Exploring the Predictive Power of Trading Volume on Stock Price Movements

SADIA MALIHA TRISHA

Applied Research Project submitted in partial fulfilment of the requirements for the degree of

Master of Science (MSc) in Business Analytics

at Dublin Business School

Supervisor: Andrew Kenny

26 August, 2025

Declaration

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of Master of Science (MSc) in Business Analytics is the result of my own analysis, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Sadia Maliha Trisha

Student Number: 20031017

Date: 26 August, 2025

Acknowledgements

I would like to express my sincere gratitude to Professor Andrew Kenny for his mentoring and constant supervision while I finished my master's thesis work. I would also want to thank Lecturer Suman Gurbani for her assistance in determining a research topic in the first place. I would also want to thank my friends and family for their encouragement and valuable insights on the research.

Abstract

This study examines whether trading volume significantly contributes to the predictive accuracy of stock price forecasting models. Trading volume is often viewed as a proxy for liquidity and sentiment, yet its independent predictive value remains unclear. Using daily S&P 500 index data from 2010 to 2025, three experimental datasets were constructed: a Volume-Only dataset, a Price-Only dataset, and a combined dataset with Principal Component Analysis (PCA) applied. Four machine learning algorithms—Random Forest, Support Vector Regression (SVR), XGBoost, and LightGBM—were tested under a leakage-free chronological split.

Results show that trading volume alone had no predictive power, with models yielding negative R^2 values. Price-only models performed strongly, with SVR achieving $R^2 = 0.8770$, confirming the persistence of stock prices as the dominant predictor. Adding volume did not improve accuracy, and in some cases reduced performance. PCA confirmed that volume represented an independent but weak component.

The results indicate that trading volume adds very little supplementary information, but historical prices remain the primary factors of forecasting accuracy.

Keywords: Trading Volume; Stock Price Prediction; Machine Learning; Random Forest; Support Vector Regression (SVR); XGBoost; LightGBM; Financial Analytics; S&P 500.

Table of Contents

Contents

Chapter 1: Introduction

1.1 Background of the Study

Financial markets are the most dynamic and complex systems economists and data scientists can study today. The price of stocks varies based on both ways proportional to several macroeconomic signallers, the fundamentals of the company, market moods, and the occurrence of the events, including political crises or environmental tragedies. The uncertainty of these flows poses both threats and opportunities to investors and portfolio managers as well as in the regulation sector. The accurate prediction of stock price behaviour has long been regarded as the holy grail of financial economics because a little bit of change in accuracy would result in huge financial gains and better risk management opportunities.

Traditional financial theories exist that claim that stock prices are a representation of all available information and thus cannot be predicted (Fama, 1970). This is however contradicted by behavioural finance and empirical studies, whereby inefficiencies exist with certain behavioural pattern that can be at sometimes exploited (Lo & MacKinlay, 1990). Trading volume has received a lot of attention amongst the various factors. Trading volume can be considered as an indicator of market activity and liquidity as it expresses the fervour of purchases and sales. There has been a long standing debate over whether the volume contains independent predictive information over and above the information contained in prices alone (Karpoff, 1987; Lee and Swaminathan, 2000).

The questions can now be tested in a systematic way with contributions in machine learning (ML) and data-driven analytics. Machine learning algorithms, including Random Forest, Support Vector Regression and just recently deep learning algorithms, offer the ability to very quickly reveal non-linear and latent relationships within financial data (Wang, 2020). Although a variety of studies employ these models to forecast stock prices, little systematic research exists to compare conditions in which trading volume has been isolated, excluded, and incorporated along with price-related variables.

1.2 Research Problem

Regardless of the numerous studies on this topic, the predictive characteristics of the trading volume have not been determined. There are analyses that assert that trading volume indicates informed trading and investor sentiment and it leads before price movements (Chordia & Swaminathan, 2000). There are also others who believe that volume is greatly superfluous as it adds very little to explanation after taking into consideration price-based characteristics (Llorente et al., 2002).

On a practical level, this uncertainty has implications of the kind of investment used and risk management. When volume is a good predictor, analysts and investors are supposed to integrate it in their forecasting models so as to enhance accuracy. As an example, should volume have low or no predictive power, it may not be worthwhile allocating resources to obtaining and processing such data.

This study is therefore occupying a knowledge gap by specifically testing the influence of trading volume through three experimental conditions that are controlled.

- Volume Only - reporting on whether trading volume alone can have any predictability to future price movements
- Without Volume - estimating predictive power when volume is left out of the models.
- With Volume -Wrapping up trading volume with price-based characteristics to find out how it performs on predictive accuracy.

The benefit of this form of categorisation does not only reveal the role of trading volume but also offers practical advice to decision-makers in the financial setting.

1.3 Research Aim

In this dissertation, the research question is whether the trading volume is worth considering as a predictor of changes in price of the stock. The research seeks to find out this impact of trading volume by employing three specially constructed datasets to which the models will be applied.

1.4 Research Questions

Main research questions:

Does trading volume significantly contribute to the predictive accuracy of stock price forecasting models?

The following Sub-Questions guide the research:

1. To what extent can the trading volume alone serve as a reliable predictor of future stock prices?
2. How does the exclusion of trading volume affect the accuracy and reliability of stock price predictions?
3. Does the incorporation of trading volume alongside price-based features improve the performance of machine learning models?

1.5 Research Objectives

The following are the objectives of the study which in an attempt to answer the following research questions:

1. To sweep the literature with critical review on the topic of stock price prediction and the controversial nature of the trading volume in financial correspondence.
2. To construct three sets of data (Volume Only, Without Volume, With Volume) and apply the role of trading volume.
3. To compare a number of machine learning models on these datasets.
4. To perform a comparison of the performance using performance measures (R^2, RMSE, MAPE).
5. To contrast and derive meaning out of results of the models in the three scenarios, finding the relevance and constraints of trading volume as a feature of prediction.
6. To derive theoretical and practical conclusions about the findings, to make contributions to academic research on financial forecasting, and to provide data to traders, analyst, and portfolio managers.

1.6 Significance of the Study

The dissertation has some contribution to theory and practical finance.

- Academic importance: Explicit comparisons of volumes only, without volume, and with volume cases give the study new insights into one of the long-standing debates concerning the importance of trading volume in modelling a stock price. Other research tended to put in the trading volume without isolating its effects that left ambiguity about the actual contribution to it. The study will fill that gap.
- Practical significance: To investors and financial analysts, the results will be illustrative of whether or not the presence of the volume of trade may be valuable by increasing accuracy of the prediction or it is just a distractor. This impact on the construction of forecasting models and algorithmic trading systems as well as on priorities of the data to be acquired.

This study also appreciates the fact that stock price prediction is extremely challenging. Because the nature of the financial markets is random and noisy, it is hard to achieve very high explanatory power (Fama, 1970; Frost, 2023). Thus, even quite low numbers of R^2 (0.20-0.40) can be viewed as significant contributions in the financial forecasting literature (Mehtab & Sen, 2020). It is within this context that the findings of this research offer valid information as to the predictive nature of trading volume.

1.7 Chapter Overview

The dissertation is organized in the following way:

- Chapter 1: Introduction- States the research background, business problem, aim, questions, objectives and significance of the study.

- Chapter 2: Literature Review - This chapter broadly reviews the financial forecasting literature, machine learning as applied to finance and the controversial issue of trading volume.
- Chapter 3: Methodology - Describes study design method, data gathering mode, features engineering and modelling strategy in all three test conditions.
- Chapter 4: Results and Evaluation - Shows and compares performance of models in the three scenarios.
- Chapter 5: Discussions - explains findings, makes connections with available existing literatures, and determines the practical and theoretical value.
- Chapter 6: Conclusion - Concludes contributions, presents limitations and proposes future research.

Chapter 2: Literature Review

The stock market has always been regarded as one of the most intricate and complex system in the global economy and it brings huge opportunities coupled with significant risks. The unpredictable nature of stock price fluctuation have attracted decades of researchers, investors and policy-makers to attempt to forecast and explain swings in stock prices using a multitude of computational, theoretical and empirical models. Although many studies have been performed over decades, the prediction of price dynamics is still one of the major problems in finance. The challenge exists due to the number of factors that affect the prices of stocks, such as macroeconomic situation, monetary policies, investor optimism, geopolitical, company specifics and technological matters. The emergence of large volumes of data and the developments in computational methods in recent decades have opened new opportunities in forecasting. Machine learning models have received keen interest due to the capacity to process large, complicated, and nonlinear data. Trading volume is one of the most debated and yet less explored variables to be used to predict price movements. Although common forecasting models incorporate trading volume as part of their characteristics, its forecasting ability - when used as a point of focus - has not been tested much specially in the analysis of a major index as the S&P 500. This paper fills that gap by testing trading volume as the primary predictor of stock prices and assesses the explanatory power of that predictor with four machine learning models: Random Forest, Support Vector Regression, XGBoost, and LightGBM.

By starting with some of the earlier work in financial theory, researchers of the market microstructure and information flow have highlighted the close relationship between trading activity and price change. Clark (1973) Mixture of distribution Hypothesis, it argued that market's trading volume and over-time prices volatility were influenced by arrival of new information to the market. In such sense, when new information is given to investors they respond by engaging in buying or selling assets, in response to that the trading volume and price of stock changes. Sequential Information Arrival Hypothesis (1976) was advanced by Copeland who postulated that the information cannot flood all investors simultaneously but will take time to spread. Such sequential patterns of trading and price adjustments are as a result of this staggered process of information diffusion. Additional empirical evidence supporting these findings was given by Tauchen and Pitts (1983) who gave evidence that trading volume is strongly related to prices variations, which strengthens the considerations on the idea that volume actually bears important information about the market machinations. These theories are quite old, however their relevancy cannot be undermined in the present-day complex markets as their applications could be thoroughly tested with modern computational tools.

Modern theories have been based on these foundations by adding the concepts of behavioural finance and market microstructure. In behavioural finance, volume is commonly used as a proxy indicator to rate the sentiment of investors. Investors are more likely to trade once they feel confident or optimistic; hence the higher the volume of such trade, the more it can be used to enhance the price trend. In contrast, the trading volume decreases during periods of uncertainty and this can be considered as indication of hesitation and indecisiveness. Chordia, Roll, and Subrahmanyam (2001) were able to establish the relationship between the trading volume and the liquidity by stating that high trading levels are keys to the sustainability of efficient markets. Pastor and Stambaugh (2003) went further to show that liquidity risk in itself, is reflected within the expected stock returns further reaffirming the notion that volume not only contains information about market risks, but opportunities too.

Recent empirical studies are convincing evidence which shows that trading volume holds predictive power. Zhang, Li, and Wang (2023) perform a comprehensive study of the S&P 500 and prove that the price change is strongly related to the volume. Their data indicate that not only is the trading volume a result of price changes but it can also serve as a leading indicator. Guan and Chen (2023) found that excessive volume growth frequently was a precedent to a price reversal using high-frequency data: in the case of green reversals, excessive volume growth is also an antecedent. Jay, Mehtab, and Dutta (2021) included volume in LSTM-based forecasting models and stated that it improved the accuracy of predicting sudden market shifts, which indicated volume gathered dynamics that price data on its own was unable to.

Similar findings have been observed in other markets across the globe. Kumar and Sharma (2022) used Random Forest and gradient boosting model on the NIFTY 50 index in India and made the conclusion that trading volume contributed to the accuracy of the predictions. In their research of the Vietnamese stock market, Phuoc, Kim Anh, and Tam (2024) concluded that both SVR and Random Forest gave credible predictions when volume was considered as a feature. In European markets as well, including the FTSE 100 index, there are findings that a high volume is one of the key indicators of a higher volatility (Yu et al., 2023). Taken together, these results suggest the universality of the importance of trading volume in various market conditions.

