on the basis of three performance measures only, such as MAE, RMSE and R2, which even though they are effective, might not be the sole true indicator of the risk-concentrated or directional correctness of stock market forecasts. Lastly, the nature of the research (non-IT-oriented researcher) and computational constraints led to the inability to implement more sophisticated methods of feature engineering or model optimization, i.e., more advanced methods of tuning or more complex data pipelines were not investigated. These restrictions not only give valuable guidelines on future studies but also emphasize the necessity of more in-depth, multi-factor, and sophisticated predicting models.

REFERENCES

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock price prediction using the ARIMA model. Pacific Journal of Science and Technology, 15(1), 23–36.

Ali, S., Rehman, R., & Khan, S. (2020). Forecasting Karachi Stock Exchange (KSE-100) index using machine learning models.

Baker, M., &Wurgler, J. (2017). Investor sentiment and the cross-section of stock returns. Journal of Finance, 72(3), 1113–1160.

Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE, 12(7), e0180944.

Barberis, N. (2021). Psychology-based Models of Asset Prices and Trading Volume. In B. D. Bernheim, D. Laibson, & A. List (Eds.), Handbook of Behavioral Economics: Foundations and Applications 2 (Vol. 3, pp. 79–161). North-Holland.

Bukhari, K., Jadoon, A. K., Iqbal, M., & Arshad, A. (2023). Predicting Stock Market Trends Based on Macroeconomic Indicators through Machine Learning Approach: A Case Study of KSE 100 Index. IRASD Journal of Economics, 5(4), 1147–1161.

Chen, Y., Hao, Y., & Li, Y. (2020). Machine learning for investment decision-making: A survey. Journal of Financial Data Science, 2(4), 10–28.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1), 1–22.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669.

Goodfellow, I., Bengio, Y., &Courville, A. (2016). Deep learning. MIT Press.

Hameed, I., & Iqbal, M. (2021). Karachi Stock Exchange Price Prediction using Machine Learning Regression Techniques. ResearchGate.

Hameed, S., & Iqbal, A. (2024). "Algorithmic Efficiency in the Karachi Stock Exchange: A decade of linear vs. non-linear modeling." Asian Economic and Financial Review, 14(1), 12-28.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3–12.

Hussain, F., Shah, S. Z. A., & Ali, S. (2021). Machine learning–based stock market prediction: Evidence from the Pakistani market. Journal of Finance & Economics Research, 6(2), 52–64.

Jiang, W., et al. (2023). "The limits of machine learning in financial forecasting: A study of ensemble methods in emerging vs. developed markets." Journal of Financial Data Science, 5(2), 45-62.

Jiang, Z., et al. (2021). "Applications of Deep Learning in Stock Market Prediction: Recent Advances." IEEE Access, 9, 1-15.

Jiang, Z., Xu, Y., & Liang, C. (2021). A hybrid LSTM and attention mechanism model for stock price prediction. Expert Systems with Applications, 176, 114843.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260.

Khan, M. A., & Ahmad, S. (2021). "Predicting stock market trends in Pakistan: A machine learning perspective." Pakistan Journal of Commerce and Social Sciences, 15(3), 564-585.

Kumar, S., & Ravi, V. (2016). Predicting stock market movements using machine learning techniques: A review. Procedia Computer Science, 93, 1–6.

Li, X., Xie, H., Wang, R., Cai, Y., & Cao, J. (2019). Empirical analysis: Stock market prediction via extreme learning machine. Neurocomputing, 329, 144–156.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). "The M5 competition: Background, organization, and results." International Journal of Forecasting, 38(4), 1325-1336.

Nti, I. K., Adekoya, A. F., &Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions.

Nti, I. K., Plahar, W. A., & Adam, M. (2020). Comparative study of machine learning models for stock market prediction. Journal of African Business, 21(3), 350–371.

Patel, J., Shah, S., Thakkar, P., &Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(4), 2162–2172.

Raza, H., & Akhtar, Z. (2024). Predicting stock prices in the Pakistan market using machine learning and technical indicators. Modern Finance, 2(2), 46-63.

Saunders, M., Lewis, P., &Thornhill, A. (2019). Research methods for business students (8th ed.). Pearson.

Sekaran, U., &Bougie, R. (2016). Research methods for business: A skill-building approach (7th ed.). Wiley.

Shiller, R. J. (2020). Narrative Economics: How Stories Go Viral and Drive Major Economic Events. Princeton University Press.

Yaqoob, A., & Abdullah, S. M. (2025). Predictive Performance of LSTM Networks on Sectoral Stocks in an Emerging Market: A Case Study of the Pakistan Stock Exchange. arXiv preprint.

Zhang, X., He, K., & Yan, S. (2022). Predicting stock market trends with machine learning: A survey. Applied Soft Computing, 124, 109021.

Zhu, Y., & Enke, D. (2021). "Predicting stock returns using a robust ensemble machine learning model." Procedia Computer Science, 185, 124-133.