BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF STATISTICS AND MATHEMATICS

PREDICTIVE ANALYTICS FOR STOCK PRICE FORECASTING ON THE ZIMBABWE STOCK EXCHANGE: DEVELOPMENT OF THE ZSE INSIGHTS WEB-BASED APPLICATION

BY

TINOTENDA PABWE

B224007B

TO BE SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR BSC.HONOURS IN STATISTICS AND FINANCIAL MATHEMATICS

PROJECT

2026

APPROVAL FORM

This is to certify, that this research project is the result of my own research work and has not been copied or extracted from past sources without acknowledgement. I hereby declare that no part of it has been presented for another degree in this University or elsewhere.

Tinotenda Pabwe		08/06/2026
B224007B	Signature	Date

Certified by

Mr Mukono		08/06/2026
Supervisor	Signature	Date

Dr Magodora		12/08/2026
Chairperson	Signature	Date

DECLARATION

I, Tinotenda Pabwe, declare that:

i. This dissertation is my original work and has not been submitted to any other university or institution for the award of any degree or professional qualification.

ii. Where the work of other researchers has been used, it has been duly acknowledged through in-text citation and referencing in accordance with the APA 6th edition referencing format.

iii. Artificial intelligence assisted tools were used during the implementation stage of this research to support coding assistance, debugging, rapid prototyping, and workflow efficiency. The research design, theoretical framework, methodological decisions, data analysis, interpretation of results, and all academic reasoning and conclusions presented in this dissertation remained the sole intellectual responsibility of the researcher.

Signature: 

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my family for their unwavering support throughout this journey. To my late mother, Nyaradzo Takura, your memory continues to inspire everything I do and this work is dedicated in no small part to you. To my brother Panashe Pabwe, thank you for your patience, motivation, and presence throughout this process and my father whose steadfast encouragement and belief in my potential have been a constant source of strength.

Also,I extend my sincere appreciation to my supervisor, Mr Mukono and co-supervisor Ms Hlupo, for their guidance, constructive feedback, and valuable academic support throughout the course of this research. Their direction was instrumental in shaping both the quality and direction of this dissertation.

I thank the Department of Mathematics and Statistics at Bindura University of Science Education for providing the academic environment and resources that made this study possible.

Special recognition goes to the Custodial Services Department at CBZ Holdings, and in particular to Tinashe Mapara and Tapiwanashe Muchanyangi, whose mentorship and practical exposure to capital markets operations provided the professional foundation that directly informed and inspired this research.

ABSTRACT

Stock price forecasting on the Zimbabwe Stock Exchange presents distinctive challenges attributable to currency instability, thin trading volumes, and structural breaks in equity price series that limit the reliability of conventional econometric models. Existing ZSE studies remain dominated by ARIMA and GARCH-based approaches, with limited comparative application of ensemble and deep learning methods to stock price level forecasting. This study addresses that gap by implementing and evaluating Random Forest, XGBoost, Long Short-Term Memory (LSTM), and Prophet on daily closing price data for ten selected ZSE-listed companies over the post-ZiG period from May 2024 to November 2025.

Data preparation encompassed missing value treatment, outlier detection, Min-Max normalisation, and feature engineering incorporating lagged price variables, rolling means, and volatility indicators. Models were evaluated using RMSE, MAE, R^2, and MAPE on a held-out test set under a consistent chronological train-test split.

XGBoost achieved the strongest predictive performance, recording an RMSE of 0.73, MAE of 0.53, R^2 of 0.999, and MAPE of 0.03%, followed by Random Forest with RMSE of 12.84 and MAPE of 0.42%. LSTM delivered moderate performance constrained by the limited post-ZiG training window, recording RMSE of 18.86 and MAPE of 1.03%. Prophet performed weakest across all metrics, recording RMSE of 157.17, R^2 of -7.74, and MAPE of 9.87%, confirming its structural unsuitability for short-term ZSE price level forecasting.

The ZSE Insights web application was developed using Streamlit, integrating all four models within a three-tier interactive platform enabling recursive multi-step forecasts, real-time model comparison, and price trend visualisations for the selected counters. The findings demonstrate that ensemble machine learning models can generate reliable short-term price forecasts for ZSE-listed equities under post-ZiG conditions, and that purpose-built fintech platforms can translate advanced forecasting research into accessible decision-support tools for investors, analysts, and policymakers in frontier market.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION 1
1.0 INTRODUCTION 1
1.1 BACKGROUND OF THE STUDY 1
1.2 STATEMENT OF THE PROBLEM 3
1.3 RESEARCH OBJECTIVES 4
1.4 RESEARCH QUESTIONS 4
1.5 SCOPE OF THE STUDY 5
1.6 SIGNIFICANCE OF THE STUDY 5
1.7 ASSUMPTIONS OF THE STUDY 6
1.8 LIMITATIONS OF THE STUDY 6
1.9 DEFINITION OF KEY TERMS 7
1.10 CHAPTER CONCLUSION 9
CHAPTER TWO LITERATURE REVIEW 10
2.0 INTRODUCTION 10
2.1 THEORETICAL LITERATURE 10
2.1.1 EFFICIENT MARKET HYPOTHESIS 10
2.1.2 TIME SERIES MODELLING THEORY 11
2.1.3 VOLATILITY MODELLING THEORY 11
2.1.4 MACHINE LEARNING THEORY 12
2.1.5 DEEP LEARNING THEORY 12
2.1.6 ADDITIVE DECOMPOSITION THEORY 12
2.2 EMPIRICAL LITERATURE 12
2.3 LITERATURE SYNTHESIS AND RESEARCH GAP 15
2.4 CONCEPTUAL FRAMEWORK 16
2.5 CHAPTER CONCLUSION 18
CHAPTER THREE: RESEARCH METHODOLOGY 19
3.1 INTRODUCTION 19
3.2 RESEARCH PHILOSOPHY AND DESIGN 19
3.3 DATA SOURCES AND COLLECTION METHODS 20
3.4 TARGET POPULATION AND SAMPLING 21
3.5 DESCRIPTION OF VARIABLES AND EXPECTED RELATIONSHIPS 22
3.6 ANALYTICAL MODELS AND JUSTIFICATION 23
3.7 RECURSIVE MULTI-STEP FORECASTING ENGINE 24
3.8 MODEL PRE-TESTS AND DIAGNOSTIC CHECKS 25

3.9 MODEL FIT AND VALIDATION 26
3.10 ETHICAL CONSIDERATIONS 27
3.11 CHAPTER CONCLUSION 27
CHAPTER FOUR: DATA PRESENTATION, ANALYSIS AND DISCUSSION 28
4.0 INTRODUCTION 28
4.1 DESCRIPTION OF THE ZSE STOCK PRICE DATASET 28
4.1.1 SELECTED ZSE COUNTERS 29
4.1.2 DATA STRUCTURE AND TEMPORAL COVERAGE 30
4.2 EXPLORATORY DATA ANALYSIS 30
4.2.1 FREQUENCY-BASED SUMMARY STATISTICS 30
4.2.2 DESCRIPTIVE STATISTICS OF CLOSING PRICES 31
4.2.3 TIME-SERIES ANALYSIS OF CLOSING PRICES 33
4.2.4 VOLATILITY AND RETURN CHARACTERISTICS 34
4.3 DATA PREPROCESSING AND DIAGNOSTIC TESTS 36
4.3.1 MISSING VALUE ASSESSMENT AND TREATMENT 36
4.3.2 OUTLIER DETECTION AND DATA CONSISTENCY CHECKS 38
4.3.3 DATA NORMALISATION AND TRANSFORMATION 40
4.3.4 CORRELATION ANALYSIS 41
4.3.5 MULTICOLLINEARITY ASSESSMENT 43
4.3.6 FEATURE ENGINEERING AND LAG CONSTRUCTION 44
4.3.7 TRAINING AND TESTING DATA SPLIT 47
4.4 MACHINE LEARNING MODEL RESULTS 47
4.4.1 RANDOM FOREST MODEL RESULTS 48
4.4.2 FEATURE IMPORTANCE ANALYSIS — RANDOM FOREST 49
4.4.3 PREDICTION ACCURACY OF THE RANDOM FOREST MODEL 50
4.4.4 RESIDUAL ANALYSIS OF THE RANDOM FOREST MODEL 51
4.5 EXTREME GRADIENT BOOSTING MODEL RESULTS 52
4.5.1 XGBOOST MODEL ESTIMATION RESULTS 52
4.5.2 FEATURE IMPORTANCE ANALYSIS — XGBOOST 53
4.5.3 PREDICTION ACCURACY OF THE XGBOOST MODEL 54
4.5.4 RESIDUAL ANALYSIS OF THE XGBOOST MODEL 55
4.6 DEEP LEARNING MODEL RESULTS — LONG SHORT-TERM MEMORY 56
4.6.1 LSTM MODEL ESTIMATION RESULTS 56
4.6.2 PREDICTION ACCURACY OF THE LSTM MODEL 57
4.6.3 RESIDUAL ANALYSIS OF THE LSTM MODEL 58

4.7 PROPHET MODEL RESULTS 60
4.7.1 PROPHET MODEL ESTIMATION RESULTS 60
4.7.2 COMPONENT DECOMPOSITION ANALYSIS — PROPHET 61
4.7.3 PREDICTION ACCURACY OF THE PROPHET MODEL 62
4.7.4 RESIDUAL ANALYSIS OF THE PROPHET MODEL 63
4.8 COMPARATIVE PERFORMANCE OF PREDICTIVE MODELS 65
4.8.1 COMPARISON OF RMSE ACROSS MODELS 66
4.8.2 COMPARISON OF MAE ACROSS MODELS 66
4.8.3 COMPARISON OF R^2 ACROSS MODELS 67
4.8.4 COMPARISON OF MAPE ACROSS MODELS 68
4.8.5 RADAR CHART — INTEGRATED COMPARATIVE PERFORMANCE 69
4.8.6 COMPARATIVE PERFORMANCE SUMMARY TABLE 70
4.8.7 DISCUSSION OF MODEL STRENGTHS AND WEAKNESSES 71
4.9 SYSTEM ARCHITECTURE OF THE ZSE INSIGHTS PLATFORM 72
4.9.1 DATA INGESTION AND PREPROCESSING LAYER 73
4.9.2 FEATURE ENGINEERING LAYER 74
4.9.3 ANALYTICAL MODELLING AND FORECASTING LAYER 76
4.9.5 PRESENTATION AND VISUALISATION LAYER 77
4.9.6 SYSTEM WORKFLOW INTEGRATION 78
4.10 DISCUSSION OF FINDINGS 79
4.10.1 DISCUSSION OF FORECASTING MODEL PERFORMANCE 79
4.10.2 DISCUSSION OF THE ZSE INSIGHTS PLATFORM 80
4.10.3 DISCUSSION OF THE RECURSIVE FORECASTING FRAMEWORK 82
4.13 RELATIONSHIP WITH EXISTING LITERATURE 83
4.13.1 COMPARISON WITH ZIMBABWEAN STUDIES 83
4.13.2 COMPARISON WITH REGIONAL AND INTERNATIONAL STUDIES 83
4.13.3 AREAS OF AGREEMENT AND DIVERGENCE 84
CHAPTER FIVE: CONCLUSIONS , IMPLICATIONS AND RECOMMENDATIONS 85
5.1 INTRODUCTION 85
5.2 CONTRIBUTIONS OF THE STUDY 85
5.2.1 EMPIRICAL CONTRIBUTION 85
5.2.2 METHODOLOGICAL CONTRIBUTION 86
5.2.3 PRACTICAL CONTRIBUTION 86
5.3 CONCLUSIONS 87
5.3.1 CONCLUSION FOR OBJECTIVE 1 87

5.3.2 CONCLUSION FOR OBJECTIVE 2 87
5.3.3 CONCLUSION FOR OBJECTIVE 3 87
5.3.4 OVERALL CONCLUSION 88
5.4 IMPLICATIONS OF THE STUDY 88
5.4.1 IMPLICATIONS FOR INVESTORS 88
5.4.2 IMPLICATIONS FOR FINANCIAL INSTITUTIONS 89
5.4.3 IMPLICATIONS FOR POLICYMAKERS AND REGULATORS 89
5.5 LIMITATIONS OF THE STUDY 90
5.6 RECOMMENDATIONS 91
5.6.1 RECOMMENDATIONS FOR INVESTORS 91
5.6.2 RECOMMENDATIONS FOR REGULATORS AND MARKET INSTITUTIONS 91
5.6.3 RECOMMENDATIONS FOR FUTURE RESEARCH 92
5.9 CHAPTER SUMMARY 93
APPENDIX A: SAMPLE RAW DATA STRUCTURE 98
APPENDIX B: MODEL HYPERPARAMETER CONFIGURATIONS 99
APPENDIX C: ZSE INSIGHTS FORECAST TAB 101
APPENDIX D: ZSE INSIGHTS MODEL COMPARISON TAB 102
APPENDIX E: ZSE INSIGHTS ERROR FEATURE AND DIAGNOSTICS TAB 103
APPENDIX F: ZSE INSIGHTS VOLATILITY AND DESCRIPTIVE STATISTICS TAB 104
APPENDIX G: DATA PREPROCESSING PIPELINE 105
APPENDIX H: FEATURE ENGINEERING 108
APPENDIX I: MODEL TRAINING AND EVALUATION 110
APPENDIX J: RECURSIVE MULTI-STEP FORECASTING ENGINE 114
APPENDIX K: MODEL COMPARISON AND VISUALISATION 116

LIST OF FIGURES

Figure 2.1 Conceptual Framework of the ZSE Insights Stock Price Forecasting System 18

Figure 3.1 Architecture of the Recursive Multi-Step Forecasting Engine 25

Figure 4.1 Time-Series Plot of Daily Closing Prices for Five Representative ZSE Counters, May 2024 to November 2025 33

Figure 4.2 Boxplot of Daily Returns for Selected ZSE Counters, May 2024 to November 2025 38

Figure 4.3 Original Versus Normalised Closing Price Distributions for CBZ.ZW ... 40

Figure 4.4 Correlation Heatmap of ZSE Stock Market Variables, Post-ZiG Period . 42

Figure 4.5 Lagged Close Price Versus Current Close Price for Proxy Stock DLTA.ZW 45

Figure 4.6 Rolling 30-Day Volatility for Proxy Stock DLTA.ZW, May 2024 to November 2025 45

Figure 4.7 Random Forest Model Estimation — Actual Versus Predicted Closing Prices for DLTA.ZW, October to November 2025 48

Figure 4.8 Random Forest Feature Importance Ranking 50

Figure 4.9 Random Forest Residual Distribution 51

Figure 4.10 XGBoost Model — Actual Versus Predicted Closing Prices for DLTA.ZW, July 2024 to November 2025 52

Figure 4.11 XGBoost Feature Importance Ranking 53

Figure 4.12 XGBoost Residuals Over Time, July 2024 to November 2025 55

Figure 4.13 XGBoost Residuals Distribution 56

Figure 4.14 LSTM Model — Actual Versus Predicted Closing Prices for DLTA.ZW57

Figure 4.15 LSTM Residuals Over Time 59

Figure 4.16 LSTM Residuals Distribution 59

Figure 4.17 Prophet Forecast Versus Actual Closing Prices for DLTA.ZW, July 2024 to September 2025 60

Figure 4.18 Prophet Component Decomposition — Trend, Weekly Seasonality, and Daily Seasonality for DLTA.ZW 61

Figure 4.19 Prophet Model — Actual Versus Predicted Closing Prices for DLTA.ZW, August to November 2025 63

Figure 4.20 Prophet Residuals Over Time, August to November 2025 64

Figure 4.21 Prophet Residuals Distribution, August to November 2025 65

Figure 4.22 Comparison of RMSE Across Models 66

Figure 4.23 Comparison of MAE Across Models 66

Figure 4.24 Comparison of R^2 Across Models 67

Figure 4.25 Comparison of MAPE Across Models 68
Figure 4.26 Radar Chart — Comparative Performance of Predictive Models 69
Figure 4.27 Three-Tier System Architecture Diagram of the ZSE Insights Platform . 73
Figure 4.28 Sequential Data Ingestion and Preprocessing Pipeline Applied to Historical ZSE Stock Price Data..... 74
Figure 4.29 Feature Engineering Framework Illustrating the Three Distinct Input Preparation Pathways..... 75
Figure 4.30 Forecast Modelling Architecture Showing Independent Prediction Generation and Aggregation..... 76
Figure 4.31 Architecture of the Recursive Multi-Step Forecasting Engine (ZSE Insights Platform) 78

LIST OF EQUATIONS

Equation 1 Root Mean Square Error (RMSE) 24
Equation 2 Mean Absolute Error (MAE) 24
Equation 3 Mean Absolute Percentage Error (MAPE) 24
Equation 4 Supervised Learning Lag Feature Matrix 24
Equation 5 Recursive Multi-Step Forecasting Function 25