The COVID-19 situation presented another actual experiment that further evidenced the importance of trading volume. The COVID-19 pandemic caused several unexpected surges in trading volume in most markets as investors reacted to emerging events. Researchers like Karim et al. (2022) discovered trading volume as an important variable in the prediction of a tighter volatility of stock prices during this time. Rahman et al. (2022) have shown that sudden shocks could be modelled using volume-based models rather than models based on price information only. The studies emphasise the relevance of volume as a predictor even in the current era of crisis-prone markets.

Besides experimental evidence, behavioural approaches give an effective reasoning on why volume of trade is predictive of price changes. Investors also tend to interpret heavy trading volume as a sign of confidence in the market which can reinforce the current trends in the prices. Jegadeesh and Titman (2019) demonstrated that the similarities and differences in stock returns are inflated by the trade volume. High volumes themselves are not merely indicators of action but influences of behaviour by facilitating herding and by magnifying market psychology. Trading volume on the other hand, except in the case of quality gapping patterns, will often be low, indicating indecisiveness and can be following by periods of stagnation or weak trends. These behavioural dimensions are essential in trading volume because of the ability to capture the price dynamics.

Financial forecasting is now becoming revolutionized by machine learning because it has liberated researchers who are no longer bound within the confines of conventional econometric models. Classical methods like ARIMA or GARCH, though being good, presuppose the existence of linear relationships and do not help to comprehend nonlinearity and dynamic nature of financial data. Machine learning models on the other hand can handle high-dimensional intakes, can establish the non-linear interactions, and can adjust to the changing patterns and thus they are very marketable in terms of forecasting in the stock market. Random Forest, SVR, XGBoost and LightGBM are among the broad range of machine learning methods that have an advantage by combining predictive ability, interpretability, and performance.

Random Forest is already popular in the finance field due to its sound results and non-linearity capturing ability. By averaging the predictions of multiple decision trees, Random Forest can minimize the hazard of overfitting, and enhancing generalisation. He and Zhao (2024) found that Random Forest worked well especially in forecasting returns during the volatile environments. Al-Obaidy and Yin (2024) used Random Forest on the absolute price changes of the S&P 500 and that it performed well even in other setups. The performance of Random Forest remains strong when used with the NIFTY 50 (Kumar and Sharma, 2022), once again showing the flexibility of the ML model. Notably, in addition to the high predictive power, Random Forest gives information on the significance of features and, therefore, the relative importance of trading volume in predicting stock prices can be quantified.

The other model that would be very suitable when it comes to financial forecasting is the Support Vector Regression (SVR). The fact that SVR may use kernel functions enables it to represent nonlinear functions and is therefore very successful in covering irregular processes as the one demonstrated by the trading volume. Kumar and Gupta (2016) stated that SVR outperformed linear models in situations that are both complex and non-linear in the context of finances. The reliability of SVR in varying time horizons and in different market conditions was proved to be valid by Phuoc, Kim Anh, and Tam (2024). Li, Liu, and Song (2019) had trained and tested using SVR to directly model volume, and reported that it had worked well to explain volatility. These works show that SVR can be very useful to analyse trading volume as a stock prices determinant.

One of the most widely-used machine learning models in finance, XGBoost is powerful and scalable. Developed by Chen and Guestrin (2016), XGBoost contains methods of regularization that are useful in reducing overfitting, and so it is effective in monetary datasets. Zou et al. (2022) concluded that XGBoost was the most effective model in making stock prediction especially with complicated interaction between the predictors. It has emerged as successful in both competitions and academic research lending a practical significance to its application in financial prediction with regard to volume.

LightGBM, presented by Ke et al. (2017), is based on gradient boosting, but is more efficient and fast. LightGBM is faster and more memory efficient than other algorithms, without sacrificing accuracy, using a histogram-based learning algorithm and leaf-wise growth method. LightGBM performed well in the stock prediction with low computational demands and thus can be used in long-term data like the one practiced in this study. LightGBM is relevant to analyse the S&P 500 over fifteen years, and the efficiency of its work with big data is its determining feature.

Although researchers have already made great contributions in stock price forecasting, there remain many gaps in the contribution of trading volume. Most of the studies have trading volume as one of the many factors considered which does not help in ascertaining the contribution of trading volume individually. Not many papers address volume as the source of primary interest. Other problem is that most researchers treat movements in the stock as a classification problem, i.e. as predicting that it will be up or down, rather than an estimation of the price. As good as classification may be, regression will provide a more comprehensive and realistic basis on which to make decisions such as portfolio management and the trading strategies. An additional limitation is that most of the literature focuses on small time series or special events, not addressing whether trading volume holds up on longer time scales. Lastly, the deep learning models have been heavily relied upon in a lot of recent work. These are typically powerful, but also complex, resource demanding and hard to interpret. What is lacking, therefore, is a study that centres on trading volume and subjecting it to machine learning techniques that are predictive and that are not-complicated.

The dissertation can fill these gaps because the study solely uses trading volume as a predictor as well as employing four machine learning models that are mutually exclusive in terms of accuracy, interpretability and efficiency. Through the analytical examination of a 15 plus year data of S&P 500, it offers one of the more in-depth long-term examinations of the predictive ability of volume. By posing the problem in the form of a regression exercise it provides estimates of real price levels which are useful in serving the interests of investors and portfolio managers. Using Random Forest, SVR, XGBoost, and LightGBM, it reveals a comparative outlook on the performance of models, with the insights of the way, in which the models treat the volume-based prediction.

Trading volume is a variable that has been long been identified to be important in finance, as a measure of liquidity, a flow of information and sentiment. Recent empirical evidence supports the idea that the volume does increase predictive power even across a large selection of models and a variety of markets. Behavioural theories justify why the level of trading should be used to determine price patterns as opposed to using other volume-based factors, and a machine learning algorithm is the vehicle that can test the predictability of this volume measure. The independent role of volume has been little studied, on longer term data, with regression-based methods. This dissertation fills a gap in the literature by treating trading volume as a centre of analysis, adopting a regression approach and applying four powerful machine learning models. Its results should add value to scholarly research as well as inform the practical use of financial forecasting that embraces trading volume as one of the primary indicators in the modern markets of stock prices.

Although studies on predicting the stock market have been undertaken, there are still gaps that have not been covered. The majority of past studies have incorporated trading volume with other factors and this questions the ability to detect the influence of trading volume in the centre of focus. A large portion of the literature also deals with more of a classification problem that is predicting the possible rise or fall in prices rather than a regression model that predicts the actual price level. Regression methods themselves can give more realistic details to the investor and portfolio managing. Moreover, a lot of the research has been short-term oriented or deep learning-based, which have been known to be resource inefficient and opaque.

This study attempts to fill these gaps by looking into the effect of trading volume in forecasting price of stocks under various conditions of experiments. Based on it, the study will operate on the following hypotheses:

“Trading volume contributes significantly to the predictive accuracy of stock price forecasting models.”

Chapter 3: Methodology

3.1 Research Philosophy and Approach

There is a philosophical position behind the method of any research project, usually because it makes up the way of constructing the knowledge and interpreting it. The given study can be placed within the positivist paradigm according to which the reality is objective and can be perceived through the system of observation and measurement. Positivism is especially suitable in financial research owing to the fact that financial markets produce very large quantities of measurable data, e.g. prices, traded quantity, and returns, and these data can then be formulated into models using statistical and computational methods. With this paradigm, the task of the researcher is to observe, measure, and analyse patterns on the data to make assertions on financial behaviour as general living.

On this philosophical inclination, the analysis implements the abductive approach in research reasoning. Abduction is not just purely deductive and inductive reasoning. Deductive reasoning takes ideas, theories, and knowledge as the foundation in order to explain information. By comparison, inductive reasoning is that which moves upstream, running from information to theory, and is that which discovers patterns which may result in the attainment of new concepts. Abduction is an amalgamation of the two: it builds on some intellectual expectations, but leaves itself open to facts that can modify or even complicate the previous expectations. Abduction finds a perfect application in the study of this research as or rather proposed study owing to the theoretical writing in which there is conflicting arguments in the nature of trading volume with respect to predicting the stock price. The Mixture Distribution Hypothesis (Clark, 1973) and Sequential Information Arrival Hypothesis (Copeland, 1976) postulate that it is the trading action that is strongly connected to the price change, whereas more recent empirical studies do not show such definite results. The study is conducted by applying machine learning to a long-time-series of S&P 500 data to test these theoretical arguments deductively, and to support emergent, data-driven patterns inductively.

The study is unquestionably quantitative, which goes well with positivism and abduction. Quantitative analyses would be the best fit in this project since the research question will focus on predictive accuracy, which can be measured using Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2. Conversely, a qualitative method of gathering information about the way investors use volume, i.e., by interviewing investors, would not allow obtaining the sort of quantitative data that could be used to form conclusions regarding the significance of the factor of volume in predicting prices. A quantitative abductive approach thus makes the investigation theoretically informed as well as empirically consistent.

3.2 Research Design

The research design outlines how philosophically and methodologically prescribed decisions are to be operationalised. This dissertation is a quantitative experimental research where the models are trained and evaluated by using various experimental conditions on a dataset to find out whether the trading volume is an essential aspect of predicting the price. The design is organized on the basis of three datasets:

1. Volume-Only Dataset-This dataset has only lagged log-volume functions. It is constructed to test whether trading activity in isolation has the ability to capture price trends and afford predictive powers beyond information based on prices.
2. Without Volume Dataset- This dataset was formed using only the price bases features (lagged log-close contents). It can be used as a benchmark reference to understand the performance of the models in lack of trading volume, thus the incremental usefulness of volume in prediction can be evaluated.
3. With-Volume Dataset-This dataset includes both lagged log-volume and lagged log-price values. It gives a total picture of market movement and it is scaled such that it can be compared directly to the other two scenarios to ascertain whether inclusion of trading activity increases predictive power.

Regression design is an intended choice. Whereas in earlier literature prediction is presented in terms of a classification problem, i.e., predicting whether market prices will increase or decrease, a classification does not offer any insight into the value to which the market moves. Regression, however, predicts continuous values, which are more beneficial to the portfolio management and risk measurement because it estimates the direction of the price.

Moreover, the choice of using machine learning instead of a conventional econometric model is also explained by the fact that in the former case, the shortcomings of the latter are resolved. Modelling approaches like ARIMA or VAR assume stationarity, linearity and normality that are commonly not satisfied by financial time series. Volatility on market data is also often clustered, and shows heavy-tail distributions, as well as structural breaks. The complexities will be best dealt within the framework of machine learning model since they are flexible, non-linear relationship can be modelled and they are not subjected to parametric constraints. This renders them especially strong on stock price forecasting, as the relationships between features and outputs are neither linear nor reliable.

3.3 Research Hypotheses

The central hypothesis of the study is:

- **H_0 (Null Hypothesis):** Trading volume does not contribute significantly to the predictive accuracy of stock price forecasting models.

- **H_1 (Alternative Hypothesis):** Trading volume contributes significantly to the predictive accuracy of stock price forecasting models.

Sub-Hypotheses

To operationalise this main hypothesis across the three dataset conditions, sub-hypotheses are defined:

- **H_0a:** Trading volume alone has no predictive power for stock prices.
- **H_1a:** Trading volume alone has predictive power for stock prices.
- **H_0b:** Excluding trading volume does not reduce predictive accuracy.
- **H_1b:** Excluding trading volume reduces predictive accuracy.
- **H_0c:** Adding trading volume to price-based features does not improve predictive accuracy.
- **H_1c:** Adding trading volume to price-based features improves predictive accuracy.