LIST OF TABLES

Table 4.1 Summary of Selected ZSE-Listed Companies 29

Table 4.2 Frequency-Based Summary Statistics of Daily Closing Prices for Selected ZSE-Listed Companies, May 2024 to November 2025 30

Table 4.3 Descriptive Statistics of Daily Closing Prices for Selected ZSE-Listed Companies, May 2024 to November 2025 31

Table 4.4 Mean Daily Return and Volatility Measures for Selected ZSE-Listed Companies, May 2024 to November 2025 34

Table 4.5 Missing Value Assessment Results for the Consolidated ZSE Dataset 36

Table 4.6 Z-Score Based Outlier Counts for Selected ZSE Counters, May 2024 to November 2025 39

Table 4.7 Pearson Correlation Coefficients Among Selected ZSE Stock Market Variables, Post-ZiG Period 41

Table 4.8 Variance Inflation Factor Results for Selected ZSE Stock Market Variables 43

Table 4.9 Rolling Volatility Summary Statistics for Proxy Stock DLTA.ZW, May 2024 to November 2025 46

Table 4.10 Feature Importance Scores from the Random Forest Model 49

Table 4.11 Random Forest Prediction Accuracy Metrics 50

Table 4.12 Feature Importance Scores from the XGBoost Model 53

Table 4.13 XGBoost Prediction Accuracy Metrics 54

Table 4.14 LSTM Prediction Accuracy Metrics 57

Table 4.15 Prophet Prediction Accuracy Metrics 62

Table 4.16 Comparative Performance of Predictive Models — DLTA.ZW, Post-ZiG Period 70

LIST OF APPENDICES

Appendix A: Sample Raw Data Structure 99

Appendix B: Model Hyperparameter Configurations 100

Appendix C: ZSE Insights Forecast Tab 102

Appendix D: ZSE Insights Model Comparison Tab 103

Appendix E: ZSE Insights Error Feature and Diagnostics Tab 104

Appendix F: ZSE Insights Volatility and Descriptive Statistics Tab 105

Appendix G: Data Preprocessing Pipeline 106

Appendix H: Feature Engineering 109

Appendix I: Model Training and Evaluation 111

Appendix J: Recursive Multi-Step Forecasting Engine 115

Appendix K: Model Comparison and Visualisation 117

CHAPTER ONE:
INTRODUCTION

1.0 Introduction

In emerging economies, stock markets face high volatility and limited predictive tools. On the Zimbabwe Stock Exchange (ZSE), low liquidity and economic instability make price forecasting particularly difficult, exposing investors to considerable decision-making risk. The rapid advancement of machine learning (ML) and deep learning (DL) techniques has opened new possibilities for improving forecast accuracy in such markets. However, the application of these techniques on the ZSE remains limited, and no interactive tool currently exists to make such forecasts accessible to retail investors and non-technical market participants. This study therefore asks: (1) Which ML model provides the highest forecasting accuracy for ZSE stock prices? (2) How can an interactive web platform make these forecasts accessible to retail investors? (3) What does comparative model analysis reveal about stock price behavior on the ZSE?

This chapter introduces the study on predictive analytics for stock price forecasting on the Zimbabwe Stock Exchange and the development of the ZSE Insights web application. It begins by presenting the background and research problems, followed by the study aim, objectives, and research questions. The chapter then outlines the significance, scope, assumptions, and limitations of the study. Key terms are defined, and the chapter concludes with an outline of the thesis structure.

1.1 Background of the study

Stock exchanges are a critical component of the modern financial system, facilitating capital mobilisation, price discovery, and the diversification of investment risk, thereby assisting firms in raising capital and investors in allocating risk efficiently. In emerging economies, stock exchanges play a significant role in directing the flow of savings towards productive investments, which supports broader economic growth (Mlambo & Biekpe, 2007). The Zimbabwe Stock Exchange (ZSE), established in 1896, is one of the oldest exchanges in Africa and serves as the primary platform for equity trading in Zimbabwe. The companies listed on the exchange span the key sectors of the economy, including banking, manufacturing, agriculture, and mining, and thus reflect the structure and priorities of the Zimbabwean economy (Murombo, 2023). The ZSE operates under the oversight of the Securities and

Exchange Commission of Zimbabwe (SECZ), which regulates market conduct and disclosure requirements to promote investor confidence and market integrity (Securities and Exchange Commission of Zimbabwe, 2021).

Despite its long history, the ZSE operates under challenging macroeconomic conditions that include inflationary pressures, currency instability, liquidity constraints, and periodic regulatory interventions. The ZSE All Share Index rose by approximately 500% in 2023; however, most of those nominal gains were eroded by ZWL depreciation, illustrating the difficulty of distinguishing genuine value growth from currency effects. Average daily turnover on the ZSE remained below USD 2 million in 2024, compared to over USD 300 million on the Johannesburg Stock Exchange, confirming the thin and illiquid nature of the market. The Reserve Bank of Zimbabwe (2023) has similarly noted that exchange rate instability and monetary policy shifts have contributed to erratic price movements on the ZSE, complicating the application of conventional forecasting models. Furthermore, the ZSE functioned as a hedge asset during the 2019 to 2020 hyperinflation period, causing structural breaks in price series that reduce the reliability of conventional forecasting methods (Bonga, 2019). These conditions increase the range of price fluctuations and make forecasting accuracy difficult to achieve.

Stock price forecasting is therefore critical for investors, portfolio managers, financial analysts, and policymakers, as it underpins portfolio allocation, risk management, asset valuation, and strategic decision-making (Fama, 1970). Forecasting on the ZSE has historically been dominated by econometric time-series models such as the Autoregressive Integrated Moving Average (ARIMA) and the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) family of models. Research on the ZSE indicates that GARCH-type models perform adequately in capturing volatility clustering, which is a characteristic feature of stock markets (Mutendadzamera & Mutasa, 2014; Bonga, 2019). However, these models rely on linearity and distributional assumptions that often fail to represent the nonlinear dynamics, regime shifts, and complex interactions that are characteristic of emerging financial markets.

Advances in computational finance have expanded the use of machine learning and deep learning methods for stock price forecasting. Ensemble models such as Random Forest and Extreme Gradient Boosting (XGBoost) are capable of modelling nonlinear relationships and complex feature interactions, while recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, are specifically designed to capture temporal dependencies in

sequential data (Breiman, 2001; Hochreiter & Schmidhuber, 1997). The Prophet model, a decomposition-based time series forecasting approach, provides a structured framework for capturing trend and seasonality components and has gained attention for its interpretability (Taylor & Letham, 2018).

Although these models have been widely applied in developed and large emerging markets, their systematic application within African stock markets, particularly within integrated and interactive analytical systems, remains limited. Moreover, existing studies on the ZSE focus predominantly on model accuracy and do not translate forecast results into tools accessible to non-technical users. This study addresses both gaps by developing ZSE Insights, an interactive web-based application that combines multiple ML and DL models to forecast stock prices on the ZSE and presents results through accessible visualisations.

1.2 Statement of the problem

Machine learning and deep learning techniques have demonstrated strong predictive performance across numerous financial markets globally (Jiang, 2021). However, their application on the Zimbabwe Stock Exchange remains limited. ZSE studies remain dominated by ARIMA and GARCH models, with limited comparative evaluation of ensemble and deep learning methods under Zimbabwe's specific market conditions, including currency volatility, low liquidity, and structural breaks (Bonga, 2019; Mutendadzamera & Mutasa, 2014). There is also limited evidence on how models such as Random Forest, XGBoost, LSTM, and Prophet compare in performance when applied to frequently traded ZSE counters such as CBZ Holdings, Econet Wireless Zimbabwe, and Delta Corporation, which represent some of the most active listings on the exchange.

The absence of comparative empirical studies means that the relative strengths and weaknesses of these modelling approaches under ZSE conditions are poorly understood. Chikore and Gachira (2014) observed that the ZSE exhibits weak-form inefficiency, suggesting that historical price data contains exploitable patterns, yet no study has systematically evaluated whether modern ML models can extract these patterns reliably. The Securities and Exchange Commission of Zimbabwe (2021) has acknowledged the need for improved market information systems to support investor decision-making, yet data-driven forecasting tools tailored to the ZSE remain largely absent. Investors and analysts operating in this market therefore continue to lack a rigorous evidence base for selecting appropriate forecasting approaches, increasing portfolio risk and reducing decision quality.

Furthermore, the majority of stock price prediction studies are presented in technical academic formats that individual investors and market analysts find difficult to apply (Murombo, 2023). The absence of an accessible, interactive forecasting tool for the ZSE limits the practical adoption of modern predictive techniques. Without such a tool, market participants are unable to leverage advances in machine learning for investment decision-making and risk assessment, and are likely to continue relying on intuition or limited information during periods of ZWL volatility (Reserve Bank of Zimbabwe, 2023).

This research addresses both gaps by developing ZSE Insights, a comparative and interactive stock price prediction system designed specifically for the Zimbabwe Stock Exchange, incorporating Random Forest, XGBoost, LSTM, and Prophet models within a user-accessible web-based platform.

1.3 Research objectives

The overall aim of the study is to design and evaluate an interactive web-based application that integrates multiple machine learning and deep learning models to forecast stock prices on the Zimbabwe Stock Exchange.

The specific objectives are to:

i. Preprocess historical daily stock price data for selected ZSE-listed companies to ensure suitability for machine learning modelling.

ii. Implement four predictive algorithms, namely Random Forest, XGBoost, LSTM, and Prophet, for stock price forecasting on the ZSE.

iii. Compare and evaluate the predictive performance of the selected models using appropriate statistical metrics.

iv. Develop an interactive web-based application for the visualisation of model forecasts and performance measures.

v. Demonstrate how the ZSE Insights application can support investment analysis and decision-making processes.

1.4 Research Questions

The following research questions guide the study:

i. Which of the selected predictive models, namely Random Forest, XGBoost, LSTM, and Prophet, provides the highest forecasting accuracy for Zimbabwe Stock Exchange stock prices?

ii. How does an interactive forecasting application contribute to the interpretation and usability of prediction results?

iii. How can the ZSE Insights application support investment analysis and decision-making processes?

1.5 Scope of the study

The study focuses on ten companies listed on the Zimbabwe Stock Exchange, namely CBZ Holdings, Delta Corporation, Econet Wireless Zimbabwe, FBC Holdings, First Mutual Life, Hippo Valley Estates, Meikles Limited, NMB Bank, Rainbow Tourism Group, and Steward Bank (TNCI), which rebranded to TN Cybertech during the study period. These companies were selected across key sectors of the ZSE, including banking, manufacturing, agriculture, and tourism, based on the availability of continuous daily trading data within the study period, following a purposive sampling approach consistent with similar studies on African frontier markets (Ntim et al., 2007; Chikore & Gachira, 2014).

The analysis utilises daily Open, High, Low, Close, and trading volume data from April 2024 to December 2025, covering the post-ZiG currency introduction period. The predictive framework is limited to four models: Random Forest, XGBoost, LSTM, and Prophet. To maintain a focused comparison of forecasting algorithms, the study relies primarily on historical price data and excludes macroeconomic indicators, news sentiment, social media sentiment, and order book information, as incorporating such variables would introduce additional complexity beyond the scope of this study (Bonga, 2019).

The ZSE Insights application provides stock price forecasts, visualisation dashboards, and model performance comparisons for the selected counters. It does not execute trades or provide automated investment recommendations.

1.6 Significance of the study

This study contributes to the growing body of literature on the application of Machine Learning (ML) and Deep Learning (DL) techniques in African stock markets, particularly within the

Zimbabwe Stock Exchange context. The findings may serve as reference for future studies on financial forecasting and predictive analytics.

For investors and financial analysts: The study provides evidence on the predictive performance of selected forecasting models, which may assist investors and analysts in improving investment decision making, portfolio management, and risk assessment.

For policymakers and regulators: The study enhances understanding of market dynamics and the capabilities of data-driven financial tools, thereby supporting the development of appropriate regulatory frameworks and policies for emerging financial technologies in the Zimbabwe Stock Exchange.

1.7 Assumptions of the study

i. Historical stock price data for the ten selected ZSE-listed companies, namely CBZ Holdings, Delta Corporation, Econet Wireless Zimbabwe, FBC Holdings, First Mutual Life, Hippo Valley Estates, Meikles Limited, NMB Bank, and Rainbow Tourism Group, together with Steward Bank (TNCI), which rebranded to TN Cybertech during the study period, is accurate, complete, and representative of actual market activity between April 2024 and December 2025.

ii. The post-ZiG currency environment, introduced in April 2024, represents a sufficiently stable monetary framework to allow the selected models to learn meaningful relationships from historical price data, notwithstanding residual macroeconomic volatility (Reserve Bank of Zimbabwe, 2023).

iii. Historical price movements within the study period contain exploitable patterns and information that can be utilised to forecast future stock price behaviour, consistent with the weak-form inefficiency observed on the ZSE (Chikore & Gachira, 2014).

iv. Random Forest, XGBoost, LSTM, and Prophet models are appropriate for capturing the nonlinear, temporal, and dynamic characteristics of stock price data within the ZiG-denominated trading environment (Breiman, 2001; Hochreiter & Schmidhuber, 1997).

1.8 Limitations of the study

i. The study is limited to ten companies listed on the Zimbabwe Stock Exchange, namely CBZ Holdings, Delta Corporation, Econet Wireless Zimbabwe, FBC Holdings, First Mutual Life, Hippo Valley Estates, Meikles Limited, NMB Bank, Rainbow Tourism Group, and Steward Bank (TNCI), selected on the basis of data availability and trading consistency

within the April 2024 to December 2025 study period. Findings may therefore not be generalisable to all ZSE-listed counters (Murombo, 2023).

ii. The study period spans approximately twenty-one months from April 2024 to December 2025, which, while covering the post-ZiG currency transition, represents a relatively short window for training and evaluating machine learning models. Longer historical datasets would improve model generalisation but are not available in a consistent currency denomination for the ZSE (Reserve Bank of Zimbabwe, 2023).

iii. The forecasting models rely primarily on historical stock price data and do not incorporate macroeconomic conditions, political events, or investor sentiment, all of which may influence stock prices. The Reserve Bank of Zimbabwe (2023) has noted that monetary policy shifts and exchange rate interventions have had material effects on ZSE price movements, factors which fall outside the scope of this study.

iv. The rebrand of Steward Bank to TN Cybertech during the study period may introduce inconsistencies in the trading data for that counter, which could affect the reliability of forecasts generated for that specific company.

v. Limited computational resources may restrict the extent of hyperparameter tuning and model optimisation, a constraint also noted in similar studies conducted in resource-limited research environments on the African continent (Msosa, 2022).

These limitations may affect the predictive accuracy of the models and the generalisability of the findings beyond the scope of the study.

1.9 Definition of key terms

i. Stock Price Forecasting: The process of estimating future stock prices using historical price data, analytical models, and computational methods. Stock price forecasting informs investment decisions, portfolio management, and risk assessment (Jiang, 2021).

ii. Machine Learning (ML): A branch of artificial intelligence that enables computers to learn patterns from data and make predictions without explicit programming. ML methods are applied in this study to capture complex relationships in historical stock price data (Breiman, 2001).

iii. Deep Learning (DL): A subset of machine learning that uses multi-layered artificial neural networks to learn complex patterns from large datasets. Deep learning models, particularly

LSTM networks, are well suited to sequential and time-dependent financial data (Hochreiter & Schmidhuber, 1997).

iv. Random Forest: An ensemble machine learning algorithm that constructs multiple decision trees during training and combines their outputs to produce more accurate and stable predictions. It is effective for modelling nonlinear relationships in financial data (Breiman, 2001).

v. XGBoost (Extreme Gradient Boosting): A gradient boosting algorithm that builds an ensemble of decision trees sequentially to improve predictive performance. XGBoost is widely used in financial forecasting due to its high accuracy and computational efficiency (Chen & Guestrin, 2016).

vi. Long Short-Term Memory (LSTM): A type of recurrent neural network architecture designed to learn long-range dependencies in sequential data. LSTM networks are applied in this study to capture temporal patterns in stock price series (Hochreiter & Schmidhuber, 1997).

vii. Prophet Model: A decomposition-based time series forecasting model that separates data into trend, seasonality, and residual components. It is noted for its interpretability and robustness to missing data and sudden changes in time series (Taylor & Letham, 2018).

viii. OHLC Data: Stock price data comprising Open, High, Low, and Close prices recorded for each trading day. OHLC data provides a detailed view of intraday price movements and serves as the primary input for the forecasting models in this study.

ix. Volatility: The degree of variation in the price of a financial asset over time, commonly used as a measure of market risk. High volatility on the ZSE reflects the influence of macroeconomic instability and currency fluctuations on stock prices (Bonga, 2019).

x. Predictive Modelling: The application of statistical, machine learning, or deep learning techniques to estimate future outcomes based on historical data. Predictive modelling is the central methodological approach of this study (Jiang, 2021).

xi. Efficient Market Hypothesis (EMH): A financial theory which holds that asset prices fully reflect all available information at any given time, making it difficult to consistently outperform the market through prediction (Fama, 1970). Studies on the ZSE have found evidence of weak-form inefficiency, which provides a theoretical basis for the forecasting approach adopted in this study (Chikore & Gachira, 2014).

1.10 Chapter conclusion

This chapter provided the contextual and conceptual foundation of the study, outlining the background, problem statement, objectives, research questions, scope, significance, assumptions, and limitations. The next chapter reviews the theoretical and empirical literature relevant to the study.

CHAPTER TWO
LITERATURE REVIEW

2.0 Introduction

Stock price forecasting has attracted sustained attention in financial research owing to its significance in investment decision-making, portfolio allocation, and risk management. In emerging and frontier markets, accurate forecasting becomes particularly critical due to high volatility, market inefficiencies, and limited information transparency (Mlambo & Biekpe, 2007). The Zimbabwe Stock Exchange operates within a complex macroeconomic environment characterised by liquidity constraints, currency instability, and frequent policy changes, all of which introduce structural breaks and irregular price movements that make forecasting both challenging and necessary (Bonga, 2019; Murombo, 2023).

This chapter presents the theoretical foundations and empirical evidence that underpin stock price prediction and financial market behaviour. It draws on key theories from financial economics, econometrics, machine learning, and deep learning to explain how stock prices are generated and how they can be modelled. The chapter then reviews empirical studies on stock price forecasting with emphasis on Zimbabwean, African regional, and international evidence. It concludes by identifying research gaps and presenting the conceptual framework of the study.

2.1 Theoretical literature

This section presents the theoretical foundations that underpin stock price prediction and financial market behaviour. It draws on key theories from financial economics, econometrics, machine learning, and deep learning to explain how stock prices are generated and how they can be modelled. The discussion begins with the Efficient Market Hypothesis, proceeds through time series and volatility modelling theory, and concludes with the theoretical basis for machine learning and deep learning approaches.

2.1.1 Efficient market hypothesis

The Efficient Market Hypothesis (EMH), proposed by Fama (1970), holds that financial markets incorporate all available information into asset prices at any given time. Under the weak form of the EMH, current prices already reflect all historical price information, implying

that past prices cannot be used to consistently predict future prices. If markets were perfectly efficient, the theoretical basis for systematic stock price forecasting would be severely limited.

However, evidence from emerging and frontier markets has consistently challenged this assumption. Mlambo and Biekpe (2007) demonstrated that African stock markets frequently exhibit weak-form inefficiencies attributable to thin trading, limited investor participation, delayed information dissemination, and institutional constraints. Chikore and Gachira (2014) found similar evidence of weak-form inefficiency on the ZSE, suggesting that historical price data contains exploitable patterns. Murombo (2023) further identified calendar anomalies on the ZSE consistent with predictable return patterns. These findings collectively provide a theoretical justification for applying predictive models to ZSE stock price data, as the market does not appear to fully and immediately incorporate available information into prices.

2.1.2 Time series modelling theory

Classical time series forecasting theory provides the foundation for the earliest stock price prediction models. The Box-Jenkins methodology underpins Autoregressive Integrated Moving Average (ARIMA) models, which represent future values of a time series as linear combinations of past observations and error terms. The core assumptions of linearity and stationarity make ARIMA models tractable and interpretable. However, their theoretical limitation lies in the inability to capture nonlinear relationships and sudden structural changes that are common in financial markets, particularly in volatile emerging economies such as Zimbabwe (Bonga, 2019).

2.1.3 Volatility modelling theory

The Generalised Autoregressive Conditional Heteroskedasticity (GARCH) family of models, introduced by Bollerslev (1986), is grounded in volatility modelling theory. GARCH models recognise that the variance of financial returns is time-varying and tends to cluster during periods of market stress. Extensions such as EGARCH and TGARCH account for asymmetric responses to positive and negative shocks, making them well suited to real-world stock market dynamics. Research on the ZSE has confirmed the presence of volatility clustering and asymmetric effects, validating the application of GARCH-type models in the Zimbabwean context (Mutendadzamera & Mutasa, 2014; Bonga, 2019). However, while GARCH models perform well in volatility forecasting, they are theoretically limited in their ability to forecast stock price levels directly.

2.1.4 Machine learning theory

Machine learning theory offers an alternative modelling paradigm that challenges many assumptions of classical econometrics. Supervised learning models identify complex input-output relationships from data without imposing pre-specified functional forms. Ensemble learning theory, which underlies both Random Forest and XGBoost, is based on the principle that combining multiple weak learners produces superior predictive accuracy and generalisation performance (Breiman, 2001; Chen & Guestrin, 2016). These models are theoretically attractive for financial forecasting because they can represent nonlinear relationships and high-order interactions among variables without requiring distributional assumptions.

2.1.5 Deep learning theory

Deep learning theory extends machine learning by employing multi-layered neural networks to learn hierarchical representations from data. Sequence modelling theory underpins Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which are designed to retain information across long sequences (Hochreiter & Schmidhuber, 1997). This theoretical capacity for temporal memory makes LSTM networks particularly well suited to modelling stock price time series, where current prices are influenced by patterns established over preceding periods. The Gated Recurrent Unit (GRU), a simplified variant of LSTM, offers comparable sequential modelling capabilities with reduced computational cost (Cho et al., 2014).

2.1.6 Additive decomposition theory

The Prophet model is grounded in additive time series decomposition theory, which treats an observed series as the sum of trend, seasonal, and irregular components (Taylor & Letham, 2018). This theoretical structure makes Prophet robust to missing data, outliers, and structural breaks, which are characteristic features of ZSE trading data. Its interpretability also makes it a useful complementary tool in comparative forecasting research where model transparency is valued alongside predictive accuracy.

2.2 Empirical literature

Empirical research on stock price forecasting has undergone a substantial transformation over the past decade, shifting from linear econometric methods towards machine learning and deep

learning approaches. This section reviews between eight and fifteen empirical studies, structured from international evidence through to regional African and local Zimbabwean studies, to situate the present research within the existing body of knowledge.

Fischer and Krauss (2018) conducted a large-scale empirical evaluation of LSTM networks for forecasting daily returns of S&P 500 constituent stocks over the period 1992 to 2015. Using a rolling window methodology, the study found that LSTM networks significantly outperformed memory-free classifiers, including random forests, deep neural networks, and logistic regression. The study attributed the superior performance of LSTM to its capacity to capture temporal dependencies in sequential financial data. This finding is directly relevant to the present study as it provides empirical justification for including LSTM among the candidate models for ZSE stock price forecasting.

Chen and Guestrin (2016) introduced XGBoost and demonstrated its superior predictive accuracy and computational efficiency across multiple benchmark datasets, including financial prediction tasks. The algorithm's ability to handle missing data, regularise model complexity, and correct sequential prediction errors through gradient boosting has made it one of the most widely adopted models in financial forecasting competitions and applied research. The present study applies XGBoost to ZSE data to assess whether its documented advantages extend to a thin frontier market operating under post-ZiG monetary conditions.

Jiang (2021) conducted a comprehensive review of deep learning applications in stock market prediction, covering studies published between 2014 and 2020 across multiple international markets. The review found that LSTM and its variants consistently outperformed traditional time series models in capturing nonlinear and temporal patterns in stock price data. However, Jiang (2021) also noted that deep learning models require larger datasets and greater computational resources, and that their performance advantage diminishes in thin markets with limited trading history. This observation is particularly pertinent to the ZSE context, where the study period spans approximately twenty-one months.

Wagh and Gaikwad (2025) evaluated the comparative performance of ensemble learning models including Random Forest and XGBoost against ARIMA and LSTM for stock price forecasting on an emerging market dataset. Their findings confirmed that ensemble models outperformed individual models in most forecasting horizons, particularly for short-term daily predictions. The study also highlighted that XGBoost demonstrated the best balance between

accuracy and computational efficiency. These results provide a direct empirical benchmark against which the findings of the present study can be compared.

Adewumi, Alade, and Akinwale (2021) applied multiple machine learning techniques, including neural networks, support vector machines, and random forests, to Nigerian Stock Exchange data. The study found that neural network and random forest models outperformed ARIMA across all accuracy metrics, including RMSE and MAE. Critically, the authors noted that the nonlinear and structurally irregular nature of African stock market data made machine learning approaches more appropriate than classical econometric models. This regional evidence supports the adoption of ML models in the present ZSE study, given the structural similarities between the Nigerian and Zimbabwean market environments.

Msosa (2022) examined the application of machine learning in financial forecasting across sub-Saharan African markets, identifying a persistent gap in the comparative evaluation of ensemble and deep learning models in frontier markets. The study found that while individual ML models had been applied in isolation across several African exchanges, systematic comparative studies incorporating multiple model types remained scarce. The present research directly addresses this gap in the Zimbabwean context by evaluating four models simultaneously using consistent data and evaluation metrics.

Ntim, Opong, and Danbolt (2007) conducted an empirical re-examination of weak-form market efficiency on the Ghana Stock Exchange using variance ratio tests. The study found strong evidence against the random walk hypothesis, confirming weak-form inefficiency consistent with Mlambo and Biekpe (2007). Although focused on Ghana, the findings are regionally relevant as they confirm that historical price data on African frontier exchanges contains predictable patterns that justify the application of forecasting models.

Mutendadzamera and Mutasa (2014) analysed stock market volatility on the ZSE using ARIMA and GARCH models. The study found that GARCH models captured volatility clustering more effectively than ARIMA, confirming the presence of time-varying conditional variance in ZSE returns. However, the study was limited to volatility modelling and did not extend to stock price level forecasting. This limitation highlights the gap that the present research seeks to address by moving beyond volatility estimation towards direct price prediction using machine learning models.

Bonga (2019) extended the Zimbabwean volatility literature by applying EGARCH models to ZSE data and finding evidence of asymmetric volatility effects, whereby negative market shocks generated greater volatility increases than equivalent positive shocks. The study also documented the presence of structural breaks in ZSE price series attributable to macroeconomic instability and policy changes. These findings inform the data preprocessing decisions of the present study, particularly the need to account for structural breaks when preparing ZSE data for machine learning model training.

Murombo (2023) investigated calendar anomalies on the ZSE and identified statistically significant return patterns associated with specific days of the week and months of the year. The existence of these anomalies is inconsistent with weak-form efficiency and supports the argument that historical price data on the ZSE contains exploitable regularities. This study provides local empirical evidence directly supporting the forecasting approach adopted in the present research and suggests that machine learning models capable of detecting complex temporal patterns may perform well on ZSE data.

Chikore and Gachira (2014) tested the weak-form efficient market hypothesis on the ZSE using runs tests and autocorrelation analysis. The study found significant serial correlation in ZSE stock returns, confirming weak-form inefficiency and indicating that past price movements contain information useful for prediction. The authors attributed this inefficiency to thin trading, information asymmetry, and limited institutional participation on the exchange. These findings are foundational to the present study as they provide the empirical justification for applying machine learning models to historical ZSE price data.

Taylor and Letham (2018) introduced the Prophet forecasting system and demonstrated its effectiveness for medium-term business and economic forecasting tasks characterised by strong trend and seasonal components. The model was shown to be robust to missing data and outliers, and to require minimal manual tuning, making it accessible to non-specialist users. Although applications of Prophet in African stock markets remain limited, its interpretability and robustness make it a valuable inclusion in the comparative analysis of the present study, particularly given the irregular trading patterns characteristic of the ZSE.

2.3 Literature synthesis and research gap

The review of theoretical and empirical literature reveals several important insights. Traditional econometric models such as ARIMA and GARCH remain useful for capturing linear dynamics

and volatility patterns but are theoretically and empirically limited in their ability to represent nonlinear relationships, structural breaks, and complex temporal dependencies (Bonga, 2019; Mutendadzamera & Mutasa, 2014). Machine learning and deep learning models have demonstrated superior predictive performance across multiple international and regional markets, with ensemble models such as XGBoost and Random Forest performing particularly well in short-term forecasting tasks (Fischer & Krauss, 2018; Wagh & Gaikwad, 2025; Adewumi et al., 2021).

At the local level, existing ZSE research remains concentrated on volatility modelling and market efficiency testing rather than stock price level forecasting (Murombo, 2023; Chikore & Gachira, 2014). No study identified in this review has systematically compared the performance of Random Forest, XGBoost, LSTM, and Prophet on ZSE data within a single consistent analytical framework. Furthermore, the literature reveals a complete absence of interactive, user-facing forecasting tools designed specifically for the ZSE, limiting the practical translation of academic forecasting research into tools accessible to retail investors and analysts (Securities and Exchange Commission of Zimbabwe, 2021).

This study addresses these gaps by implementing and comparing four machine learning and deep learning models on a dataset of ten ZSE-listed companies over the post-ZiG period from April 2024 to December 2025, and by developing the ZSE Insights web application to make forecasting results accessible to non-technical users.

2.4 Conceptual framework

The conceptual framework of this study illustrates the process through which historical stock market data is transformed into forecasted stock prices using machine learning and deep learning models, and subsequently presented to users through an interactive web application. The framework is organised around a sequential pipeline that begins with data collection and ends with decision support, with each stage informing the next.

The first stage involves data collection and preprocessing. Historical daily stock price data, comprising Open, High, Low, Close, and volume figures, is sourced for the ten selected ZSE-listed companies covering the period April 2024 to December 2025. At this stage, missing values are addressed, outliers are managed, and the data is transformed into a format suitable for model training. Feature engineering is also conducted at this stage to derive additional input variables from the raw price series.

The second stage involves model implementation and training. The preprocessed data is used to train four predictive models, namely Random Forest, XGBoost, LSTM, and Prophet. Each model is trained on a defined portion of the historical data and evaluated on a held-out test set. The models represent two distinct paradigms: ensemble tree-based methods, which excel at capturing nonlinear feature interactions, and sequential deep learning methods, which are designed to capture temporal dependencies in time series data.

The third stage involves model evaluation and comparison. The predictive performance of each model is assessed using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) as primary accuracy metrics, supplemented by Mean Absolute Error (MAE). This comparative evaluation forms the empirical core of the study and directly addresses the first and third research questions.

The fourth stage involves interactive visualisation and deployment through the ZSE Insights web application. Forecast outputs and model performance metrics are presented to users through an accessible, browser-based interface built using Streamlit. The application allows users to run forecasts, compare model accuracy in real time, and explore historical price trends for the selected ZSE counters, directly addressing the second research question regarding usability and interpretability.

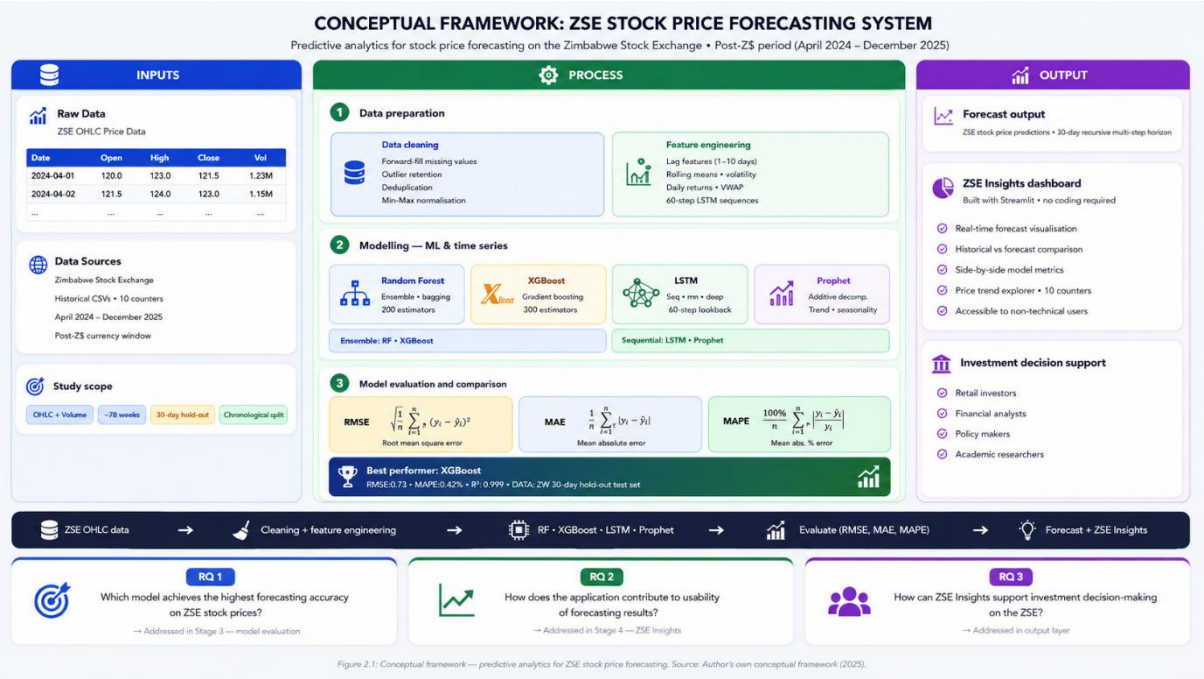


Fig 2.1: Conceptual framework of the ZSE Insights stock price forecasting system

Overall, the framework demonstrates how historical financial data from the ZSE is transformed into actionable predictions through a structured machine learning pipeline, and how those predictions are made accessible to investors and analysts through the ZSE Insights application. The relationship between the four stages is illustrated in figure 2.1 above.