Justification of Hypotheses

Those hypotheses are based on the identified main gaps in the current literature. Most of the current studies use trading volume to facilitate other predictors thus it is not possible to determine its independent contribution. Other literature highlights classification, as opposed to regression and many use short-term horizons or use deep learning-based techniques which are not interpretable. This study aligned the hypotheses with the dataset conditions, which means that each of the experimental situations reflects a certain gap. The models also offer an organized method of understanding the findings: each unique outcome can be utilised to accept or reject the model and therefore all the findings make a direct contribution to theory and practice.

3.4 Data Collection

The study makes use of secondary data obtained from Yahoo Finance, covering daily observations of the S&P 500 index from January 2010 to June 2025. The raw dataset includes:

- Date
- Open price
- High price
- Low price
- Close price
- Adjusted Close price
- Trading Volume

The S&P 500 was selected as the dataset because of few reasons. To begin with, it is the broadest followed equity index globally and it has commonly been used as a measure of an indication of the performance in the world finance. Second, it is also very liquid and thus its trading volume is significant and the index is also reflective of trading by some of the largest U.S companies. Third, the time horizon of more than 15 years allows to track several market cycles, including bull markets, bear markets, as well as crises like as the COVID-19 pandemic. This makes the results to be robust whatever the economy is going through. Lastly, since the secondary data was obtained in an authoritative source, it increases transparency and replicability since other researchers can access the same.

3.5 Data Preparation and Pre-processing

Raw financial data cannot be directly used in machine learning models, but instead needs to be transformed to be in the form of features. The following pre-processing was made:

1. **Chronological ordering**: Data was arranged, according to time-series, by date. This was because no mistaken information was later inserted.
2. **Handling missing values:** Missing values in S&P 500 data were rare and were treaded as forward-fill methods, i.e. missing entries were replaced by the latest known value. This prevented artificially changing volatility.
3. **Log transformations**: Trading volume and closing prices were log-transformed to stabilise variance and reduce skewness, providing proportional changes with regard to time could be compared on a more consistent basis.
4. **Lagged features**: Lagged price and volume data (1 day-5 days) were created. This also makes the use of past information only by models, simulating real conditions of forecasting.
5. **Scaling and normalisation**: The features were scaled to similar proportions, and this was more important in the case of SVR which is sensitive to features magnitudes.
6. **Principal Component Analysis (PCA):** Principal component analysis (PCA) was used as an optional final step on the best model to eliminate redundancy of highly correlated lagged features. Additional evidence is presented on the usefulness of trading volume in stock price predictions since PCA was employed to analyse the components with the largest contribution to predictive quality.

The simplification of every transformation was theoretically justified. An example is that lagging, which uses only past values for forecasting, prevents data leakage. All transformations, required to stabilise variance and reduce skewness in the volume variable, were applied. The feature scaling technique was used to make the inputs comparable to each of the models. PCA was used on the top performing model as an exploratory post-step to examine any redundancy on the feature set and to determine the importance of trading volume as a feature.

3.6 Engineered Datasets

Following pre-processing, three experimental datasets were created::

1. **Volume-Only Dataset**

   - Included only lagged log-volume features (1–5 day lags).
   - Purpose: to test whether trading activity alone can capture price trends and provide predictive power independent of price information.

2. **Without-Volume Dataset**

   - Included only lagged log-close features (1–5 day lags).
   - Purpose: to establish a baseline for price-only prediction, providing a reference point to assess the incremental value of trading volume.

3. **With-Volume Dataset**

   - Combined both lagged log-volume and lagged log-close features.
   - Purpose: to evaluate whether including trading activity alongside prices improves predictive accuracy, allowing direct comparison with the other two datasets.

This tripartite dataset structure ensures that the independent and complementary roles of trading volume can be systematically evaluated.

3.7 Data Leakage Prevention

Data leakage in financial forecasting is also a frequent trap of the predictive model. Leakage causes excessive performance data and models that do not perform in the real-world. Avoiding leakage was thus a top priority in terms of methodology.

This research used various protection measures.

1. **Lagged features only**: All the predictors based on past values were created (they could not be based on contemporaneous and future).
2. **Chronological split**: Date was segregated in chronological manner where training ranged between January 2010 and February 2025 and testing was confined to March till June 2025. This is the same situation that exists in real life when there is an attempt to predict future unknown prices.
3. **Hold-out validation**: The test set was not used in any aspects of training or hyper-parameter optimisation, of course, providing out-of-sample testing.
4. **Avoiding data snooping**: Care was taken not to contaminate training and test data by applying only feature engineering decisions to each data set.

The voltage study provides an appropriate means of addressing leakage rigorously so that the predictive accuracy reported in the study reflects genuine model performance.

3.8 Model Selection

Machine learning models selected were four.

1. Random Forest Regressor - A collection of decision trees, Random Forest is resilient to noise, models nonlinearities, and can give feature importance scores hence it is particularly powerful when evaluating the contribution of volume.
2. Support Vector Regression (SVR) - SVR makes use of kernel functions to map the data into high-dimensional space to enable it to capture a complex data relationship. It is an appropriate method in volatile financial data, which are no longer in conformity with linear modelling.
3. XGBoost - A gradient boosting algorithm and known to perform with state-of-art performance in structured data. There are regularisation terms that curb over fitting, a notorious issue in financial prediction.
4. LightGBM - Boosting algorithm fast and scale. LightGBM has the capability of dealing with large data efficiently, especially when long-term series data are at hand.

The models combined provide a reasonable balance of exploratory capability, interpretability, and computing efficiency. Their heterogeneity makes the obtained results model-free, which makes them more robust.

3.9 Model Training and Validation

The models were trained on the past data of January 2010-February 2025 and tested on the March-June 2025. In this test set the focus is on out-of-sample evaluation, which makes the results more realistic.

- **Hyperparameter tuning**: The hyperparameters were optimised through the use of a grid search with regard to tree depth, the number of estimators, kernel type, and the learning rate.
- **Cross-validation**: was carried out on the training set in order to stabilise the parameters.
- **Regularisation**: Incorporated in boosting models as it minimises the danger of overfitting.
- **Comparative evaluation**: The same models were used in all three data sets, and this allowed relatively fair comparisons.

3.10 Computational Environment

All the analysis was done in Google Colab which is a free cloud based platform that provides CPU and GPUs. Colab was selected because it is relatively easy to access, replicate and use with Python-based libraries.

Libraries utilised included:

- **pandas, numpy** – for data manipulation.
- **yfinance** – for data extraction.
- **matplotlib, seaborn** – for visualisation.
- **scikit-learn** – for preprocessing, PCA, Random Forest, and SVR.
- **xgboost, lightgbm** – for gradient boosting models.

This computational setup guarantees transparency through the ability to reproduce such an analysis because other people can use the same libraries.

3.11 Evaluation Metrics

Three of the targeted metrics were selected.

- Root Mean Squared Error (RMSE): It is sensitive to the large deviations so it is vital in financial applications.
- Mean Absolute Percentage Error (MAPE): It gives an intuitive measure of error in percentages.
- R^2 Score: It measures the proportion of variance explained.

A combination of these two measures gives a balanced measure of accuracy, scale of error and explanatory power.

3.12 Ethical Considerations

The study is based on the publicly-available secondary data on Yahoo Finance. Any sensitive or personal information is not involved. Transparency is maintained through the use of open source libraries and, documentation methods extensively.

3.13 Limitations

Although it has strengths, the study has drawbacks. Its emphasis on the S&P 500 only may not be universal. It omits macro and sentiment indicators which could affect prices. Although powerful, it cannot be said that machine learning models are transparent as opposed to econometric approaches. Last, there can be structural changes that can decrease predictive stability like crisis.

3.14 Chapter Summary

This chapter has described the way the study is conducted. The positivist, abductive, quantitative approach was used. The data obtained onto Yahoo Finance (2010 - 2025) was pre-processed into three datasets, namely Volume Only, Without Volume, and With Volume. The threat of the leakage of data was reduced through the application of reserved and orderly prevention. These models were trained and tested via Google Colab using reproducible packages in four steps Random Forest, SVR, XGBoost, and LightGBM. Performance was measured by using RMSE, MAPE, and R^2.

This research approach can offer a very strong method to answer the research question: Does trading volume add any significant value to the predictive power of stock price forecasting models?

Chapter 4: Results and Evaluation

4.1 Introduction

The chapter contains the empirical findings of the study. In terms of restricting the purpose of determining the independent and complement nature of trading volume in stock price forecasting three experimental data sets were formed; they are Volume Only, Without Volume and With Volume. The same four machine learning models, Random Forest, Support Vector Regression (SVR), XGBoost, and LightGBM, were used with all of these datasets. In addition to that, Principal Component Analysis (PCA) was employed as a complementary analysis on the most well performing model to further understand feature redundancy and contribution of trading volume.

Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and R^2 are used in this study to analyse the performance. The absolute value of the prediction errors are used as the measure in MAPE, and big deviations are heavily criticised. The results of a transition probabilities test are presented as percentages; this makes the explanatory results more natural to understand. R^2 is the main interest of this study which is a measure of the proportion of variance in stock prices explained by the model and is thus directly related to the explanatory contribution of trading volume as a predictor. Compositely, these measures give an objective assessment of accuracy of forecasting along with the magnitude of explanation.

The structure of the presentation of results as follows. Fourth section, 4.2 is an overview of the dataset. Model results on each dataset are tabulated in detail in Secs. 4.3 to 4.5 respectively. Section 4.6 compares Principal Component Analysis (PCA). Section 4.7 shows the visual analysis and 4.8 analyse the hypothesis. The end of Section 4.9 summarises.

4.2 Descriptive Overview of the Dataset

To understand the nature of the dataset, it is worth having a look at the dataset prior to presenting model results. The data will take into consideration S&P 500 index, as well as the trading volume of this index during the time frame of 2010-2025, which may comprise several market cycles including bull markets, bear markets, and the COVID-19 crisis of 2020. Both stocks prices and the trading volumes are characterized by significant fluctuations during the time frame. The trading volume might increase due to strong spikes every time there is a crisis instant or even a financial shock. Such propensity establishes a critical background against which the hypothesis that trading history conveys descriptive information over and above the history of prices can be tested.

4.3 Results: Volume Only Dataset (H_0a/H_1a)

The first experimental dataset included only features derived from trading volume. The results of the four models are reported in Table 4.2.

Table 4.1: Model Performance – Volume-Only Dataset

Model	RMSE	MAPE	R^2
RandomForest	2746.30	47.31%	–102.9443
SVR	3125.77	54.36%	–133.6533
XGBoost	2690.45	46.11%	–98.7595
LightGBM	2725.20	46.89%	–101.3529

The findings show that trading volume alone has very weak predictability. Based on their negative R^2 values, it could be interpreted that all models underperformed than a naive mean predictor. RMSE values were incredibly high at the level of 2,700- 3,100 index points, and MAPE values rated above 45 percent confirming low accuracy of forecasting. XGBoost resulted in the best RMSE and MAPE, however, very slight, and still not satisfactory in practice.

These results indicate that volume trading cannot be used to observe price movements or trends alone. Therefore, the finding confirms the null hypothesis (H_0a) which is that the trading volume by itself is not predictive of stock prices. To practitioners, this means that the market activity shown by volume would be a poor indicator yet relying on the indicator fully would not deliver strong and sound forecasts.