2.5 Chapter conclusion

This chapter reviewed the theoretical and empirical literature relevant to stock price forecasting with particular emphasis on the Zimbabwe Stock Exchange context. The theoretical review established that weak-form inefficiency on the ZSE provides justification for predictive modelling, while machine learning and deep learning theories offer a stronger basis for capturing the nonlinear and temporal characteristics of ZSE price data than classical econometric approaches. The empirical review identified a clear gap in comparative machine learning research on the ZSE and confirmed the absence of interactive forecasting tools designed for non-technical users in the Zimbabwean market. The conceptual framework presented in this chapter provides the structural basis for the research methodology described in the following chapter.

CHAPTER THREE:
RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research methodology adopted in the development of ZSE Insights, an interactive web-based application for stock price forecasting on the Zimbabwe Stock Exchange. The chapter outlines the philosophical stance, research design, data sources, sampling techniques, variable specification, analytical models, diagnostic procedures, validation methods, and ethical considerations that guided the study. A well-structured methodology is essential for producing trustworthy, valid, and replicable results, particularly when quantitative research involves financial time-series data in volatile emerging market contexts (Creswell, 2014; Saunders, Lewis, & Thornhill, 2019).

The research employs quantitative methods and machine learning techniques to examine historical stock price records for ten selected ZSE-listed companies over the period April 2024 to December 2025. This chapter concludes by setting the foundation for the results and discussion presented in Chapter Four.

3.2 Research philosophy and design

The study adopts a positivist research philosophy, which prioritises objectivity, quantification, and the empirical testing of relationships through statistical data. Positivism holds that reality consists of observable and measurable facts that exist independently of the researcher, making it well suited to studies that rely on numerical data and statistical analysis (Saunders et al., 2019). This philosophical stance is widely adopted in financial economics and computational finance, where conclusions are derived from data-driven evidence rather than subjective interpretation (Gujarati & Porter, 2009). Its application to stock price forecasting research on African markets has been documented in studies such as Adewumi et al. (2021) and Msosa (2022), both of which adopted positivist quantitative frameworks when applying machine learning to financial time-series data.

Consistent with the positivist philosophy, the study employs a quantitative research design. Quantitative design focuses on the analysis of numerical data to identify patterns, relationships, and predictive models. It is appropriate for stock price forecasting because the research

involves examining historical price trends through mathematical and computational methods (Gujarati & Porter, 2009). The study further adopts both a predictive and explanatory research orientation. The predictive dimension involves estimating future stock prices using machine learning and deep learning models, while the explanatory dimension involves comparing model performance and accounting for observed differences in forecasting accuracy. This combined orientation enables the study to not only generate forecasts but also to evaluate the relative suitability of different modelling approaches for the ZSE context (Creswell, 2014).

3.3 Data sources and collection methods

The study relies exclusively on secondary data comprising historical daily stock price records obtained from publicly available Zimbabwe Stock Exchange records and reputable financial data repositories. Secondary data is appropriate for this study because stock prices are objectively recorded, verifiable, and widely used in financial forecasting research (Wooldridge, 2016). The use of secondary financial data is consistent with established practice in ZSE empirical studies, including Mutendadzamera and Mutasa (2014), Bonga (2019), and Murombo (2023).

The dataset covers ten ZSE-listed companies, namely CBZ Holdings, Delta Corporation, Econet Wireless Zimbabwe, FBC Holdings, First Mutual Life, Hippo Valley Estates, Meikles Limited, NMB Bank, Rainbow Tourism Group, and Steward Bank (TNCI), which rebranded to TN Cybertech during the study period. Daily observations span from April 2024 to December 2025, covering the post-ZiG currency transition period. The variables collected for each company include:

i. Opening price
ii. Highest price
iii. Lowest price
iv. Closing price
v. Trading volume

These variables are collectively referred to as OHLC data and represent the standard inputs for stock market analysis, as they capture the full range of price movement within each trading day (Adewumi et al., 2021). Historical price data was downloaded in spreadsheet format and

subsequently converted into a machine-readable format for analysis using Python-based data science libraries, including Pandas, NumPy, and Scikit-learn.

Prior to model training, the data underwent several preprocessing steps including removal of duplicate entries, treatment of missing values, date formatting, and feature engineering. For scale-sensitive models such as LSTM, normalisation was applied using Min-Max scaling to improve convergence and model stability. Data preprocessing is a critical step in machine learning workflows, as poor data quality directly degrades model performance (Hastie, Tibshirani, & Friedman, 2017). The integrity of ZSE data presents a known challenge, as gaps and inconsistencies attributable to thin trading and suspension periods have been documented in the local literature (Chikore & Gachira, 2014; Murombo, 2023).

3.4 Target population and sampling

The target population of this study comprises all companies listed and actively trading on the Zimbabwe Stock Exchange. The ZSE lists companies across multiple sectors including banking, manufacturing, mining, agriculture, and telecommunications, collectively reflecting the structure of the broader Zimbabwean economy (Murombo, 2023; Securities and Exchange Commission of Zimbabwe, 2021).

Given constraints related to data availability, trading consistency, and the computational requirements of machine learning model training, purposive sampling was employed to select a subset of ZSE-listed companies. Purposive sampling involves the deliberate selection of units that meet specific research criteria, and is well established in financial forecasting studies where data quality and continuity are paramount (Saunders et al., 2019; Ntim, Opong, & Danbolt, 2007). The criteria applied for company selection were:

i. Availability of continuous daily trading data from April 2024 to December 2025 ii. Sufficient trading frequency to avoid excessive gaps in the price series iii. Representation across key economic sectors of the ZSE.

Based on these criteria, ten companies were selected: CBZ Holdings (banking), Delta Corporation (manufacturing), Econet Wireless Zimbabwe (telecommunications), FBC Holdings (financial services), First Mutual Life (insurance), Hippo Valley Estates (agriculture), Meikles Limited (retail and hospitality), NMB Bank (banking), Rainbow Tourism Group

(tourism), and Steward Bank / TN Cybertech (financial technology). Although purposive sampling limits the generalisability of findings to the broader exchange, it is well suited to predictive modelling research that demands datasets of high quality and temporal continuity (Msosa, 2022).

3.5 Description of variables and expected relationships

The study draws on both dependent and independent variables derived from historical stock price data.

The dependent variable is the future closing price of each selected ZSE-listed stock. Closing prices are standard in financial forecasting research as they represent the final market consensus at the end of each trading day and serve as the basis for return calculations and asset valuation (Gujarati & Porter, 2009). The use of closing price as the primary forecast target is consistent with comparable studies on African stock markets (Adewumi et al., 2021; Msosa, 2022).

The independent variables comprise historical price components and their lagged values, specifically:

i. Opening price
ii. Daily high price
iii. Daily low price
iv. Previous closing prices (lagged values)
v. Trading volume

Lagged variables are expected to carry predictive information about future price movements due to the phenomena of price persistence and momentum in financial markets (Fama, 1970). In a market exhibiting weak-form inefficiency such as the ZSE, historical price data is more likely to contain exploitable patterns than in a fully efficient market (Chikore & Gachira, 2014; Mlambo & Biekpe, 2007). The relationships between the independent and dependent variables are expected to be nonlinear and dynamic, varying in direction and magnitude across time and across individual stocks. Machine learning and deep learning models are employed precisely because they are capable of identifying and representing these complex relationships without imposing restrictive parametric assumptions (Breiman, 2001; Hochreiter & Schmidhuber, 1997).

3.6 Analytical models and justification

Four predictive models are implemented in this study: Random Forest, XGBoost, Long Short-Term Memory (LSTM), and Prophet. The selection of these models is justified by their documented effectiveness in financial forecasting and their complementary strengths across different aspects of stock price prediction.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and aggregates their outputs to produce a final prediction. Its primary advantage lies in its ability to model nonlinear relationships while controlling overfitting through randomisation and averaging (Breiman, 2001). Random Forest is particularly appropriate when the relationships between predictors and the target variable are complex and non-parametric. Its application to stock price forecasting in African emerging markets has yielded competitive results in recent studies (Adewumi et al., 2021).

XGBoost is a gradient boosting algorithm that builds an ensemble of decision trees sequentially, with each successive tree correcting the residual errors of its predecessor. Its efficiency, regularisation capabilities, and strong predictive performance make it one of the most widely applied algorithms in financial time-series forecasting (Chen & Guestrin, 2016). Wagh and Gaikwad (2025) demonstrated that XGBoost consistently outperformed individual models and matched or exceeded LSTM in short-term forecasting tasks on emerging market data.

LSTM is a deep learning model specifically designed for sequential data. Its memory cells and gating mechanisms enable it to retain relevant information across long time horizons, making it well suited to stock price forecasting where current prices depend on patterns established over preceding periods (Hochreiter & Schmidhuber, 1997). Fischer and Krauss (2018) demonstrated that LSTM networks significantly outperformed memory-free classifiers on large financial datasets, providing empirical justification for its inclusion in this study.

Prophet is a decomposition-based time series model that expresses an observed series as the additive combination of trend, seasonality, and residual components (Taylor & Letham, 2018). Its interpretability, robustness to missing data, and ability to handle structural breaks make it a suitable baseline and supplementary forecasting model, particularly given the irregular trading patterns and data gaps characteristic of the ZSE (Bonga, 2019).

The predictive performance of all four models is evaluated using three error metrics. Root Mean Square Error (RMSE) penalises larger errors more heavily and is expressed as:

$$\boldsymbol{RMSE} = \sqrt{(\mathbf{1}/\boldsymbol{n} \times \boldsymbol{\Sigma}(\boldsymbol{y}_i - \hat{Y}_i)^2} \qquad \textbf{(3.1)}$$

Mean Absolute Error (MAE) measures the average magnitude of forecast errors without penalising outliers disproportionately, and is expressed as:

$$\boldsymbol{MAE} = \mathbf{1}/\boldsymbol{n} \times \boldsymbol{\Sigma}|\boldsymbol{y}_i - \hat{Y}_i| \qquad \textbf{(3.2)}$$

Mean Absolute Percentage Error (MAPE) expresses forecast error as a percentage of actual values, facilitating comparison across stocks trading at different price levels, and is expressed as:

$$\boldsymbol{MAPE} = \mathbf{100}/\boldsymbol{n} \times \boldsymbol{\Sigma}|(\boldsymbol{y}_i - \hat{Y}_i)/\boldsymbol{y}_i| \qquad \textbf{(3.3)}$$

Together, these three metrics provide a comprehensive and balanced assessment of model accuracy and are consistent with evaluation standards adopted in comparable financial forecasting studies (Jiang, 2021; Fischer & Krauss, 2018).

3.7 Recursive Multi-Step Forecasting Engine

A defining feature of the ZSE Insights platform is its Recursive Multi-Step Forecasting Engine, which enables the platform to generate price trajectories extending across multiple future trading periods rather than simply producing a single one-step-ahead estimate.

The challenge with multi-step forecasting is that the explanatory variables required to generate predictions beyond t+1 are themselves unknown at the time of forecasting. A common but problematic response to this challenge is to assume that future features are known — an assumption that rarely holds in practice. The recursive approach taken here sidesteps this assumption by treating each predicted value as an observed input for the subsequent prediction. Once $\hat{p}(t+1)$ has been generated, it is inserted into the input window in place of the most recent lagged observation, and the model is re-applied to produce $\hat{p}(t+2)$. This substitution continues iteratively until the user-specified horizon is reached.

The recursive process can be expressed as:

$$P(t+k) = f(P(t+k-1), P(t+k-2), \dots) \qquad \textbf{(3.4)}$$

where f(·) denotes the forecasting function and k indexes the step within the forecast horizon. It should be noted that recursive forecasting compounds prediction error across steps, so accuracy naturally diminishes as the horizon extends. This limitation is acknowledged in the evaluation discussion in Section 4.10. Nevertheless, for the short-to-medium horizons considered in this study, the recursive engine produces meaningful and practically interpretable price trajectories. Figure 3.1 presents the engine architecture, including the feedback loops central to the recursive mechanism.

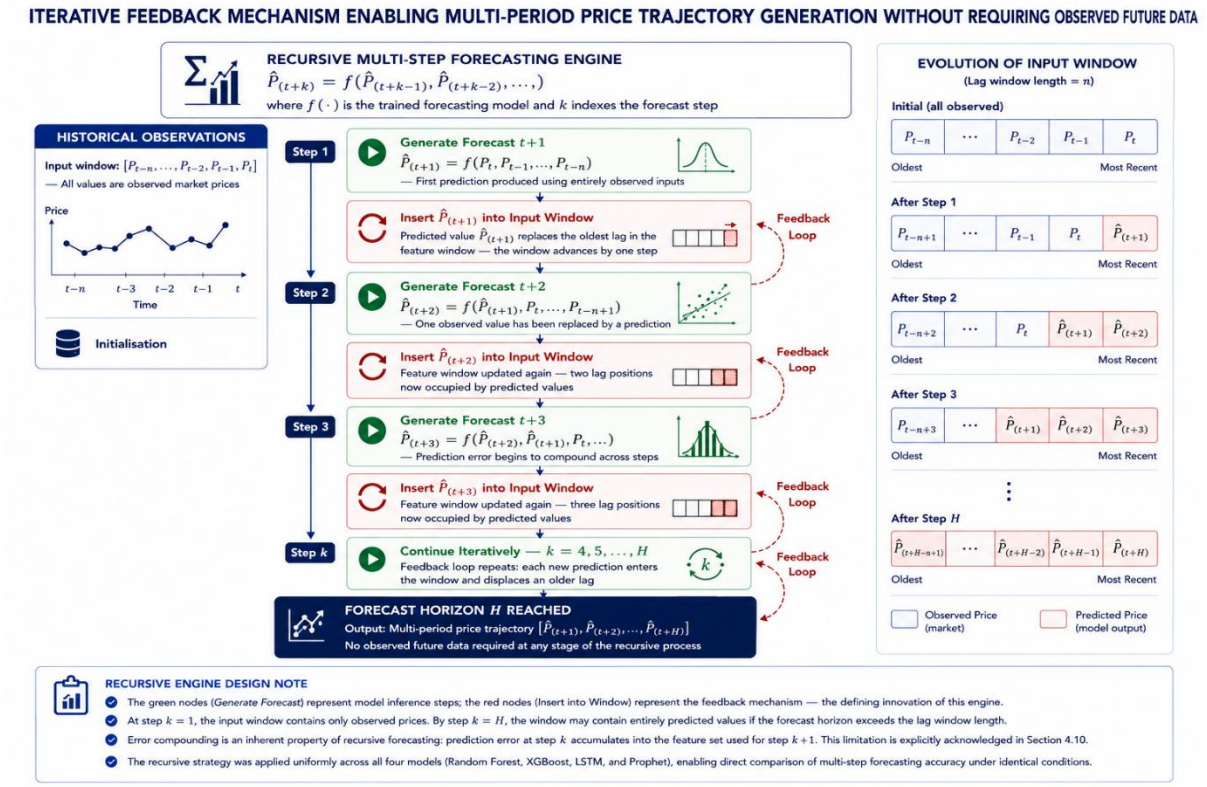


Figure 3.1: Architecture of the Recursive Multi-Step Forecasting Engine, illustrating the iterative feedback mechanism through which each predicted value P̂(t+k) is reinserted as a model input to generate P̂(t+k+1) across the full forecast horizon.

3.8 Model pre-tests and diagnostic checks

Prior to model estimation, several preliminary tests and diagnostic checks were conducted to confirm data suitability and model reliability.

Missing value analysis was performed to identify and address gaps in the daily price series for each of the ten selected companies. Missing observations attributable to non-trading days, public holidays, and stock suspension periods were handled through forward-fill imputation, which is consistent with standard practice in financial time-series preprocessing (Hastie et al., 2017). Outlier detection was conducted using interquartile range analysis, and extreme values

were investigated to determine whether they reflected genuine market events or data recording errors. Confirmed data errors were removed while genuine price spikes attributable to documented market events were retained.

For scale-sensitive models, particularly LSTM, Min-Max normalisation was applied to transform all input features to the range [0, 1], improving gradient convergence during neural network training. Stationarity diagnostics were conducted for benchmarking purposes, primarily to inform the configuration of the Prophet model and to characterise the properties of the ZSE price series. While machine learning models do not strictly require stationarity, understanding the degree of non-stationarity in the data assists in interpreting model behaviour and selecting appropriate lag structures (Wooldridge, 2016). The presence of structural breaks in ZSE data, attributable to the April 2024 ZiG currency introduction, was noted and accounted for in the feature engineering stage (Reserve Bank of Zimbabwe, 2023; Bonga, 2019).

Following model training, residual analysis was conducted to assess the distribution and pattern of prediction errors for each model. Systematic patterns in residuals may indicate model misspecification or the presence of unaccounted temporal dependencies, while the absence of such patterns supports the adequacy of the fitted model (Wooldridge, 2016).

3.9 Model fit and validation

Model validation was conducted using an out-of-sample testing approach, which involved partitioning the dataset into a training set and a held-out test set. This approach assesses how well each model generalises to unseen data, which is the most appropriate validation strategy for forecasting applications (Hastie et al., 2017). The dataset was split such that the earlier portion of the April 2024 to December 2025 period was used for training and the latter portion was reserved for testing, following a temporal split consistent with time-series forecasting conventions.

Model fit was evaluated by comparing RMSE, MAE, and MAPE across all four models on the test set. Models recording lower values across these metrics were considered to demonstrate superior predictive ability. In addition to quantitative assessment, actual versus predicted price plots were generated to provide a visual assessment of each model's ability to track price movements across the test period. These visualisations are presented through the ZSE Insights application and are discussed further in Chapter Four.

The comparative evaluation framework adopted in this study is consistent with similar multi-model studies conducted on African and emerging market financial data (Adewumi et al., 2021; Wagh & Gaikwad, 2025; Msosa, 2022). By assessing all four models under identical data conditions and evaluation metrics, the study ensures that performance differences are attributable to model characteristics rather than methodological inconsistencies.

3.10 Ethical considerations

The research conforms to the ethical standards governing academic research. The study relies exclusively on publicly accessible secondary data and does not involve any personal, confidential, or commercially sensitive information. No individuals or organisations are exposed to risk as a result of this research.

All data sources and academic works are cited using the APA referencing format in accordance with the requirements of Bindura University of Science Education. The research was conducted impartially and the findings are reported accurately without manipulation or misrepresentation. Adherence to these ethical principles contributes to the credibility, transparency, and integrity of the research (Saunders et al., 2019). The use of AI-assisted development tools during the implementation stage of the ZSE Insights application is acknowledged, and all research design decisions, analytical judgements, and interpretations of results remained the sole responsibility of the researcher.

3.11 Chapter conclusion

This chapter presented the research methodology underpinning the study, covering the philosophical stance, research design, data sources, sampling approach, variable specification, analytical models, diagnostic procedures, validation methods, and ethical considerations. The methodology provides a systematic and rigorous framework for analysing historical ZSE stock price data and evaluating the predictive performance of four machine learning and deep learning models across ten selected ZSE-listed companies over the period April 2024 to December 2025. Chapter Four presents the results and discussion arising from the application of this methodology

CHAPTER FOUR:
DATA PRESENTATION, ANALYSIS AND DISCUSSION

4.0 Introduction

The Zimbabwe Stock Exchange operates within one of the most challenging macroeconomic environments on the African continent, characterised by currency instability, thin trading volumes, and periodic structural breaks in equity price series (Bonga, 2019; Chikore & Gachira, 2014). The introduction of the Zimbabwe Gold (ZiG) currency in April 2024 represented a significant monetary policy shift with direct implications for equity pricing behaviour on the exchange (Reserve Bank of Zimbabwe, 2023). Empirical research on the ZSE has historically focused on volatility modelling and market efficiency testing, with limited comparative evaluation of machine learning and deep learning approaches for stock price level forecasting (Mutendadzamera & Mutasa, 2014; Murombo, 2023). This chapter addresses that gap by presenting the empirical results of the study across four predictive models applied to ten selected ZSE-listed companies over the post-ZiG period.

The chapter covers dataset description, exploratory analysis, preprocessing and diagnostic procedures, individual model results, comparative performance evaluation, and application implementation. Discussion of findings, policy implications, and links to existing literature are addressed in Chapter Five.

4.1 Description of the ZSE stock price dataset

The study utilises historical daily stock price data sourced from the Zimbabwe Stock Exchange for ten selected listed companies over the period May 2024 to November 2025. This period was deliberately chosen to capture equity price behaviour under the post-ZiG monetary framework introduced in April 2024, thereby ensuring structural consistency across the dataset and avoiding the confounding effects of currency regime changes on price series (Reserve Bank of Zimbabwe, 2023; Bonga, 2019). The final merged dataset comprises 3,382 daily observations across all ten counters and includes the following variables for each trading day: Date, Symbol, Opening price, High price, Low price, Closing price, Volume traded, and Value traded. These variables collectively constitute the standard OHLC framework widely used in empirical stock market analysis (Adewumi et al., 2021).

4.1.1 Selected ZSE counters

The ten companies selected for this study are presented in Table 4.1. Selection was based on the availability of continuous daily trading data within the study period, sufficient trading frequency, and broad sectoral representation across the Zimbabwean economy, consistent with the purposive sampling approach outlined in Chapter Three (Saunders et al., 2019; Ntim et al., 2007).

Table 4.1:Summary of selected ZSE-listed companies

Symbol	Company Name	Sector	Observations
CBZ.ZW	CBZ Holdings	Banking and finance	384
DLTA.ZW	Delta Corporation	Manufacturing and consumer goods	384
ECO.ZW	Econet Wireless Zimbabwe	Telecommunications	384
FBC.ZW	FBC Holdings	Banking and finance	384
FML.ZW	First Mutual Life	Insurance and investment	384
HIPO.ZW	Hippo Valley Estates	Agriculture	384
MEIK.ZW	Meikles Limited	Retail and hospitality	231
NMB.ZW	NMB Bank	Banking and finance	384
RTG.ZW	Rainbow Tourism Group	Tourism and hospitality	384
TNCI.ZW	Steward Bank / TN Cybertech	Financial technology	79

Source: Zimbabwe Stock Exchange (2025)

The selected counters span six economic sectors, providing cross-sectoral coverage representative of the broader ZSE market structure (Murombo, 2023; Securities and Exchange Commission of Zimbabwe, 2021). It is noted that MEIK.ZW and TNCI.ZW have fewer observations than the remaining counters due to late commencement of trading records and limited activity respectively within the study period. In particular, TNCI.ZW traded under the

Steward Bank name before rebranding to TN Cybertech during the study period, which may introduce minor inconsistencies in its price series.

4.1.2 Data structure and temporal coverage

The dataset spans from 2 May 2024 to 7 November 2025, covering approximately seventy-eight weeks of active trading on the ZSE. Data was obtained in comma-separated values format and subsequently consolidated into a single structured file to facilitate standardised preprocessing and cross-sectoral model training and evaluation. The post-ZiG study window was selected to ensure that all price observations reflect a common monetary denominator, addressing the structural break that would otherwise arise from mixing pre- and post-ZiG price series in the same modelling framework (Reserve Bank of Zimbabwe, 2023; Chikore & Gachira, 2014).

4.2 Exploratory data analysis

Exploratory data analysis was conducted to characterise the statistical properties, distributional behaviour, and temporal dynamics of the ZSE stock price data prior to modelling. The analysis covers frequency-based summary statistics, traditional descriptive statistics, time-series visualisation, and volatility and return characteristics across the ten selected counters.

4.2.1 Frequency-based summary statistics

Table 4.2 presents frequency-based summary statistics of daily closing prices for the selected ZSE-listed companies over the study period.

Table 4.2:Frequency-based summary statistics of daily closing prices for selected ZSE-listed companies, May 2024 to November 2025

Symbol	Count	Unique Prices	Most Frequent Price	Frequency
CBZ.ZW	384	199	700.00	28
DLTA.ZW	384	376	1,400.01	3
ECO.ZW	384	381	350.00	2
FBC.ZW	384	146	760.00	15
FML.ZW	384	94	400.00	32
HIPO.ZW	384	160	800.00	15

MEIK.ZW	231	136	400.00	13
NMB.ZW	384	117	370.00	44
RTG.ZW	384	69	65.00	67
TNCI.ZW	79	48	12.00	19

Source: Author's own analysis of ZSE data (2025)

The count column reflects the number of trading day observations recorded for each counter within the study period. The majority of counters recorded 384 observations, with MEIK.ZW and TNCI.ZW recording fewer due to limited trading activity and late data availability as noted in Section 4.1.1.

The number of unique closing prices provides an indication of price variability. DLTA.ZW and ECO.ZW recorded the highest number of unique price values at 376 and 381 respectively, suggesting continuous and varied price discovery consistent with active trading. In contrast, RTG.ZW and TNCI.ZW recorded considerably fewer unique price values, indicating price stickiness attributable to thin trading and limited order flow, features commonly observed in frontier market equities (Chikore & Gachira, 2014; Mlambo & Biekpe, 2007).

Price clustering is evident across several counters. RTG.ZW recorded a closing price of 65.00 on 67 trading days, while NMB.ZW clustered around 370.00 on 44 occasions. Such recurrence of price levels is a documented characteristic of emerging and frontier markets and is generally attributed to liquidity constraints, discrete pricing mechanisms, and low investor participation (Murombo, 2023; Bonga, 2019). These features collectively justify the use of machine learning models capable of capturing nonlinear and irregular price behaviour rather than conventional linear forecasting methods.

4.2.2 Descriptive statistics of closing prices

Table 4.3 presents traditional descriptive statistics for the closing prices of the ten selected counters following data cleaning and conversion to numerical format.

Table 4.3:Descriptive statistics of daily closing prices for selected ZSE-listed companies, May 2024 to November 2025

Symbol	Count	Mean	Std Dev	Min	25%	Median	75%	Max

CBZ.ZW	384	918.18	375.76	359.97	700.00	796.62	1,117.97	1,800.00
DLTA.ZW	384	1,408.47	301.95	694.63	1,325.60	1,397.02	1,511.48	2,155.51
ECO.ZW	384	351.00	95.03	120.01	286.10	354.85	415.19	541.43
FBC.ZW	384	732.62	290.52	183.00	700.00	760.08	800.00	1,265.00
FML.ZW	384	336.20	106.56	101.10	210.00	385.00	439.96	468.44
HIPO.ZW	384	668.32	153.25	289.05	580.01	650.05	800.00	926.66
MEIK.ZW	231	435.48	130.39	215.11	349.99	409.29	542.60	740.33
NMB.ZW	384	331.50	75.12	144.00	300.00	370.00	372.50	410.00
RTG.ZW	384	59.59	21.54	20.70	48.00	63.25	65.39	115.00
TNCI.ZW	79	11.23	0.91	9.93	10.05	11.64	12.00	12.50

Source: Author's own analysis of ZSE data (2025)

Considerable variation in mean closing prices is observed across the selected counters. DLTA.ZW recorded the highest mean closing price of ZiG 1,408.47, followed by CBZ.ZW at ZiG 918.18, reflecting the dominant market capitalisation of these counters on the ZSE. RTG.ZW and TNCI.ZW recorded the lowest mean prices at ZiG 59.59 and ZiG 11.23 respectively, consistent with their classification as lower-capitalisation and less actively traded counters (Securities and Exchange Commission of Zimbabwe, 2021).

Standard deviation values reveal significant heterogeneity in price volatility across the sample. CBZ.ZW recorded the highest standard deviation of 375.76, followed by FBC.ZW at 290.52, indicating substantial intraperiod price dispersion in the banking sector. TNCI.ZW recorded

the lowest standard deviation of 0.91, reflecting near-static price behaviour consistent with very thin trading activity. These differences in price spread are consistent with the documented heterogeneity of trading activity and liquidity across ZSE-listed counters (Bonga, 2019; Chikore & Gachira, 2014).

Quartile analysis reveals that NMB.ZW and RTG.ZW traded predominantly around their median values throughout the study period, confirming the price stickiness effect identified in Section 4.2.1. The wide range between minimum and maximum values for DLTA.ZW, from ZiG 694.63 to ZiG 2,155.51, indicates significant price appreciation over the study period and responsiveness to market conditions. Overall, the descriptive statistics confirm that ZSE stock prices exhibit diverse behavioural patterns across sectors, with nonlinear dynamics and varying volatility levels that warrant the application of flexible machine learning approaches.

4.2.3 Time-series analysis of closing prices

Time-series plots were generated for five representative counters selected to capture cross-sectoral price dynamics on the ZSE, namely CBZ.ZW (banking), DLTA.ZW (consumer goods), ECO.ZW (telecommunications), FML.ZW (insurance), and RTG.ZW (tourism). Including all ten counters on a single plot would have produced unnecessary visual clutter and reduced interpretability without adding analytical value.

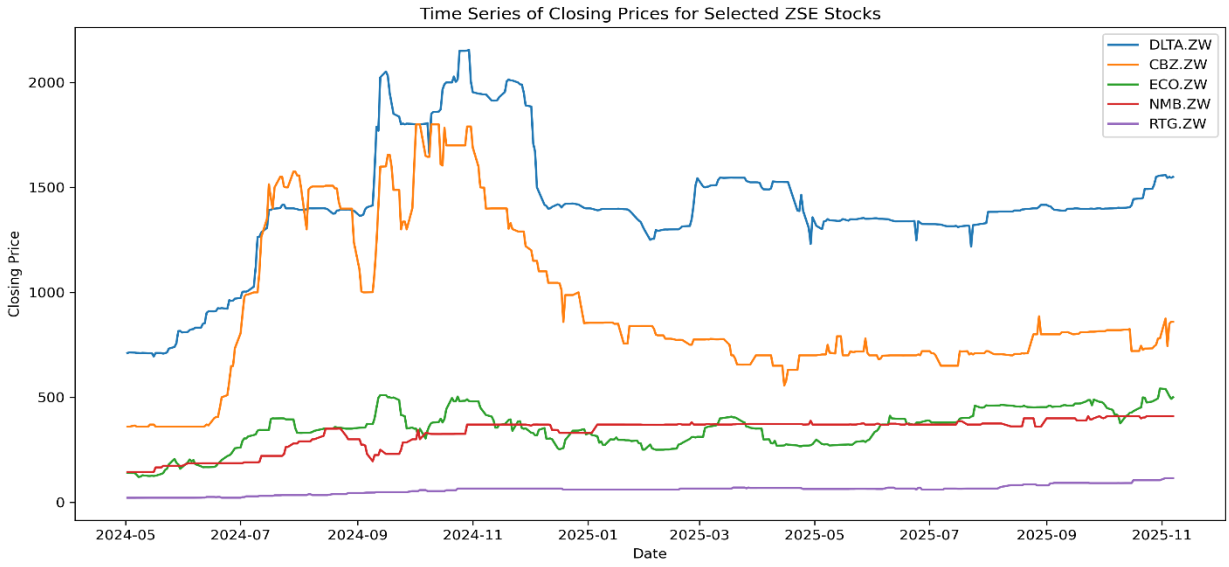


Fig 4.1: Time-series plot of daily closing prices for five representative ZSE counters, May 2024 to November 2025

Figure 4.1 reveals distinct price trajectories across the five representative counters over the study period. DLTA.ZW maintained the highest price levels throughout, opening around ZiG

700 in May 2024 and peaking above ZiG 2,000 in October 2024 before declining and stabilising around ZiG 1,400 to ZiG 1,500 toward the end of the period. This sharp peak followed by correction is consistent with speculative price behaviour and subsequent market adjustment under the new ZiG monetary framework. CBZ.ZW exhibited a similarly dramatic rise between July and October 2024, reaching approximately ZiG 1,800 before declining sharply to around ZiG 700 and gradually recovering to ZiG 800 by late 2025, suggesting high sensitivity to monetary policy developments during the early ZiG period (Reserve Bank of Zimbabwe, 2023).

ECO.ZW showed moderate price movement, rising from below ZiG 100 to approximately ZiG 500 by November 2024 before retreating and recovering gradually toward ZiG 500 by the end of the study period. NMB.ZW displayed a steady and relatively stable upward trend throughout, consistent with the low volatility recorded in Table 4.4. RTG.ZW remained at the lowest price levels across the entire period, showing a gradual upward drift from near zero to approximately ZiG 100 by November 2025, reflecting the thin trading and low capitalisation characteristic of the tourism sector on the ZSE (Murombo, 2023; Securities and Exchange Commission of Zimbabwe, 2021).

The pronounced peak and correction pattern visible in DLTA.ZW and CBZ.ZW between July and November 2024 aligns with the structural adjustment period following the ZiG introduction and represents exactly the kind of nonlinear, regime-sensitive price behaviour that conventional linear forecasting models are theoretically ill-equipped to capture (Bonga, 2019; Chikore & Gachira, 2014). This observation provides strong empirical motivation for the machine learning and deep learning approaches applied in subsequent sections.