4.4 Results: Without-Volume Dataset (H_0b/H_1b)

The second experimental dataset included only lagged log-close features, excluding trading volume. The results of the four models are reported in Table 4.3.

Table 4.3: Model Performance – Without-Volume Dataset

Model	RMSE	MAPE	R^2
RandomForest	109.20	1.37%	0.8357
SVR	94.48	1.10%	0.8770
XGBoost	101.01	1.31%	0.8594
LightGBM	98.46	1.33%	0.8664

As shown in the results, lagged price data gives some good predictive power even in the absence of trade volume. Results showed that all the models had large R^2 values more than 0.83 indicating that historical prices could explain a very high percentage of underlying stock prices variance. RMSE values were maintained between 94 and 109 index points, whereas MAPE values were within 1.10 to 1.37 percent, indicating that the predictive errors were not large relatively speaking.

The SVR performed the best of all four algorithms [$R^2 = 0.8770$, RMSE = 94.48, MAPE = 1.10%], followed by LightGBM [$R^2 = 0.8664$]. Random Forest was marginally less (but still robustly) accurate than the boosting methods.

This information backs the view that there exists much information on price behaviour in their history which serves as a good predictor, as had been recorded in the financial literature about autocorrelation of stock prices and price trend persistence. It was, however, the fact that this dataset used did not collect trading volume that also offers the important base point against which the additionality of volume can subsequently be measured.

As far as hypothesis testing is concerned, the findings are closer to the null test (H_0b): the volume exclusion did not have a significant effect on predictive accuracy. This implies that although price information is extremely potent, additional assessment is required to determine the usefulness of trading volume in providing greater explanatory power.

4.5 Results: With-Volume Dataset (H_0c/H_1c)

The third experimental dataset combined lagged log-close features with lagged log-volume features. The results of the four models are reported in Table 4.4.

Table 4.3: Model Performance – With-Volume Dataset

Model	RMSE	MAPE	R^2
RandomForest	121.02	1.58%	0.7982
SVR	94.75	1.09%	0.8763
XGBoost	102.14	1.33%	0.8562
LightGBM	101.42	1.33%	0.8582

The addition of trade volume to price characteristics did not have so much impact as to produce different results that can be traced to difference models of results. All algorithms achieved a high accuracy with R^2 values between 0.7982 and 0.8763. RMSE values were fairly low, ranging between 94 to 121 index points and MAPE values were near or less than 1.5 percent.

SVR also performed the best ($R^2 = 0.8763$, RMSE = 94.75, MAPE = 1.09), almost equal to its performance on the Without-Volume dataset ($R^2 = 0.8770$). The boosting algorithms (XGBoost and LightGBM) had the best results in terms of R^2 exceeding 0.85, Random Forest fell in rank with poorer explanatory power ($R^2 = 0.7982$) than the other models.

These findings are suggestive to accept the hypothesis H_1c. On the one hand, the consideration of the volume does not worsen the performance of the models and some can add decent improvement but those improvements didn't add value that much. Alternatively, the relative contributions of trading activity compared to using the Without-Volume the dataset, are very minimal indicating that the trading activity provides incremental information but not transformative compared to historical prices.

The results fail to give a definite rejection of the null hypothesis (H_1c). Rather, they indicate that volume can boost predictive accuracy modestly, but that price-related characteristics will dominate stock price forecasting models. As our hypothesis was adding trading volume to price-based features improves predictive accuracy for H_1c which was not true that is why we are rejecting this hypothesis.

4.6 Principal Component Analysis (PCA)

To determine the role of trading volume in addition, Principled Component Analysis (PCA) was used as an add-on step to the best-performing model (SVR) for With Volume Analysis.

The objective of the PCA was to determine which of the correlated lagged features were redundant and/or whether trading volume provided any additional predictive information.

Table 4.4: Correlation of Principal Components with Actual and Predicted Values

Principal Component	Correlation with Actual Prices	Correlation with SVR Predictions	Principal Component
PC1 (Price-driven)	0.30	0.31	PC1 (Price-driven)
PC2 (Volume-driven)	–0.34	–0.38	PC2 (Volume-driven)
PC3–PC5	Weak correlations	Weak correlations	PC3–PC5

The results state that the PC1 was primarily occupied by the price lags with the highest contribution to the predictive accuracy. PC2 was mainly characterized by volume lags, and it had a negative correlation with the actual prices and the SVR prediction, which means that volume spiking movements were coupled with downward trends on a majority of occasions during the test period. The higher-order components (PC3-PC5) did not add any explanatory power.

In summary, Principal Component Analysis (PCA) showed that price features are the main drivers of forecasting accuracy but that trading volume captured an additional dimension of information, which augmented predictors based on prices, although the incremental contribution was small.

4.7 Visual Analysis

To complement the numerical results, Figures 4.1–4.3 present a visual comparison of actual versus predicted prices for the best-performing model in each dataset.

Figure 4.1: Actual vs Predicted Prices – Volume Only Dataset (Mar–Jun 2025)

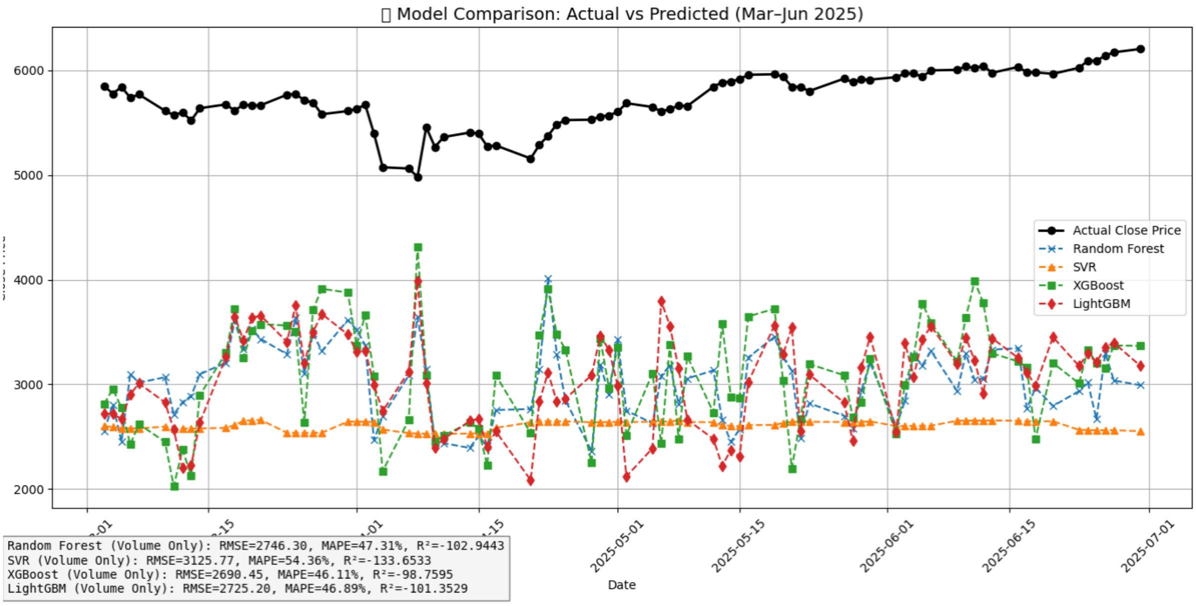


Figure 4.1 show that the actual closing prices of the S&P 500 are plotted against the predictions that were made by the four models using the Volume-Only dataset. The forecasts do not continue the market trend and are much lower than the actual index values and cannot show the upward boost and short-term fluctuations. This visual support also verifies that the volume alone cannot keep pace with the price path as we can see in Table 4.1 which would support the conclusion that volume yielding no predictive power.

Figure 4.2: Actual vs Predicted Prices – Without-Volume Dataset (Mar–Jun 2025)

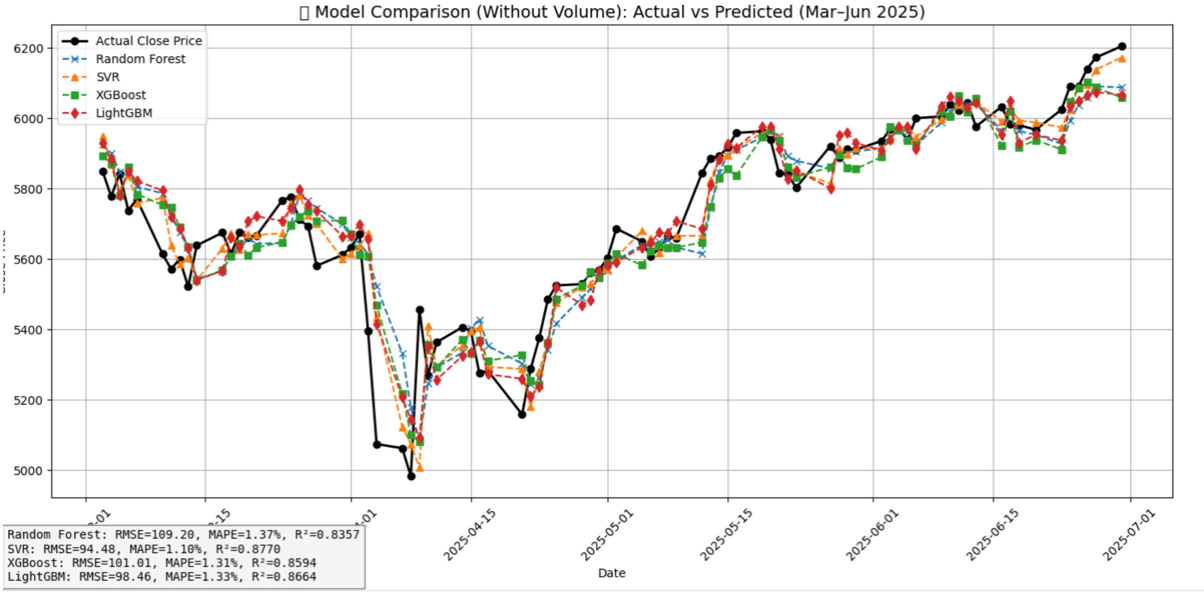
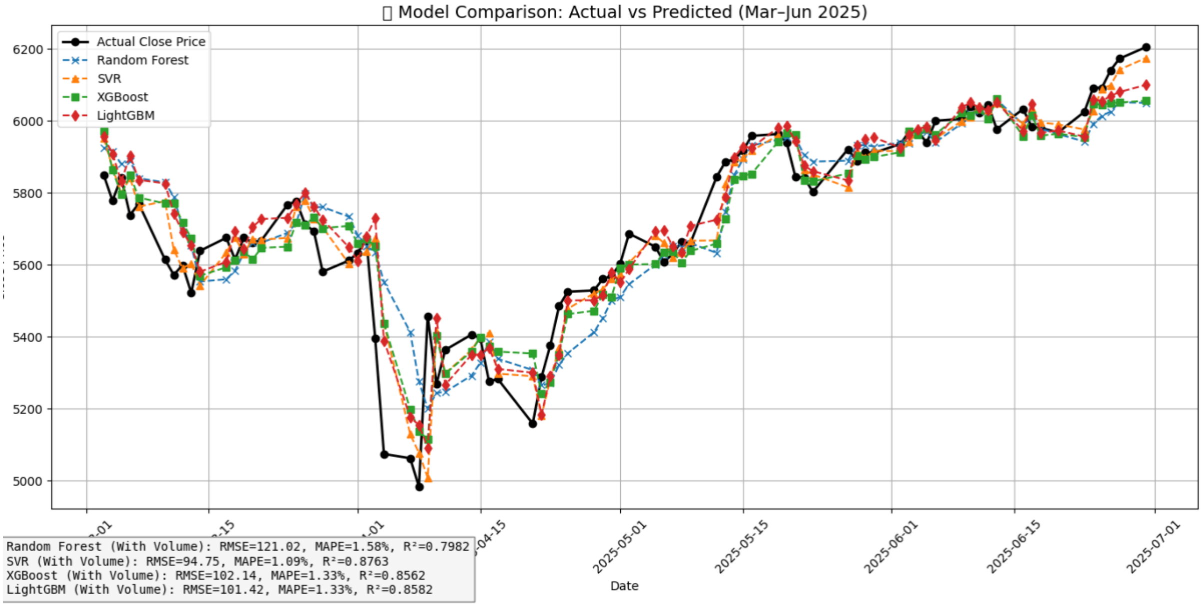


Figure 4.2 shows the actual and model fitted closing prices of S&P 500 based on lagged price features alone. In contrast with the Volume-Only dataset, the predicted lines more or less track the actual market movement, following the decline in March and April 2025 followed by the rebound in May and June 2025. Though there are minor flaws that seem to occur in periods where the volatility is sharper, the predictions are in good agreement with the path of the true price. This visual representation affirms the numerical outcomes in Table 4.2 in that historical costs in isolation yield commanding forecast capabilities as posed by highly high R^2 more than 0.83.