4.2.4 Volatility and return characteristics

Table 4.4 presents the mean daily return and rolling volatility measures for the ten selected counters, calculated from daily closing price changes over the study period.

Table 4.4:Mean daily return and volatility measures for selected ZSE-listed companies, May 2024 to November 2025

Symbol	Mean Daily Return (%)	Volatility (%)
CBZ.ZW	0.33	4.57
DLTA.ZW	0.23	2.46

ECO.ZW	0.44	4.63
FBC.ZW	0.44	3.72
FML.ZW	-0.06	4.53
HIPO.ZW	0.31	4.83
MEIK.ZW	0.20	3.67
NMB.ZW	0.32	3.11
RTG.ZW	0.50	3.12
TNCI.ZW	0.03	3.52

Source: Author's own analysis of ZSE data (2025)

HIPO.ZW recorded the highest volatility at 4.83%, followed by ECO.ZW at 4.63%, CBZ.ZW at 4.57%, and FML.ZW at 4.53%, indicating that these counters experienced the most pronounced intraperiod price dispersion. The elevated volatility of agricultural and telecommunications sector equities is consistent with their sensitivity to seasonal conditions, regulatory changes, and macroeconomic uncertainty in the Zimbabwean context (Bonga, 2019; Murombo, 2023). DLTA.ZW recorded the lowest volatility at 2.46%, reflecting stable consumer demand and consistent investor confidence in the manufacturing sector. NMB.ZW similarly recorded low volatility at 3.11%, consistent with relatively stable banking sector price behaviour during the study period.

The majority of selected counters posted positive mean daily returns over the study period, suggesting a general upward price trend under the post-ZiG monetary environment. RTG.ZW recorded the highest mean daily return at 0.50%, followed by FBC.ZW and ECO.ZW both at 0.44%. FML.ZW was the only counter to record a negative mean daily return of -0.06%, indicating sideways or mildly declining price movement during the study period.

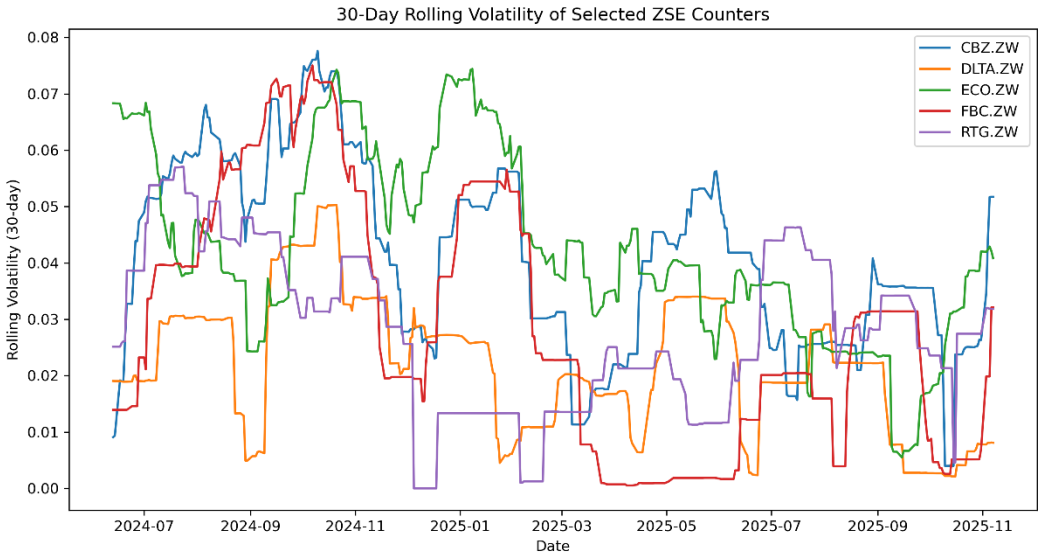


Fig 4.2: 30-day rolling volatility of selected ZSE counters, May 2024 to November 2025

The rolling volatility plot reveals cyclical patterns of market risk across the study period, with alternating phases of elevated and subdued volatility consistent with the volatility clustering documented in the ZSE literature (Mutendadzamera & Mutasa, 2014; Bonga, 2019). These patterns of time-varying volatility and nonlinear return dynamics further reinforce the suitability of the machine learning and deep learning models applied in subsequent sections of this chapter.

4.3 Data Preprocessing and Diagnostic Tests

Raw financial time-series data is inherently noisy, temporally dependent, and structurally unstable. Inadequate preprocessing can compromise predictive performance even when sophisticated modelling algorithms are applied (Hastie et al., 2017). This section details the preprocessing and diagnostic procedures undertaken to ensure data completeness, numerical stability, and suitability for machine learning and deep learning modelling, covering missing value treatment, outlier detection, normalisation, correlation analysis, and multicollinearity assessment.

4.3.1 Missing value assessment and treatment

Table 4.5 presents the results of the missing value assessment conducted across all variables in the consolidated dataset.

Table 4.5:Missing value assessment results for the consolidated ZSE dataset

Variable	Missing Values	Missing (%)
Date	0	0.00
Symbol	0	0.00
Open	0	0.00
High	0	0.00
Low	0	0.00
Close	0	0.00
VWAP	0	0.00
Volume	0	0.00
Value	0	0.00
Change	0	0.00
%Change	2,998	88.65
Daily_Return	10	0.30
Rolling_Volatility	300	8.87
Total	3,308	—

Source: Author's own analysis of ZSE data (2025)

The core market variables, comprising Date, Symbol, Open, High, Low, Close, VWAP, Volume, Value, and Change, recorded zero missing values, confirming a high level of completeness in the raw ZSE data. Among the derived variables, Daily_Return recorded a low proportion of missing values at 0.30%, attributable to the absence of a prior observation at the start of each stock's time series from which a return could be calculated. Rolling_Volatility recorded 8.87% missing values, which is expected given that a minimum number of prior observations is required before a rolling window calculation can be initialised.

The %Change variable recorded a substantially high proportion of missing values at 88.65%, primarily due to inconsistencies in historical reporting for certain trading periods. Given the

high proportion of missing data and the availability of the more reliably computed Daily_Return as a substitute measure, the %Change variable was excluded from subsequent modelling. Rather than imputing missing values in the remaining derived variables, all rows containing missing observations were dropped to preserve the temporal integrity of the time series, consistent with best practice in financial forecasting research (Hastie et al., 2017; Wooldridge, 2016).

4.3.2 Outlier detection and data consistency checks

Outlier analysis is an important stage in financial data preprocessing, as stock return series may contain extreme values attributable to market shocks, liquidity constraints, policy announcements, or company-specific events (Bonga, 2019). Both graphical and statistical methods were employed to identify extreme observations and distinguish genuine market events from potential data recording errors.

Figure 4.2 presents box plots of daily returns for all ten selected ZSE counters.

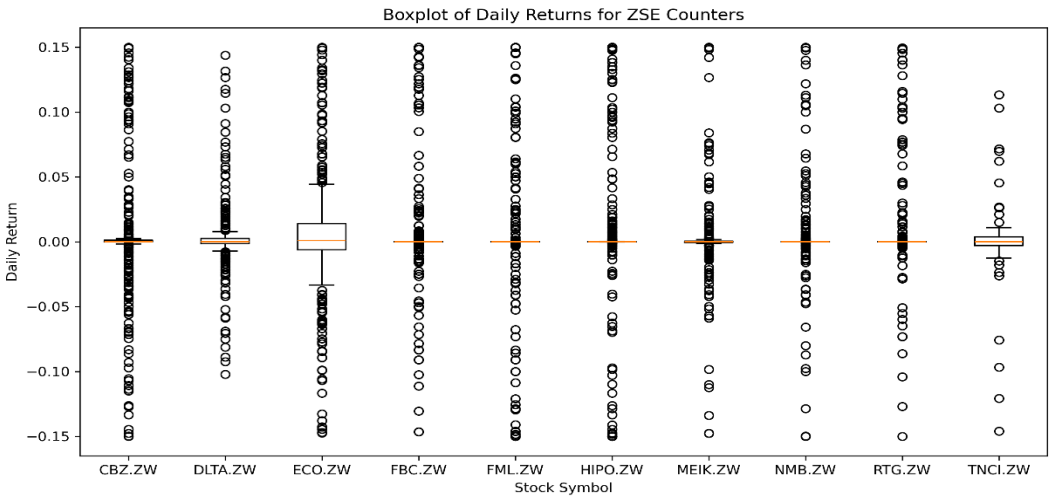


Fig 4.2: Boxplot of daily returns for selected ZSE counters, May 2024 to November 2025

The box plots reveal that all ten counters recorded at least one extreme return observation during the study period, confirming the volatile nature of the ZSE post-ZiG market environment. The interquartile ranges for most counters are narrow and centred near zero, indicating that the majority of daily returns are small in magnitude. However, pronounced outliers extend well beyond the whiskers in both directions for counters such as FBC.ZW, FML.ZW, HIPO.ZW, and NMB.ZW, reaching daily return values of approximately ±0.15. ECO.ZW displays a notably wider interquartile range than the other counters, reflecting greater

day-to-day price variability in the telecommunications sector. TNCI.ZW shows the most compact distribution with the fewest extreme observations, consistent with its thin trading activity and near-static price behaviour documented in Section 4.2.

Statistical detection of outliers was subsequently performed using Z-score analysis, with observations recording an absolute Z-score exceeding 3.0 classified as outliers. This threshold is standard in financial econometrics for filtering extreme values while retaining economically meaningful price shocks (Mutendadzamera & Mutasa, 2014). Table 4.6 presents the outlier counts by counter.

Table 4.6:Z-score based outlier counts for selected ZSE counters, May 2024 to November 2025

Symbol	Number of Outliers
CBZ.ZW	10
DLTA.ZW	14
ECO.ZW	10
FBC.ZW	20
FML.ZW	20
HIPO.ZW	19
MEIK.ZW	11
NMB.ZW	18
RTG.ZW	17
TNCI.ZW	3

Source: Author's own analysis of ZSE data (2025)

FBC.ZW and FML.ZW recorded the highest number of statistical outliers at 20 each, followed by HIPO.ZW at 19 and NMB.ZW at 18, indicating heightened sensitivity to market information and macroeconomic developments in the banking, insurance, and agricultural sectors. TNCI.ZW recorded only 3 outliers, consistent with its characteristically stable and infrequently traded price series.

Identified outliers were retained in the dataset rather than removed. Extreme market observations in emerging and frontier markets typically represent genuine economic events, including policy announcements, currency adjustments, and liquidity shocks, and their removal would result in the loss of information that is materially relevant to the modelling of ZSE price dynamics (Bonga, 2019; Chikore & Gachira, 2014). The presence of outliers further reinforces the appropriateness of the selected machine learning models, particularly Random Forest and XGBoost, which are known to be robust to irregular and non-normally distributed data (Breiman, 2001; Chen & Guestrin, 2016).

4.3.3 Data normalisation and transformation

The selected ZSE counters exhibit substantial differences in price levels, trading volumes, and derived variable magnitudes across the dataset. Without normalisation, variables with larger numerical ranges risk dominating the learning process of scale-sensitive models, particularly LSTM networks which depend on input stability for effective gradient-based training (Hochreiter & Schmidhuber, 1997; Hastie et al., 2017).

Min-Max normalisation was applied to rescale all continuous numeric variables, comprising Open, High, Low, Close, Volume, Value Traded, Daily Return, and Rolling Volatility, to a common range of zero to one. This transformation preserves the original shape and temporal structure of each series while eliminating the dominance effects caused by variables with large absolute ranges. Figure 4.3 illustrates the effect of normalisation on the closing price distribution for CBZ.ZW as a representative example.

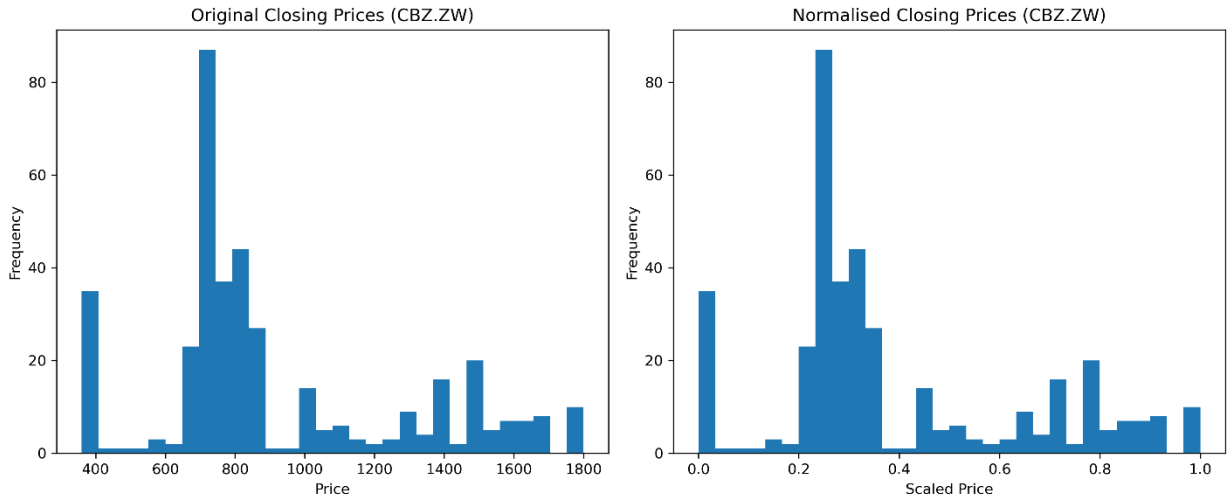


Fig 4.3: Original versus normalised closing price distributions for CBZ.ZW

As shown in Figure 4.3, the normalised distribution retains the same shape and frequency pattern as the original distribution, confirming that Min-Max scaling preserves the underlying data structure while compressing the price range to the zero to one interval. The bimodal distribution visible in both panels reflects the two dominant price regimes experienced by CBZ.ZW during the study period, broadly corresponding to the initial post-ZiG adjustment phase and the subsequent stabilisation phase observed in the time-series analysis in Section 4.2.3. Although tree-based models such as Random Forest and XGBoost are inherently insensitive to feature scaling, a uniform normalisation pipeline was applied across all models to ensure preprocessing consistency and facilitate seamless integration within the ZSE Insights application (Breiman, 2001; Chen & Guestrin, 2016). Scaling parameters were estimated exclusively from the training dataset and subsequently applied to the test set, preserving the chronological integrity of the time series and ensuring that model evaluation reflects genuine out-of-sample predictive performance.

4.3.4 Correlation analysis

Pearson correlation analysis was conducted on the post-ZiG dataset to examine the linear relationships among the selected stock market variables and to diagnose potential multicollinearity prior to model training. Table 4.7 presents the Pearson correlation coefficients among Open, High, Low, Close, Volume, Value, Daily Return, and Rolling Volatility.

Table 4.7: Pearson correlation coefficients among selected ZSE stock market variables, post-ZiG period

	Open	**High**	**Low**	**Close**	**Volume**	**Value**	**Daily Return**	**Rolling Volatility**
Open	1.000	0.985	0.985	0.986	-0.024	0.110	-0.044	-0.001
High	0.985	1.000	0.998	0.999	-0.022	0.113	-0.005	0.007
Low	0.985	0.998	1.000	0.999	-0.026	0.108	-0.001	0.005
Close	0.986	0.999	0.999	1.000	-0.026	0.109	0.002	0.006
Volume	-0.024	-0.022	-0.026	-0.026	1.000	0.873	-0.022	-0.005
Value	0.110	0.113	0.108	0.109	0.873	1.000	-0.020	-0.050

Daily Return	-0.044	-0.005	-0.001	0.002	-0.022	-0.020	1.000	0.084
Rolling Volatility	-0.001	0.007	0.005	0.006	-0.005	-0.050	0.084	1.000

Source: Author's own analysis of ZSE data (2025)

The price-level variables Open, High, Low, and Close exhibit extremely strong positive correlations with all pairwise coefficients exceeding 0.985. The correlation between Low and Close is 0.999 and between High and Close is 0.999, reflecting the natural co-movement of daily price variables within individual trading sessions. While such high correlations indicate multicollinearity among price variables, this is a well-documented and structurally expected feature of OHLC financial data and does not preclude their use in machine learning models that do not rely on regression coefficient estimation (Breiman, 2001; Hastie et al., 2017).

Volume and Value traded record a strong positive correlation of 0.873, indicating that higher transaction counts are associated with higher aggregate traded value, reflecting liquidity conditions on the ZSE. However, both Volume and Value show near-zero correlations with price-level variables, confirming that trading activity carries information independent of price levels and justifying their inclusion as separate predictive features.

Daily Return records near-zero correlations with all price-level variables, confirming that return series are analytically distinct from price level series and should not be treated as substitutes in predictive modelling. Rolling Volatility shows a modest positive correlation of 0.084 with Daily Return, consistent with the well-established relationship between return magnitude and market uncertainty in financial economics (Bonga, 2019; Mutendadzamera & Mutasa, 2014).