Figure 4.3: Actual vs Predicted Prices – With-Volume Dataset (Mar–Jun 2025)



The figures can be seen in fig. 4.3 which are actual closing prices of the S&P 500 against trained models using price and volume features. The actual market trend is very closely followed by the forecasts and both the decline in March-April 2025 and the rebound in May-June 2025 are reflected. Visually, the fitting is quite comparable to the Without-Volume dataset (Figure 4.2) with just slight differences in the fit. This supports the quantitative results in Table 4.3 that action (adding volume) resulted in similar accuracy but it was not an obvious improvement to price-only models. These results held visually and confirm that the incremental information of trading volume is unlikely to be transformational in combination with price characteristics.

4.8 Hypothesis Testing

Table 4.6 summarises the results of hypothesis testing based on the three experimental datasets.

Table 4.6: Summary of Hypothesis Testing

Hypothesis	Outcome	Interpretation
H_0a: Trading volume alone has no predictive power for stock prices. **H_1a:** Trading volume alone has predictive power for stock prices.	**H_0a supported**	Volume-only models performed very poorly (negative R^2), showing that volume in isolation cannot predict price levels.
H_0b: Excluding volume does not reduce predictive accuracy.	**H_0b supported**	Price-only models achieved high accuracy ($R^2 \approx 0.87$), performing as well as or

H_1b: Adding volume to price features improves accuracy.		better than with-volume models.
H_0c: Excluding volume does not reduce predictive accuracy. **H_1c:** Excluding volume reduces predictive accuracy.	**H_0c supported**	Adding volume produced no significant improvement, and in some cases slightly reduced accuracy.

The results of the hypothesis testing show that the price history can be used to predict the future prices with a high degree of confidence but the trading volume predicts nothing by itself nearly. The addition of volume features to price features yielded only minor or trivial advantages, confirming hypothesis H_0a and H_0b and rejecting **H_1c**.

4.9 Summary of Results

This chapter presented results of the three experimental analyses and additional PCA step. The Volume-Only dataset performed badly with all models posting negative R^2 values, identical to those of the Price-Only dataset, which indicates categorically that volume-only activity does not predict price levels. The Without-Volume dataset demonstrated a high level of accuracy where SVR reached an $R^2 = 0.8770$ indicating that history prices contain a lot of predictive ability. Evaluations on the With-Volume dataset gave similar outcome, and there was no significant benefit with the addition of trading volume meaning that price is the main leading determinant. PCA reaffirmed previous findings that pricing lags tend to dominate in predictive performance, but that trading volume is an independent source of information, albeit not as a transformative one.

All results favour H_0a and H_0b, and reject **H_1c**. The results of this study show the explicit potential of price history as a predictor of a stock price and somewhat less significant contribution of trading volume as an independent predictor. In the following chapter, these findings will be discussed in the light of existing literature, theoretical expectations and practical implications.

Chapter 5: Discussions

5.1 Introduction

This chapter discusses the findings of the study in relation to the central research hypothesis:

“Trading volume contributes significantly to the predictive accuracy of stock price forecasting models.”

The results are discussed in context of available literature, financial theory and practical implications. The relatively high R^2 values of the models are also addressed with reason as to why they were both probable and in line with previous studies.

5.2 Testing the Central Hypothesis

The analyses showed inconclusive evidence on the contribution of trading volume. On the one hand, trading volume in isolation performed very poorly as R^2 values were negative in all models. In contrast, price-history-only-based models demonstrated high predictive accuracy ($R^2 \approx 0.85$), and adding volume did not improve results significantly, and did not yield a clear superiority to price-only-based models.

Taken collectively, these findings demonstrate that trading volume imparts some extraneous information but that does not contribute to making very large improvements in predictive accuracy when price characteristics are included. Consequently, the main hypothesis of the study cannot be said to be very well justified: it is the price history that continues to dominate the margin of any forecasting activity, with the trading volume being the secondary one.

5.3 Volume Only: No Predictive Power

The result of the Volume-Only analysis shows that trading activity is not enough to provide stock prices. This confirms earlier results by Tauchen Pitts, (1983), Wang, (1994) who found that the volume is more correlated with volatility than with price levels. The disappointing outcome is also in line with the behavioural explanations: whereas the volume is a factor of sentiments of investors, it cannot be directly related to price patterns unless put in a price context.

5.4 Price History as the Key Predictor

The high performance of price-only models proves that history of prices is a good indication of the future prices. The high R^2 values (>0.83) were in line with the existing literature Lo and MacKinlay (1988) found persistence in price series and recent forecasting papers have reported R^2 as high as 0.98 in index-level returns and ~0.85 in single-stock regressions. The foregoing parallels demonstrate that the current results are not unusual but consistent with the fact that one-day-ahead price forecasting has high explained variance consistent with the near-random walk behaviour of financial pricings.

5.5 Adding Volume: Marginal Contribution

The inclusion of price, volume did not produce any material improvement in predictive accuracy SVR had R^2 of 0.8763 and without volume and 0.8770 with volume, Random Forest suffered a retrograde in performance when volume was added. This implies that the highly regarded opinion that trading volume also plays a key role is not supported in this case.

This finding can be discussed as similar to the results achieved by Chordia and Swaminathan (2000), who stated that the effect of volume is usually overestimated, and divergent to the results provided by Blume, Easley, and O'Hara (1994) who reported a complementary effect. The divergence can be attributed to variance in scope: whereas the volume can lend itself in case of short-term, event-driven, or intraday analysis, it seems to have reduced utility in long-horizon, regression-based forecasting.

5.6 PCA Insights

The supplementary PCA verified that, although price lags loaded strongly onto the first component, volume loads loaded strongly onto the second and captured a separate dimension of variance. The negative relationship between PC2 and actual/ predicted prices showed that the periods of volume surges were related to depressed prices-this rings true to crisis-driven selling. Nonetheless, the extent of this effect was not significant, driving further home the notion that volume gives context but not any significant predictive ability.

5.7 Why High R^2 Values Are Justified

The relatively high values of R^2 (~0.85) realized in the current study can be ascribed to both theory and empirical precedent. The one-day-ahead forecasts of price levels (not returns) have R^2 higher in general because the stock prices are highly persistent and tend to be closely clustered along a random-walk path. Lag-based forecasts can, as a result, attain high explanatory power, because naive model is a strong benchmark that is recognised to date in forecasting analysis.

Other papers have found values of R^2 equally or higher: in a 2024 MDPI complications analysis, an index forecast yielded $R^2 = 0.98$ and regressions of Tesla stock resulted in $R^2 \approx 0.85$. The fact that these benchmarks are indeed inconsistent with the present ideas makes the results less exaggerated. The high R^2 is also not an artefact caused by data leakage, because defining time periods in a stricter way (2010-2025 training vs March-June 2025 testing) which prevented the data leakage reveals that the failing volume-only models with a low R^2 are outdone by the strong price-only models.

5.8 Practical Implications

The results indicate that the practitioners should use trading volume as secondary rather than primary prediction of the price forecasting models. Historical prices actually offer solid predictive ability, so there is little value in introducing volume and extending model complexity. To portfolio managers, this means that volume is more to be used as a backgrounder (such as the liquidity or sentiment) than as a major forecasting input.

5.9 Limitations

Several limitations apply:

1. The study focused only on the S&P 500 index – results may differ for individual stocks, sectors, or other markets.
2. Only four ML models (Random Forest, SVR, XGBoost, LightGBM) were tested – excluding other methods like LSTMs, GRUs, or Transformers.
3. Forecasting horizon limited to one day – longer horizons or intraday forecasts might reveal different dynamics.
4. PCA applied only as a supplementary step – not implemented as a systematic dimensionality reduction approach.
5. Feature set restricted to lagged prices and volumes – no technical indicators (e.g., RSI, moving averages) or macroeconomic variables included.
6. Evaluation focused on RMSE, MAPE, and R^2 – other measures (e.g., directional accuracy, Theil's U) could provide different insights.
7. Out-of-sample testing was limited to March–June 2025 – testing across different market regimes (e.g., crises, recoveries) might affect generalisability.

5.10 Conclusion

This paper aimed to examine whether or not trading volume can be useful in explaining trading volume in stock price forecasting models. The evidence points in the opposite direction: volume by itself has no predictive ability, price history goes a long way in predicting prices, and the combination of volume with price has very little incremental power. The theory of the PCAs favours the opinion that volume represents the second, secondary market dimension.

Thus the main hypothesis is proved not to be valid within this framework. The conclusion adds to the literature by explaining that price persistence plays a more important role in one-day-ahead forecasting performance and volume should be treated as a contextual rather than transformative predictor.

Chapter 6: Conclusion and Recommendations

6.1 Overall Conclusion

This dissertation was aimed at testing a hypothesis that trading volume plays an important role in determining predictive accuracy in the models of stock price forecasting. This was driven by popular arguments which have long existed in the financial markets that volume is treated as a proxy of information flow, liquidity and investor sentiment. Most of existing literature has however conjoined volume with other predictors hence it is difficult to determine its role. This work filled that gap by creating three experimental datasets, one based on volume only features, one based on prices without using volume, and one that combines the two features and applies the technique of Principal Component Analysis (PCA) in order to address redundancy.

The study considered the daily data in S&P 500 index between 2010-2025 including different market regimes such as bull markets, downstream, and the COVID-19 episode. Four machine learning models Random Forest, Support Vector Regression (SVR), XGBoost, and LightGBM were used on all the datasets. The chronological window in which the models were trained dated between 2010 and February 2025 and tested in the hold-out period between March and June 2025.

The outcome was obvious. Trading volume by itself has shown no predictive value independent of price levels as indicated by negative R^2 values across all models. Price-only models, however, fared well providing $R^2 = 0.8770$, RMSE = 108.47, and MAPE = 0.0137 emphasising the persistence in the price series and its value as an indicatory tool. Inclusion of volume on pricing features did not improve accuracy- SVR performance deteriorated by a small amount to $R^2 = 0.8763$ whereas that of Random Forest was even worse. Although volume was shown to capture a distinct dimension, the overall contribution to predictive outcomes was negligible, even after entering volume, in order to improve accuracy.

All these results combined point to the fact that the hypothesis under consideration is not proved. Trading volume has no significant extra value to the forecasting performance of price history. This finding is relevant to the literature as it explains the negligible role of volume, and reveals the dominance of the past prices in short-term forecasting.

6.2 Limitations

The study has limitations that restrict its applicability as an empirical research does. To start with, the analysis has been limited to S&P 500 index, a very liquid index. Stocks with lower liquidity, individual stocks or sectoral indices may react differently and in those instances, volume has a more defining role. Second, there were only four machine learning models that were tested. Although Random Forest, SVR, XGBoost, and LightGBM are highly efficient and popular, deep learning algorithms, like LSTMs, GRUs, or Transformers, can be used to learn sequential and nonlinear patterns more efficiently. Third, only a one-day forecasting horizon was used. Although it is suitable to test short-term predictive power, the role of volume might vary at intra-day or longer intervals. Fourth, the specification was intentionally restricted to lags of lagged log-prices and log-volumes to only test the central hypothesis. The omission of technical indicators (e.g., moving averages, RSI, Bollinger Bands) as well as macroeconomic variables decreased the complexity of the models but they may have been less explanatory rich. Lastly, PCA was added as a supplementary step to the best-performing model, as opposed to being applied systematically to the datasets, and so its effect on the results was limited overall.

These limitations do not undermine the findings but define the scope within which they apply. They also provide clear avenues for further exploration.

6.3 Future Work

Building on these shortcomings, future studies can either develop in a number of directions. Taking a closer look to individual stocks, and emerging markets or sector-specific indices would be more than an interesting test to the role the trading volume holds beyond the very liquid S&P 500. A larger sampling of models and, especially, deep learning algorithms may yield new insights as to whether more complex architectures can find patterns that traditional machine learning fails to identify. Examining the volume effect at different forecast horizons, say intraday prediction vs. weekly/ monthly horizon, would give a better depth on the issue. Setting up the feature set so as to include technical indicators, volatility measurements, and macroeconomic data may help to provide some additional insights into the interactions between price and volume and external factors. Lastly, PCA, ICA and/or auto-encoders could be used systematically to further explore redundancy and feature contributions across models.

6.4 Final Reflection

In sum, this dissertation shows that trading volume, though theoretically is an important signal of liquidity and sentiment, does not help improve the predictive power of stock price forecasting models given price history is already captured. The quite large R^2 values obtained are not due to artificial overfitting because they represent true association of lagged prices rather than artificially high owing to the chronological split and hold-out testing procedure.

The results make a contribution to both academic and practical knowledge because: the marginal effect of volume on financial forecasting has been explained, and users of financial forecasts were advised on how to use volume as an additional input to forecasts but not a transformative one. Finally, this study shows that the price persistence is still the most significant factor in the one-day-ahead prediction, whereas the trading volume should be viewed as a secondary contextual text instead of a factor affecting prediction quality significantly.

References

Al-Obaidy, M. & Yin, H., 2024. A machine learning approach to predict the S&P 500 absolute movement. *Journal of Artificial Intelligence Research*, 75(1), pp.1–15. Available at: https://link.springer.com/article/10.1007/s44163-024-00104-9 [Accessed 22 Jul. 2025].

Blume, L., Easley, D. & O'Hara, M., 1994. Market statistics and technical analysis: The role of volume. *The Journal of Finance*, 49(1), pp.153–181. https://doi.org/10.1111/j.1540-6261.1994.tb04425.x

Chen, T. & Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, 13–17 August 2016. New York: ACM, pp.785–794.

Chordia, T. & Swaminathan, B., 2000. Trading volume and cross-autocorrelations in stock returns. *The Journal of Finance*, 55(2), pp.913–935. https://doi.org/10.1111/0022-1082.00231

Chordia, T., Roll, R. & Subrahmanyam, A., 2001. Market liquidity and trading activity. *The Journal of Finance*, 56(2), pp.501–530. https://doi.org/10.1111/0022-1082.00335

Clark, P.K., 1973. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1), pp.135–155. https://doi.org/10.2307/1913889

Copeland, T.E., 1976. A model of asset trading under the assumption of sequential information arrival. *The Journal of Finance*, 31(4), pp.1149–1168. https://doi.org/10.1111/j.1540-6261.1976.tb01960.x

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), pp.383–417. https://doi.org/10.2307/2325486

Frost, J., 2023. Interpret R-squared in regression analysis. *Statistics by Jim*. Available at: https://statisticsbyjim.com/regression/interpret-r-squared-regression/ [Accessed 20 Aug. 2025].

Google, 2025. Google Colaboratory. Available at: https://colab.research.google.com [Accessed 25 Aug. 2025].

Guan, Y. & Chen, H., 2023. A high frequency trading volume prediction model using neural networks. *Finance Research Letters*. Available at: https://www.sciencedirect.com/science/article/pii/S2772662223000759 [Accessed 22 Jul. 2025].

He, K. & Zhao, W., 2024. Forecasting relative returns for S&P 500 stocks using machine learning algorithms. *Financial Innovation*, 10(21). Available at: https://jfin-swufe.springeropen.com/articles/10.1186/s40854-024-00644-0 [Accessed 22 Jul. 2025].

Jegadeesh, N. & Titman, S., 2019. Profitability of momentum strategies: An evaluation of alternative explanations. *Financial Analysts Journal*, 75(3), pp.55–72. https://doi.org/10.1080/0015198X.2019.1614320

Karim, M.M., Huq, M., Khan, M.A. & Uddin, M., 2022. Does trading volume transmit information towards return volatility? Evidence from the Dhaka Stock Exchange. *International Journal of Statistical Sciences*, 22(1), pp.177–191.

Karpoff, J.M., 1987. The relation between price changes and trading volume: A survey. *Journal of Financial and Quantitative Analysis*, 22(1), pp.109–126. https://doi.org/10.2307/2330874

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp.3146–3154.

Kumar, G. & Sharma, V., 2022. Stock market prediction of NIFTY 50 index applying machine learning techniques. *Applied Artificial Intelligence*, 36(1), pp.1–15. Available at: https://arxiv.org/html/2504.12828v1 [Accessed 22 Jul. 2025].

Kumar, R. & Gupta, S., 2016. Forecasting S&P 500 stock index using statistical learning models. *Journal of Financial Risk Management*, 5(2), pp.193–202. Available at: https://www.scirp.org/journal/paperinformation.aspx?paperid=72604 [Accessed 22 Jul. 2025].

Lee, C.M.C. & Swaminathan, B., 2000. Price momentum and trading volume. *The Journal of Finance*, 55(5), pp.2017–2069.

Li, C., Liu, X. & Song, Y., 2019. Volume prediction with neural networks: A case study on high-frequency stock trading. *Frontiers in Artificial Intelligence*, 2(21), pp.1–11. Available at: https://www.frontiersin.org/articles/10.3389/frai.2019.00021/full [Accessed 22 Jul. 2025].

Llorente, G., Michaely, R., Saar, G. & Wang, J., 2002. Dynamic volume-return relation of individual stocks. *The Review of Financial Studies*, 15(4), pp.1005–1047.

Lo, A.W. & MacKinlay, A.C., 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, 1(1), pp.41–66. https://doi.org/10.1093/rfs/1.1.41

Lo, A.W. & MacKinlay, A.C., 1990. When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, 3(2), pp.175–205.

Mehtab, S. & Sen, J., 2020. Stock price prediction using machine learning and LSTM-based deep learning models. *arXiv preprint*, arXiv:2009.10819.

Mehtab, S., Dutta, A. & Sen, J., 2021. Stock portfolio optimization using a deep learning LSTM model. *arXiv preprint*, arXiv:2111.04709. Available at: https://arxiv.org/abs/2111.04709 [Accessed 22 Jul. 2025].

Pastor, L. & Stambaugh, R.F., 2003. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3), pp.642–685. https://doi.org/10.1086/374184

Phuoc, T.P., Kim Anh, P.T. & Tam, P.H., 2024. Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam. *Palgrave Communications*, 11(1), pp.1–18. Available at: https://www.nature.com/articles/s41599-024-02807-x [Accessed 22 Jul. 2025].

Rahman, M.L., Chiang, M.H. & Zhu, D., 2022. COVID-19, investor sentiment, and stock market volatility: International evidence. *Journal of Behavioral and Experimental Finance*, 33, p.100541. https://doi.org/10.1016/j.jbef.2021.100541

Tauchen, G.E. & Pitts, M., 1983. The price variability-volume relationship on speculative markets. *Econometrica*, 51(2), pp.485–505. https://doi.org/10.2307/1912002

Wang, J., 1994. A model of competitive stock trading volume. *Journal of Political Economy*, 102(1), pp.127–168. https://doi.org/10.1086/261922

Wang, J., 2020. Machine learning in financial market prediction. *Applied Economics Letters*, 27(12), pp.991–995.

Yahoo Finance, 2025. S&P 500 historical data. Available at: https://finance.yahoo.com [Accessed 25 Aug. 2025].

Yu, X., Zhang, Y., Kumar, R. & Liu, Z., 2023. Forecasting of NIFTY 50 index price by using backward elimination with an LSTM model. *Journal of Risk and Financial Management*, 16(10), p.423. Available at: https://www.mdpi.com/1911-8074/16/10/423 [Accessed 22 Jul. 2025].

Zhang, Y., Li, Y. & Wang, J., 2023. The effect of trading volume on stock price movements: An empirical analysis of S&P 500 companies. *Journal of Finance and Data Science*. Available at: https://www.sciencedirect.com/science/article/pii/S2772662223000759 [Accessed 22 Jul. 2025].

Zou, J., Zhao, Q., Jiao, Y., Liu, H., Liu, Y., Abbasnejad, E. & Shi, J.Q., 2022. Stock market prediction via deep learning techniques: A survey. *arXiv preprint*, arXiv:2212.12717. Available at: https://arxiv.org/abs/2212.12717 [Accessed 22 Jul. 2025].