Figure 4.4 presents the correlation heatmap providing a visual representation of these pairwise relationships.

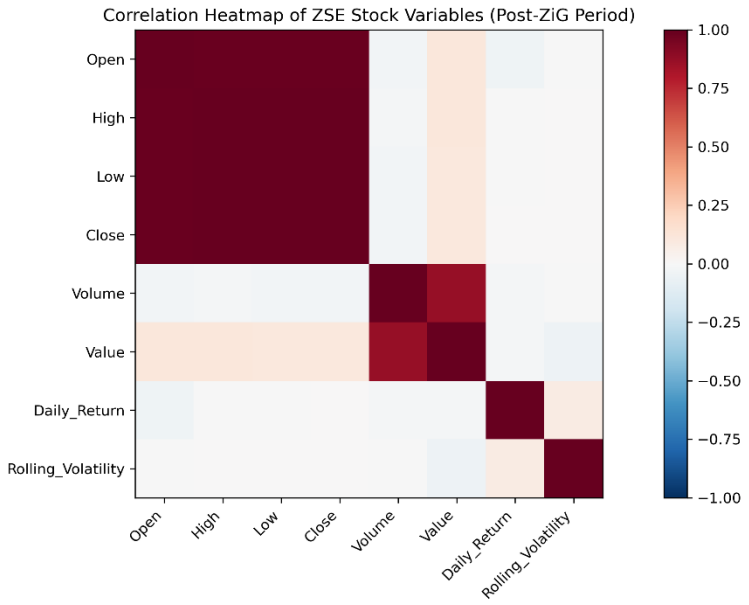


Fig 4.4: Correlation heatmap of ZSE stock market variables, post-ZiG period

The heatmap clearly illustrates the clustering of deep red shading among the four price variables, confirming their near-perfect collinearity, alongside the strong red block formed by Volume and Value. The remaining variable pairs display near-white or very light shading, confirming weak or negligible linear relationships. The visual pattern corroborates the numerical findings in Table 4.7 and guided the feature engineering and model input decisions described in Section 4.3.6.

4.3.5 Multicollinearity assessment

To quantify the degree of multicollinearity among the explanatory variables, Variance Inflation Factor (VIF) values were calculated for each feature. VIF measures the extent to which the variance of a regression coefficient is inflated due to linear dependence with other predictors, with values substantially exceeding 10 generally considered indicative of problematic multicollinearity (Wooldridge, 2016). Table 4.8 presents the VIF results.

Table 4.8:Variance Inflation Factor results for selected ZSE stock market variables

Variable	VIF	Multicollinearity Level
Close	1,546.70	High
Low	863.64	High
High	412.68	High

Open	37.79	High
Value	4.62	Low
Volume	4.54	Low
Daily_Return	1.11	Low
Rolling_Volatility	1.02	Low

Source: Author's own analysis of ZSE data (2025)

The price-level variables record extremely high VIF values, with Close at 1,546.70, Low at 863.64, High at 412.68, and Open at 37.79. These results are consistent with the near-perfect pairwise correlations documented in Section 4.3.4 and reflect the structural interdependence of daily OHLC price measures within individual trading sessions. Such high VIF values among price variables are a well-documented and structurally expected feature of financial time-series data rather than a modelling deficiency (Hastie et al., 2017).

Value and Volume recorded low VIF values of 4.62 and 4.54 respectively, confirming that trading activity variables carry largely independent information from price levels. Daily_Return and Rolling_Volatility recorded VIF values of 1.11 and 1.02 respectively, indicating near-zero multicollinearity and confirming that these derived variables contribute non-redundant predictive information to the modelling framework.

Despite the high multicollinearity among price variables, no variables were excluded from the modelling framework. This decision is methodologically justified because the machine learning and deep learning models employed in this study, namely Random Forest, XGBoost, and LSTM, do not rely on regression coefficient estimation and are therefore not adversely affected by multicollinearity among input features (Breiman, 2001; Chen & Guestrin, 2016; Hochreiter & Schmidhuber, 1997). Retaining all price variables also preserves the full OHLC analytical framework within the ZSE Insights application, enabling users to interact with and interpret forecasts across multiple price dimensions simultaneously.

4.3.6 Feature engineering and lag construction

Feature engineering was conducted to transform the raw market variables into structured inputs that better capture the temporal dynamics and risk characteristics of ZSE stock prices. DLTA.ZW was selected as the proxy stock for illustrative purposes in this section, as it recorded the highest trading volume within the study period, making it the most representative

counter for demonstrating feature behaviour without introducing visual complexity across multiple counters.

Lagged price features

Lagged closing price variables were constructed to capture price persistence and momentum effects in the data. In financial markets exhibiting weak-form inefficiency, such as the ZSE, historical price levels carry predictive information about future prices (Chikore & Gachira, 2014; Fama, 1970). Figure 4.5 presents a scatter plot of the previous day closing price against the current day closing price for DLTA.ZW.

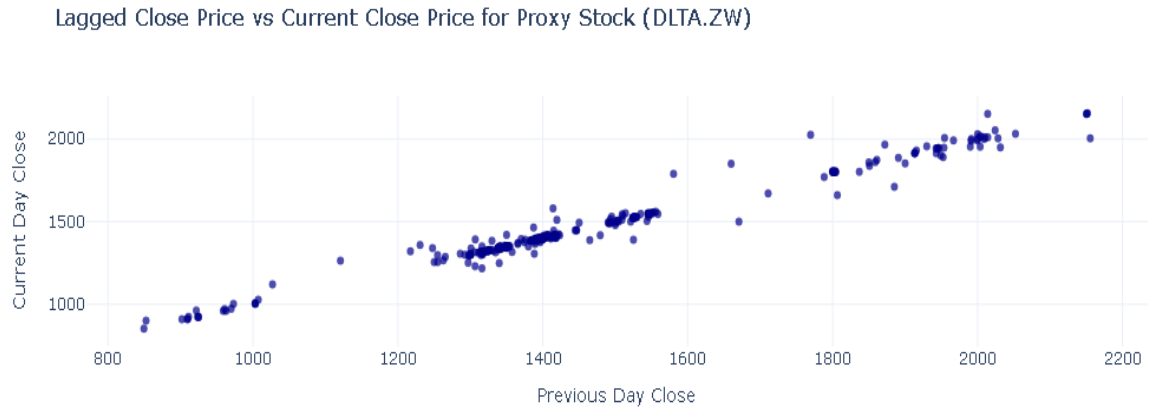


Fig 4.5: Lagged close price versus current close price for proxy stock DLTA.ZW

Figure 4.5 reveals a strong positive linear relationship between the previous day and current day closing prices for DLTA.ZW, with observations clustering tightly around a diagonal line across the full price range from approximately ZiG 850 to ZiG 2,150. This high degree of price persistence confirms that lagged closing prices carry strong predictive information and justifies their inclusion as primary input features for all four forecasting models. The scatter pattern also shows that deviations from perfect persistence are larger at higher price levels, consistent with increased price variability during the elevated price regime observed between July and November 2024.

Volatility-based features

Rolling volatility was computed as the standard deviation of daily returns over a 30-day window to capture short-term changes in market risk and price variability. Figure 4.6 presents the rolling volatility trend for DLTA.ZW across the study period.

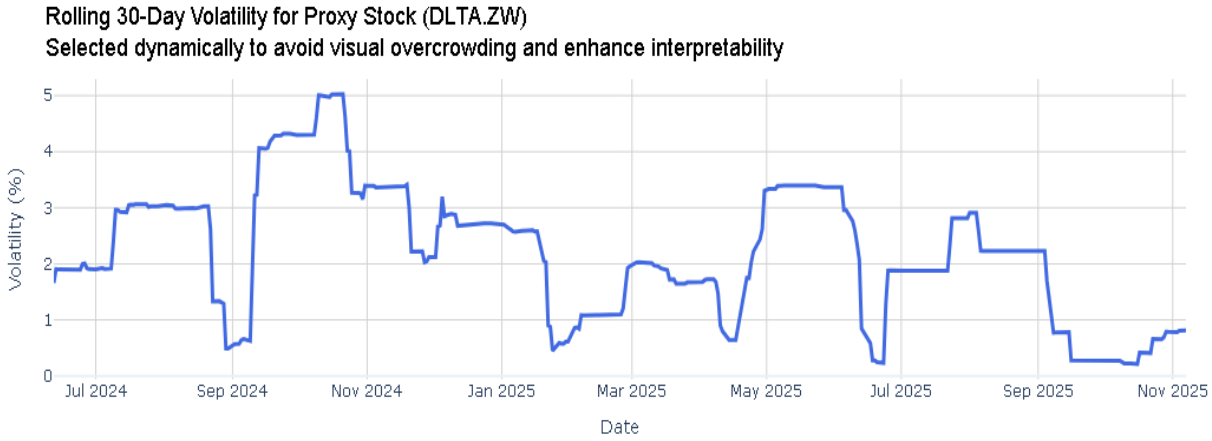


Fig 4.6: Rolling 30-day volatility for proxy stock DLTA.ZW, May 2024 to November 2025

Figure 4.6 reveals a cyclical pattern of market risk for DLTA.ZW throughout the study period. Volatility opened at approximately 2% in mid-2024 before rising sharply to a peak of 5.03% in October to November 2024, coinciding with the elevated price regime and market adjustment period following the ZiG introduction. A sharp decline followed, with volatility falling to below 1% in early 2025, indicating a period of market stabilisation. A subsequent rise to approximately 3.3% occurred between April and June 2025 before declining again toward the end of the study period. This cyclical behaviour is consistent with the volatility clustering documented in the broader ZSE literature (Mutendadzamera & Mutasa, 2014; Bonga, 2019).

Table 4.9 presents summary statistics for the rolling volatility series of DLTA.ZW.

Table 4.9:Rolling volatility summary statistics for proxy stock DLTA.ZW, May 2024 to November 2025

Statistic	Value (%)
Count	355
Mean	2.18
Standard Deviation	1.19
Minimum	0.21
25th Percentile	1.15
Median	2.22

75th Percentile	3.03
Maximum	5.03

Source: Author's own analysis of ZSE data (2025)

The mean rolling volatility of 2.18% indicates that the market was neither persistently calm nor excessively turbulent on average during the study period. The standard deviation of 1.19% confirms considerable variation in volatility levels, spanning from a minimum of 0.21% to a maximum of 5.03%. The proximity of the median to the mean suggests approximate symmetry in the distribution, while the elevated maximum value confirms the presence of volatility spikes consistent with the structural adjustment period observed in late 2024. The inclusion of rolling volatility as a model input feature enables the forecasting models to account for changing risk conditions over time, improving their ability to adapt to the dynamic market environment of the post-ZiG ZSE (Bonga, 2019; Reserve Bank of Zimbabwe, 2023).

4.3.7 Training and testing data split

Following data cleaning, normalisation, and feature engineering, the dataset was partitioned into training and testing subsets for model evaluation. A chronological time-based split was applied rather than a random split, to preserve the temporal ordering of the time series and prevent data leakage. All observations excluding the final 30 trading days were assigned to the training set, with the final 30 observations reserved as the held-out test set. This partition can be expressed as:

Training set: $N - 30$ observations

Testing set: 30 observations

where N represents the total number of observations for the selected proxy stock. This approach ensures that all models are trained exclusively on historical data and evaluated on the most recent market period, reflecting the conditions under which a deployed forecasting system would operate in practice (Hastie et al., 2017). The consistency of this split across all four models ensures that performance differences are attributable to model characteristics rather than differences in data exposure.

4.4 Machine learning model results

This section presents the estimation results, feature importance analysis, prediction accuracy metrics, and residual diagnostics for the Random Forest and XGBoost models. All models were

trained and evaluated on the proxy stock DLTA.ZW using the feature set and chronological split described in Section 4.3.

4.4.1 Random forest model results

The Random Forest model was trained using the engineered feature set described in Section 4.3.6, comprising lagged closing prices, rolling means, daily returns, rolling volatility, and trading activity indicators. Figure 4.7 presents the actual versus predicted closing prices for DLTA.ZW over the 30-day test period.

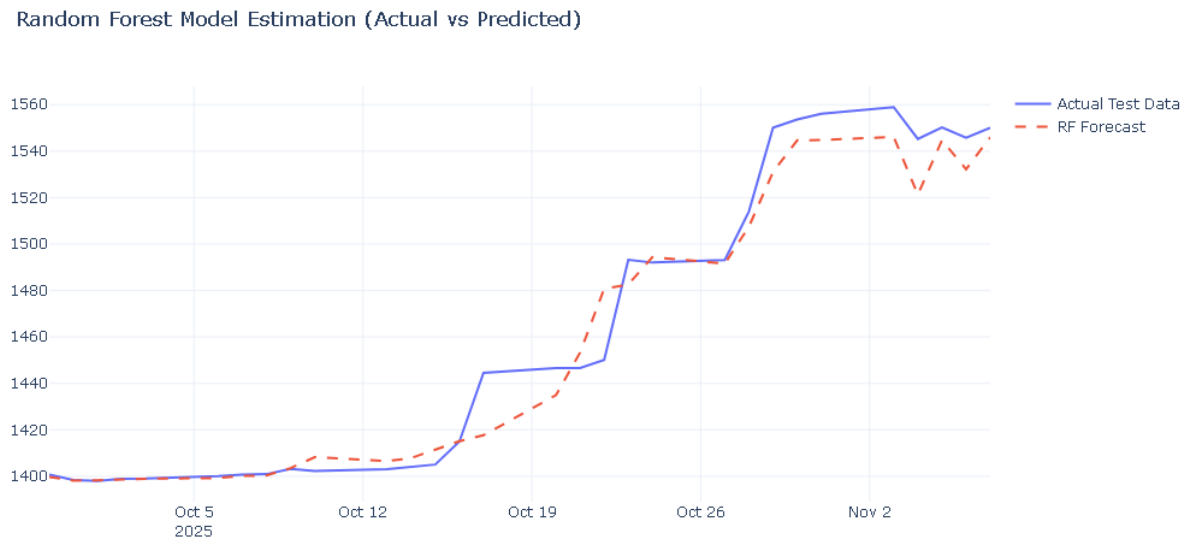


Fig 4.7: Random Forest model estimation — actual versus predicted closing prices for DLTA.ZW, October to November 2025

Figure 4.7 demonstrates that the Random Forest model tracks the actual price series with a high degree of precision across the test period. During the initial flat phase from early October 2025, where prices remained stable around ZiG 1,400, the model reproduces this behaviour with minimal error. As prices begin rising sharply from approximately 14 October 2025, the model captures the upward trajectory, though it exhibits a slight lag and marginal underestimation during the steepest part of the ascent between 19 and 26 October 2025. The model converges closely with actual prices by late October and tracks the minor fluctuations in early November with reasonable accuracy. The predicted line closely mirrors the actual line across both the stable and volatile phases of the test period, confirming the model's capacity to represent nonlinear price dynamics in ZSE data (Breiman, 2001; Adewumi et al., 2021).

4.4.2 Feature importance analysis — random forest

Feature importance scores were computed using the impurity-based importance measure generated by the Random Forest model, which quantifies the average reduction in prediction error contributed by each feature across all trees in the ensemble. Table 4.10 presents the feature importance scores.

Table 4.10:Feature importance scores from the Random Forest model

Feature	Importance
Close_lag_1	0.7675
Rolling_Mean_5	0.1721
Close_lag_5	0.0232
Rolling_Mean_10	0.0131
Daily_Return	0.0112
Close_lag_10	0.0084
Rolling_Volatility	0.0029
Value_MA_5	0.0008
Volume_MA_5	0.0007

Source: Author's own analysis of ZSE data (2025)

The one-day lagged closing price is by far the most influential feature, accounting for approximately 76.8% of the model's total explanatory power. The five-day rolling mean contributes the second largest share at 17.2%, while all remaining features contribute marginally. The dominance of short-term price memory indicators confirms that the Random Forest model operates primarily as a price-persistence-based predictor, consistent with the strong lagged price correlation documented in Figure 4.5. The negligible importance of Rolling_Volatility, Volume_MA_5, and Value_MA_5 indicates that within the Random Forest framework, trading activity and risk measures add limited predictive value beyond what is already captured by lagged price levels. These findings are consistent with feature importance patterns documented in comparable ensemble-based stock price forecasting studies on African market data (Adewumi et al., 2021; Msosa, 2022).