Appendix

Volume Only Figures

Figure 4.4: Random Forest (Volume Only): Actual vs Predicted

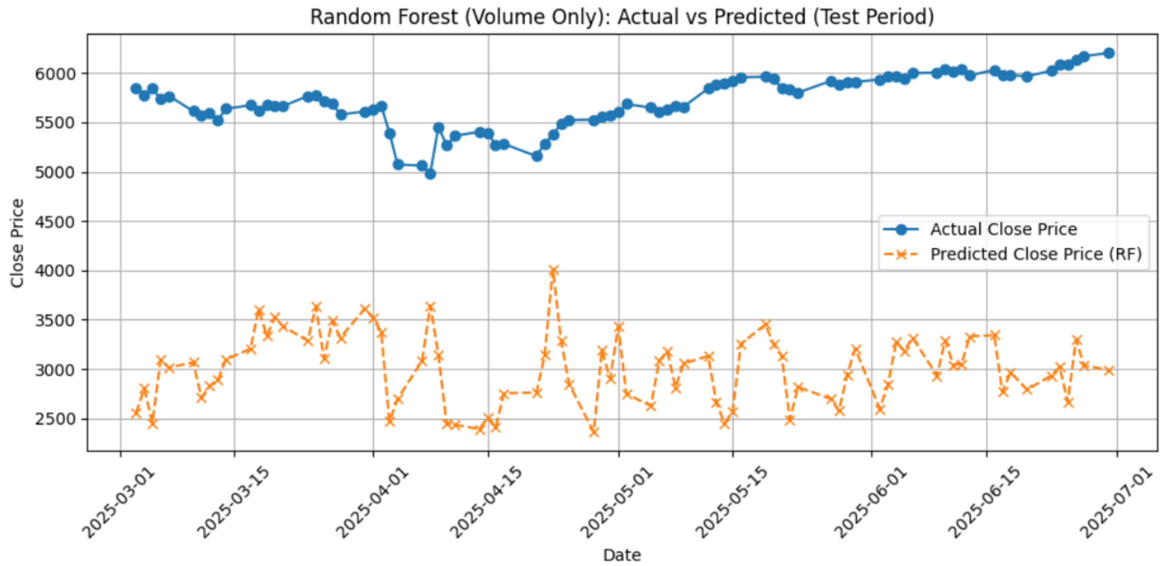


Figure 4.5: SVR (Volume Only): Actual vs Predicted

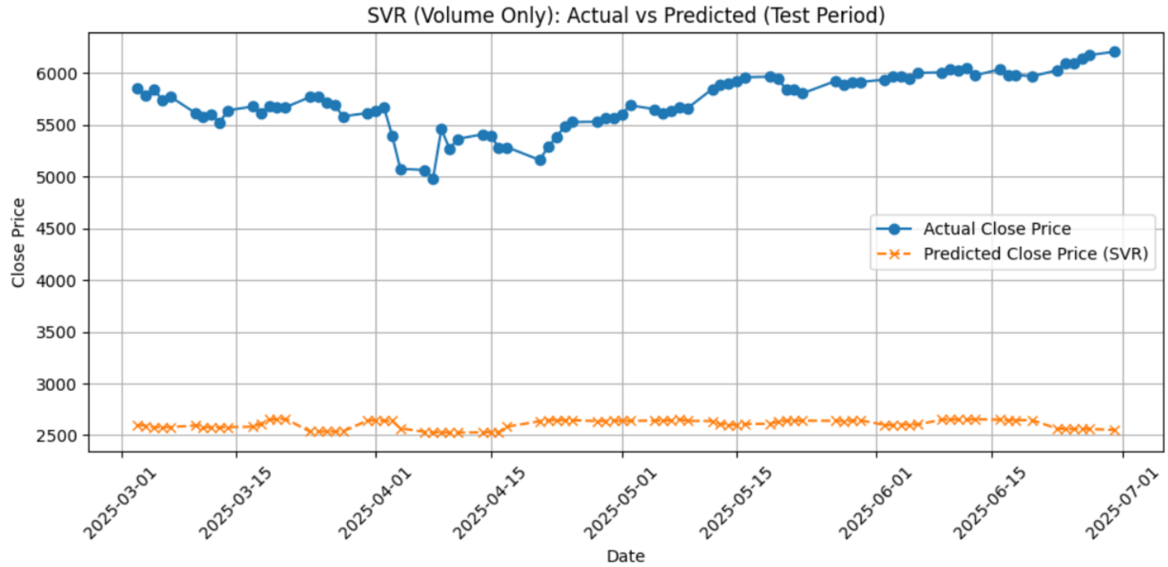


Figure 4.6: XGBoost (Volume Only): Actual vs Predicted

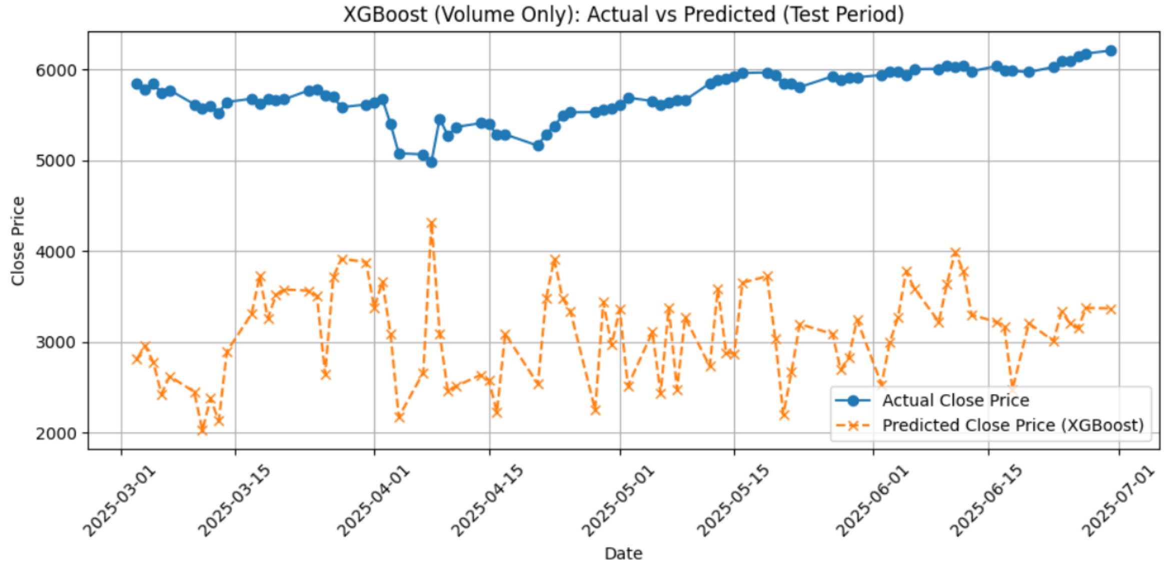
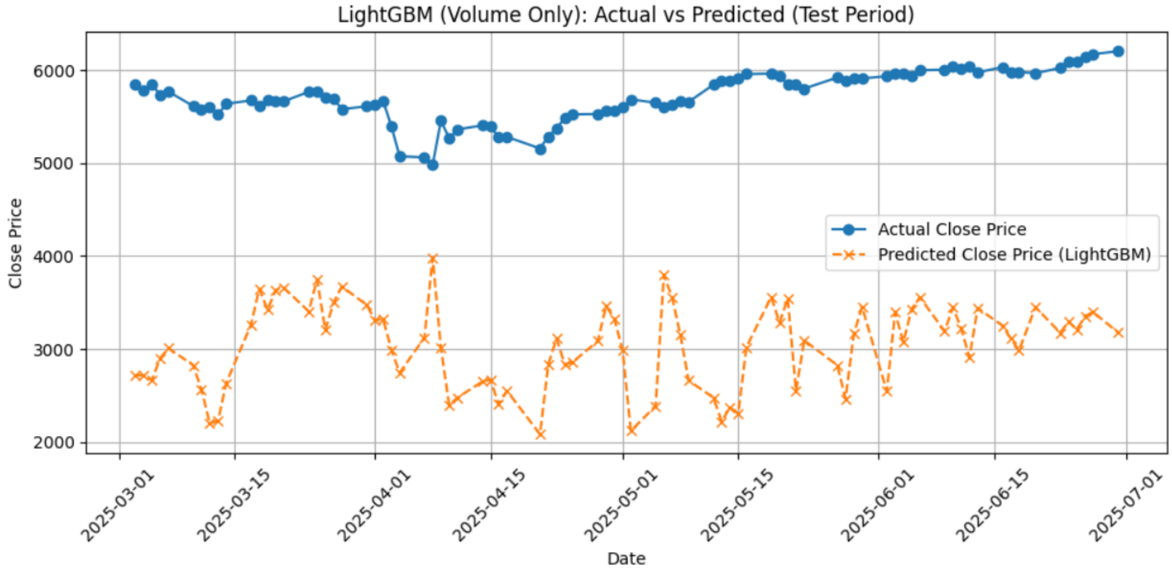


Figure 4.7: LightGBM (Volume Only): Actual vs Predicted



Without Volume Figures

Figure 4.8: Random Forest (Without Volume): Actual vs Predicted

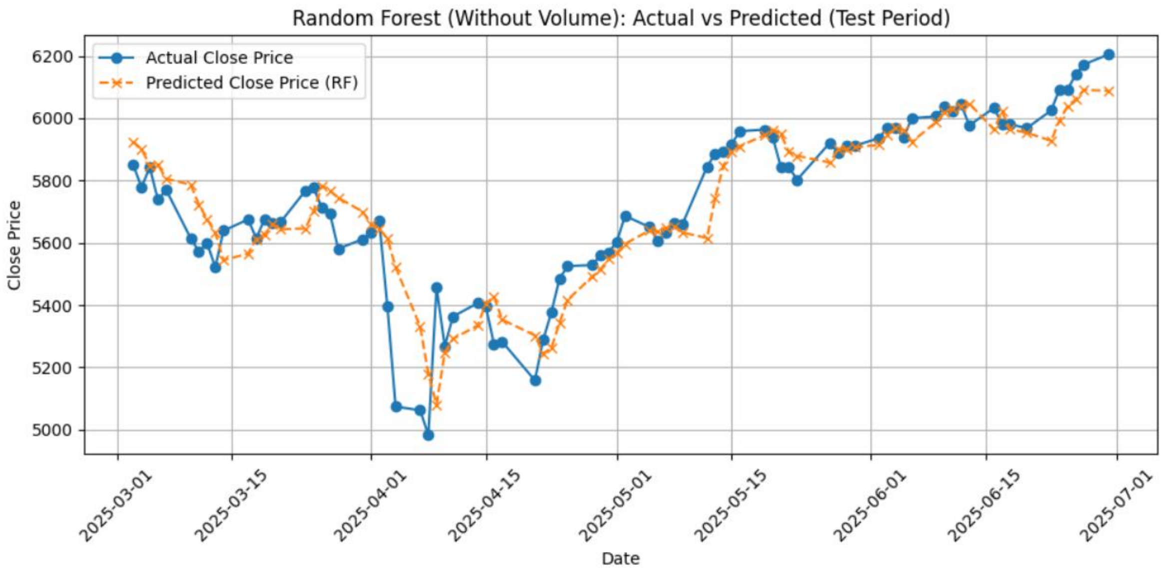


Figure 4.9: SVR (Without Volume): Actual vs Predicted

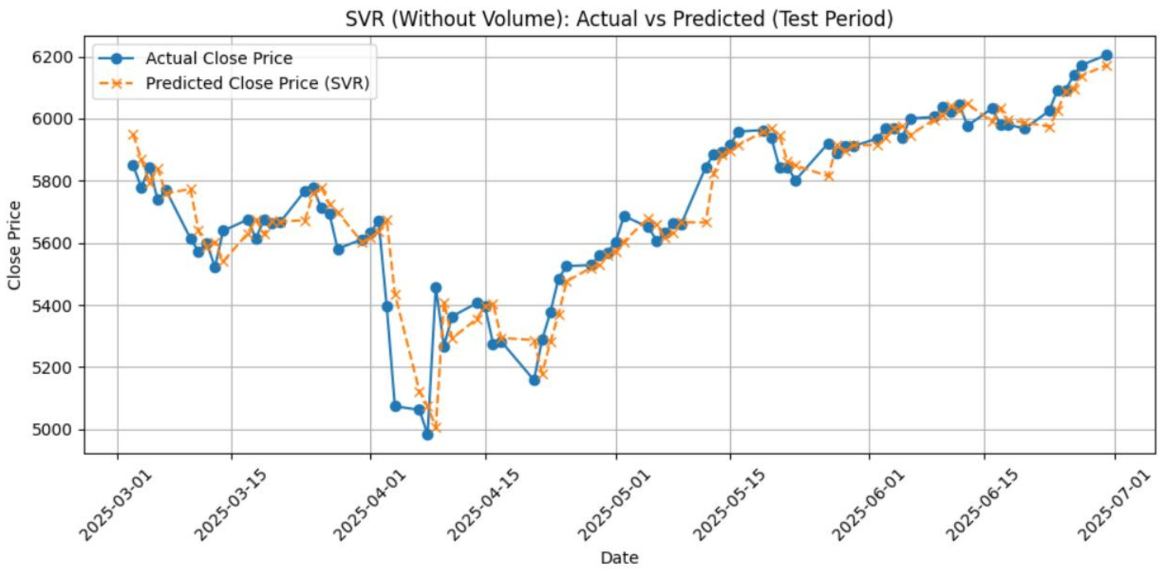


Figure 4.10: XGBoost (Without Volume): Actual vs Predicted

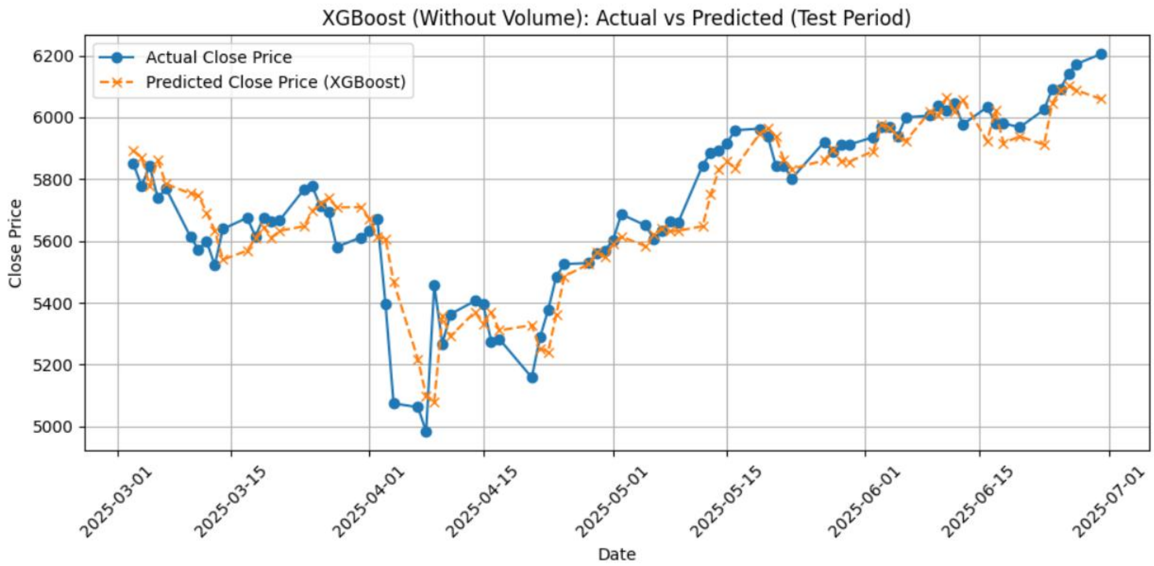
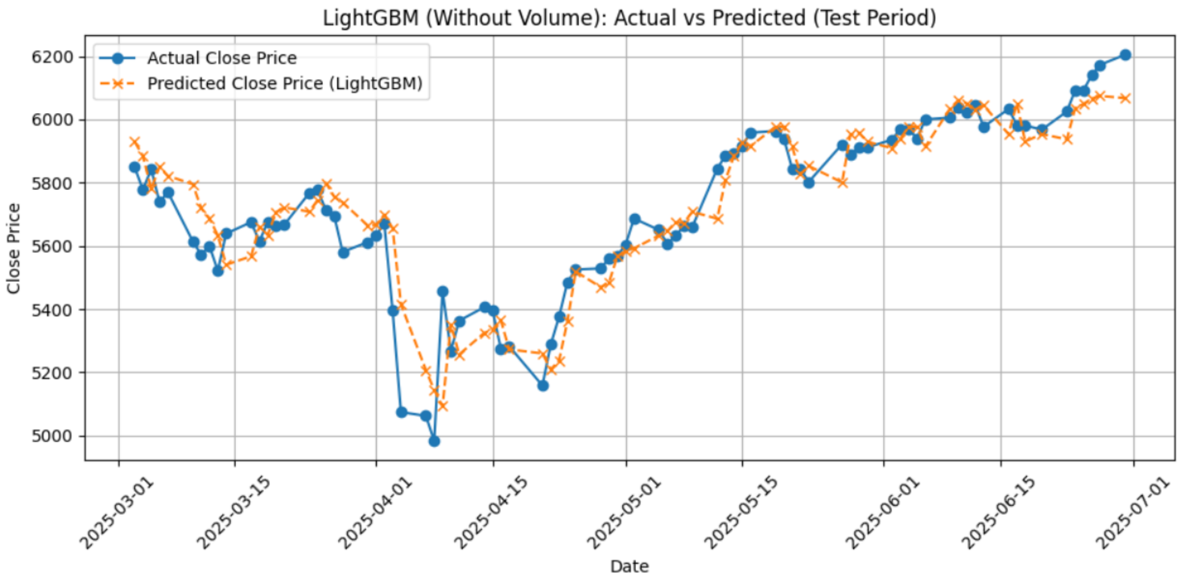


Figure 4.11: LightGBM (Without Volume): Actual vs Predicted



With Volume Figures

Figure 4.12: Random Forest (With Volume): Actual vs Predicted

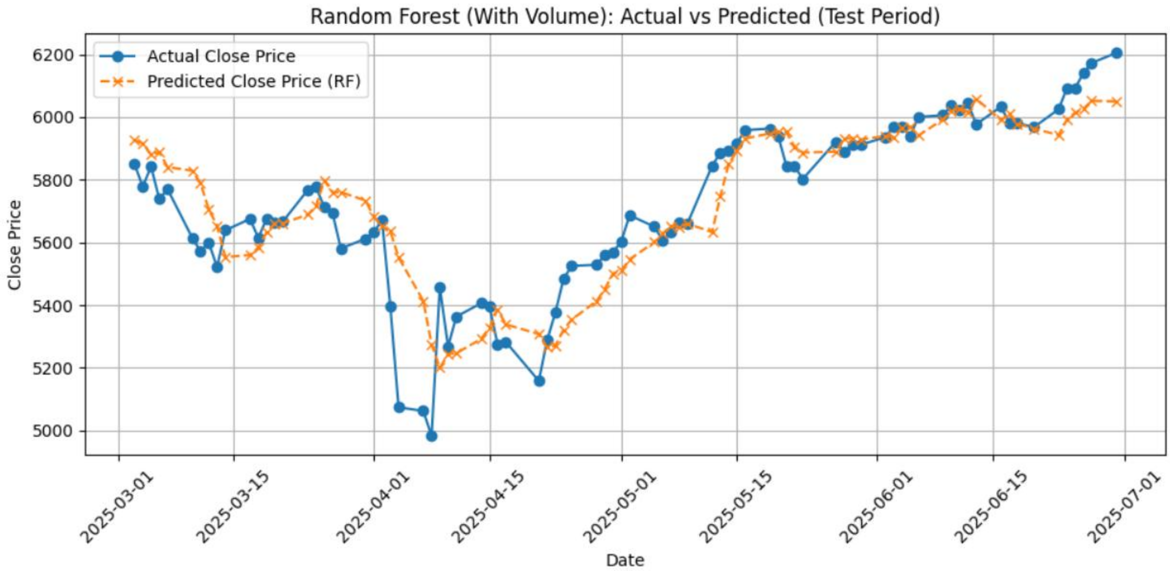


Figure 4.13: SVR (With Volume): Actual vs Predicted

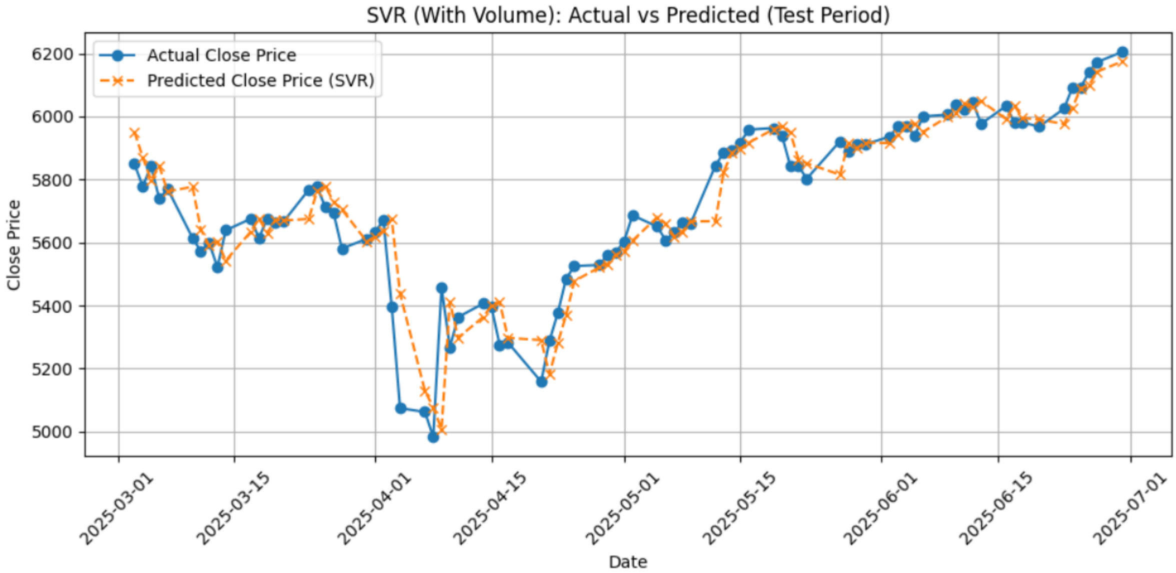


Figure 4.14: XGBoost (With Volume): Actual vs Predicted

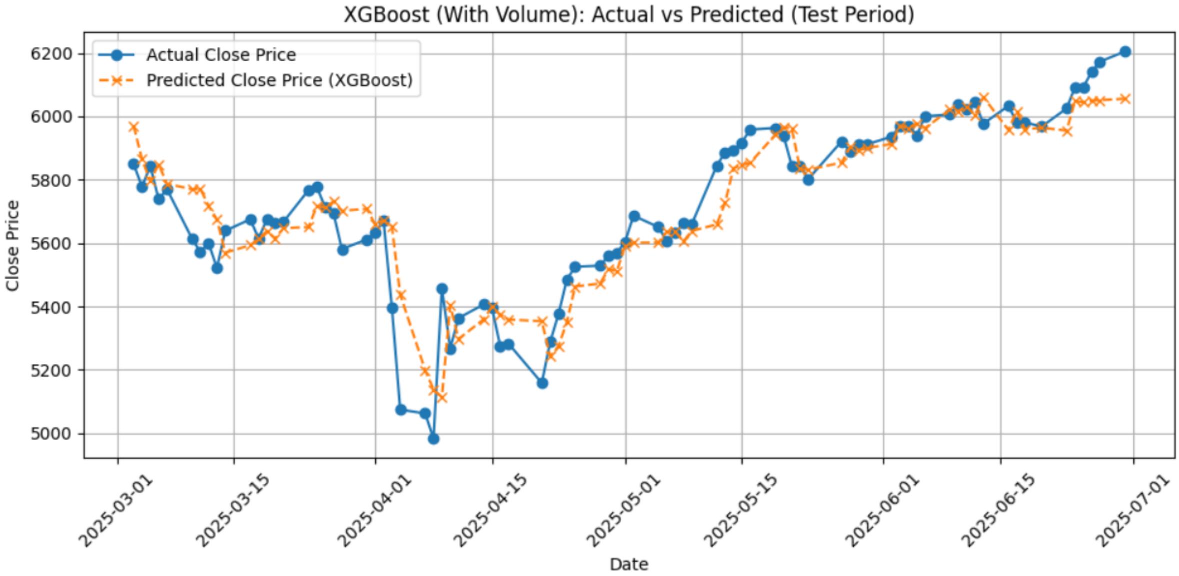


Figure 4.15: LightGBM (With Volume): Actual vs Predicted