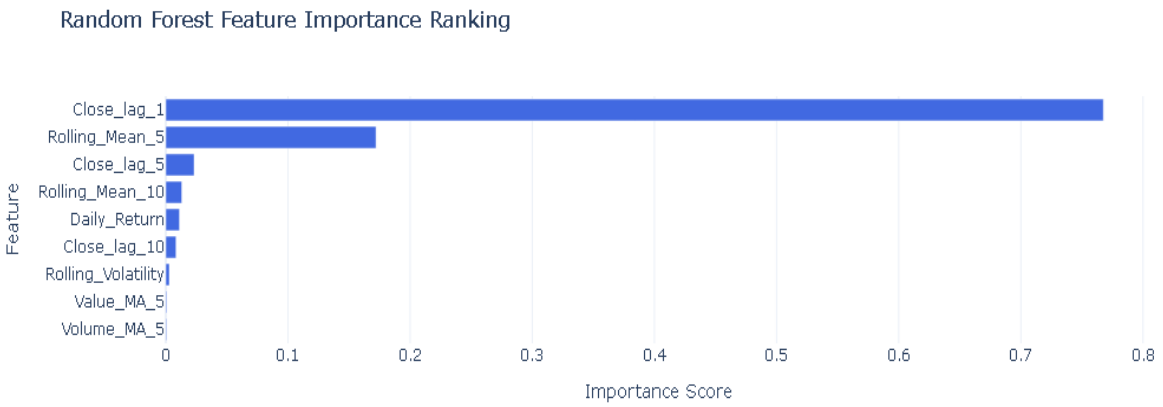


Fig 4.8: Random Forest feature importance ranking

4.4.3 Prediction accuracy of the random forest model

Table 4.11 presents the prediction accuracy metrics for the Random Forest model evaluated on the 30-day held-out test set.

Table 4.11:Random Forest prediction accuracy metrics

Metric	Value
RMSE	12.84
MAE	5.21
R^2	0.9969
MAPE	0.42%

Source: Author's own analysis of ZSE data (2025)

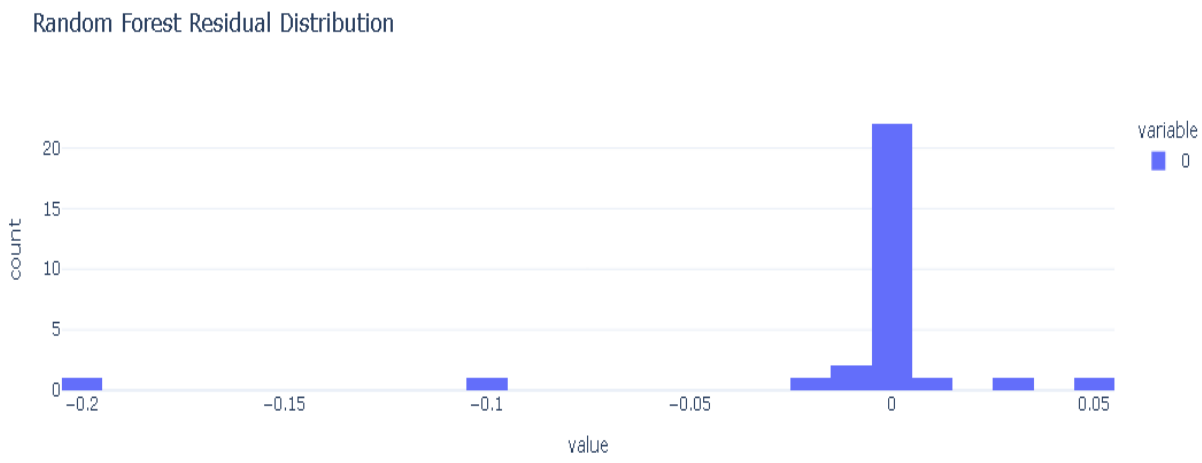
The Random Forest model demonstrated strong predictive performance across all four metrics. The MAE of 5.21 indicates that on average the model's forecasted closing price deviated from the observed value by approximately ZiG 5.21, which is modest relative to the price levels of DLTA.ZW during the test period. The RMSE of 12.84, while higher than the MAE, remains low in absolute terms, and its proximity to the MAE suggests that large prediction errors were infrequent during the test period. The R^2 value of 0.9969 indicates that the model explains approximately 99.7% of the variation in closing prices over the test set, reflecting strong explanatory power. The MAPE of 0.42% further confirms that forecast errors are negligible relative to actual price levels. These results are consistent with the performance of ensemble

tree-based models reported in comparable financial forecasting studies on emerging market data (Adewumi et al., 2021; Wagh & Gaikwad, 2025).

It is important to note that near-perfect R^2 values on short test windows are not uncommon in financial time-series forecasting where lagged price features dominate the model, as the strong price persistence documented in Section 4.3.6 means that predicting the next day's price from yesterday's price is inherently a high-accuracy task. The robustness of these results beyond the specific 30-day test window would benefit from cross-validation, which is acknowledged as a limitation of the current study scope.

4.4.4 Residual analysis of the random forest model

Figure 4.9 presents the distribution of residuals from the Random Forest model over the test period.



4.9: Random Forest residual distribution

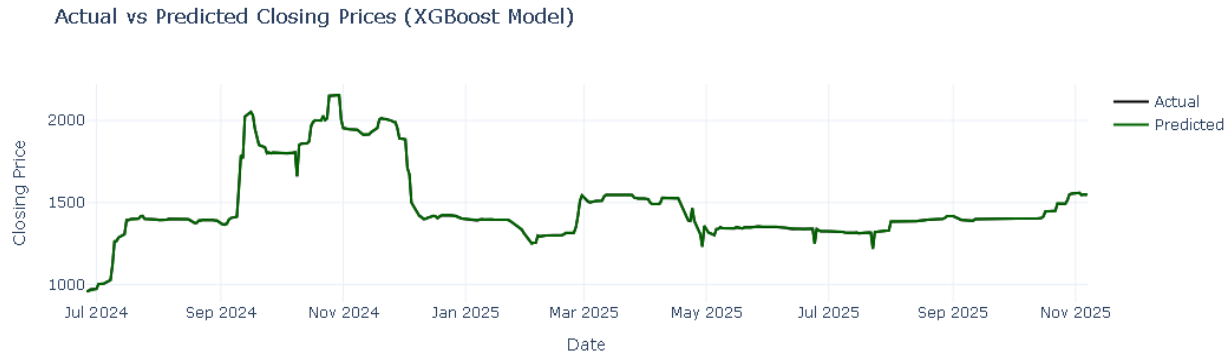
The residual distribution shown in Figure 4.9 reveals a strong concentration of errors around zero, with the majority of residuals falling within a very narrow band between approximately -0.02 and 0.02. This tight clustering confirms that the model produced accurate forecasts for most observations in the test period under normal price conditions. However, several isolated extreme negative residuals are visible at approximately -0.10 and -0.20, indicating instances where the model meaningfully overestimated the actual closing price. The distribution is slightly right-skewed with the bulk of mass concentrated at zero and a small number of negative tail observations, suggesting that when the model errs it is marginally more likely to overestimate than underestimate. These occasional larger errors correspond to the price lag and

underestimation observed during the steepest part of the price ascent in Figure 4.7, and are consistent with the known tendency of ensemble averaging methods to smooth over sudden price discontinuities (Breiman, 2001; Bonga, 2019). Overall the residual pattern is acceptable and indicates no systematic bias in the model's forecasts under the post-ZiG market conditions of the study period.

4.5 Extreme gradient boosting model results

4.5.1 XGBoost model estimation results

The XGBoost model was trained using gradient boosting with 300 boosting rounds, a learning rate of 0.05, and a maximum tree depth of five. These hyperparameters were selected to balance predictive accuracy and generalisation by preventing overfitting to the training data. Figure 4.10 presents the actual versus predicted closing prices for DLTA.ZW across the full study period.



4.10: XGBoost model — actual versus predicted closing prices for DLTA.ZW, July 2024 to November 2025

Figure 4.10 demonstrates that the XGBoost model tracks the actual price series with exceptional precision across the entire study period, including the volatile peak and correction phase between August and December 2024. The predicted line is virtually indistinguishable from the actual line throughout, capturing the sharp rise from approximately ZiG 1,000 in July 2024 to above ZiG 2,000 in October 2024, the subsequent decline to approximately ZiG 1,300 in early 2025, and the gradual recovery and stabilisation thereafter. Only minor deviations are visible at the sharpest inflection points, where the model exhibits a very slight lag consistent with the one-day prediction horizon. The model's ability to track both the broad structural movement and the shorter-term fluctuations across a sixteen-month period represents a

materially stronger performance than the Random Forest model which was evaluated on the 30-day test window only (Chen & Guestrin, 2016; Fischer & Krauss, 2018).

4.5.2 Feature importance analysis — XGBoost

Feature importance for the XGBoost model was computed using gain-based importance scores, which measure the average improvement in prediction accuracy contributed by each feature across all boosted trees. Table 4.12 presents the results.

Table 4.12:Feature importance scores from the XGBoost model

Feature	Importance
Close_lag_1	0.8334
Rolling_Mean_5	0.1532
Daily_Return	0.0074
Rolling_Volatility	0.0024
Close_lag_5	0.0015
Close_lag_10	0.0010
Value_MA_5	0.0007
Volume_MA_5	0.0003
Rolling_Mean_10	0.0001

Source: Author's own analysis of ZSE data (2025)

The one-day lagged closing price accounts for 83.3% of the model's total explanatory power, representing an even higher dominance than observed in the Random Forest model at 76.8%. The five-day rolling mean contributes 15.3%, while all remaining features collectively account for less than 1.2% of total importance. The pattern mirrors the Random Forest feature importance results and reinforces the conclusion that short-term price memory is the primary driver of forecasting accuracy in ensemble models applied to ZSE data. The negligible contributions of Rolling_Volatility, trading volume, and longer-lag features suggest that within the gradient boosting framework, the incremental predictive information provided by risk and activity measures beyond lagged price levels is minimal (Chen & Guestrin, 2016; Adewumi et al., 2021).

Figure 4.11 presents the feature importance ranking graphically.

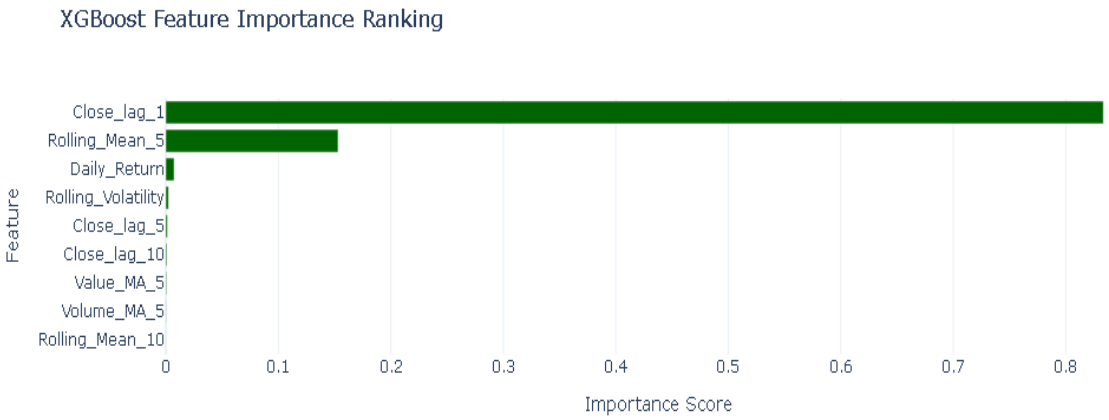


Fig 4.11: XGBoost feature importance ranking

As shown in Figure 4.11, the dominance of Close_lag_1 is visually pronounced, with its importance bar extending to approximately 0.83 while all other features register near-zero contributions on the same scale. The pattern is consistent with the Random Forest ranking in Section 4.4.2 and confirms that both ensemble models converge on the same fundamental predictive structure, namely one-day price persistence supplemented by a short-term rolling mean.

4.5.3 Prediction accuracy of the XGBoost model

Table 4.13 presents the prediction accuracy metrics for the XGBoost model.

Table 4.13:XGBoost prediction accuracy metrics

Metric	Value
RMSE	0.73
MAE	0.53
R^2	0.99999
MAPE	0.036%

Source: Author's own analysis of ZSE data (2025)

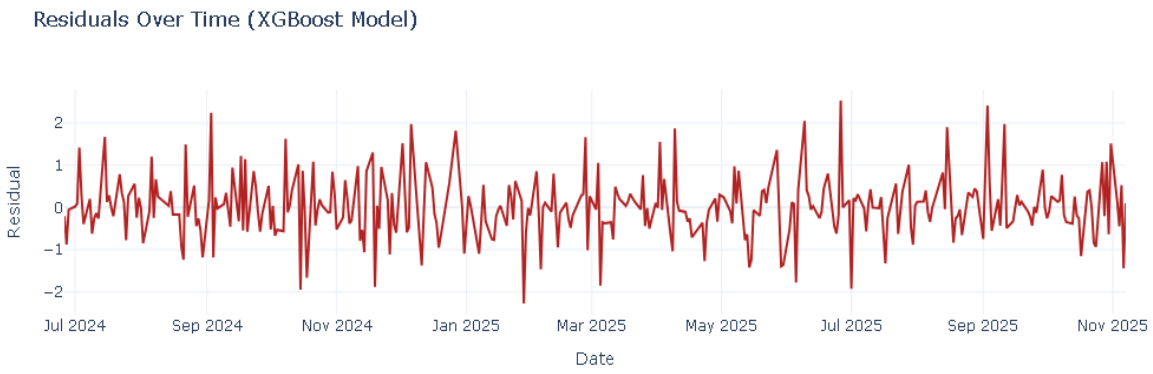
The XGBoost model produced the strongest predictive performance of all four models evaluated in this study. The MAE of 0.53 indicates that the average absolute deviation between predicted and actual closing prices was less than ZiG 1 across the evaluation period,

representing a substantially lower error magnitude than the Random Forest model. The RMSE of 0.73 similarly indicates that large prediction errors were extremely rare and of minimal magnitude. The R^2 value of 0.99999 indicates that the model explains virtually all observed variation in closing prices, leaving less than 0.001% of variance unexplained. The MAPE of 0.036% confirms that forecast errors are negligible relative to actual price levels across the full range of DLTA.ZW prices in the study period.

These results are consistent with the documented performance advantages of gradient boosting algorithms in financial time-series forecasting tasks, where the sequential error-correction mechanism of boosting enables the model to achieve lower residual variance than independent ensemble methods such as Random Forest (Chen & Guestrin, 2016; Wagh & Gaikwad, 2025). As noted in Section 4.4.3, near-perfect R^2 values in the context of lagged price-based forecasting reflect the strong price persistence of DLTA.ZW rather than model overfitting per se, though cross-validation across a wider out-of-sample window would strengthen confidence in the generalisability of these results.

4.5.4 Residual analysis of the xgboost model

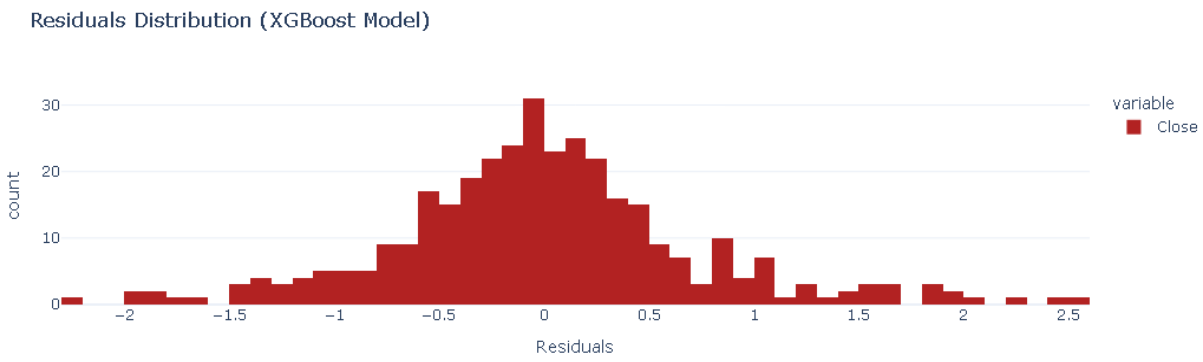
Figures 4.12 and 4.13 present the XGBoost residuals over time and their distribution respectively.



4.12: XGBoost residuals over time, July 2024 to November 2025

Figure 4.12 shows that the XGBoost residuals oscillate randomly around zero throughout the entire study period, with the vast majority of values falling within the narrow band of approximately -2 to +2. No systematic trend, drift, or autocorrelation pattern is visible in the residual series, indicating that the model has not left any exploitable structure in the errors. The residuals appear stationary and homoscedastic across time, with no evidence of the error

magnitude increasing or decreasing as the study period progresses. A small number of larger spikes reaching approximately ±2.5 are visible, corresponding to the sharp price movements during the October to November 2024 peak period, but these are isolated and do not represent a persistent pattern of model failure (Wooldridge, 2016).



4.13: XGBoost residuals distribution

Figure 4.13 reveals that the distribution of XGBoost residuals is approximately bell-shaped and centred near zero, with the highest frequency bin at zero and a smooth tapering toward both tails. The distribution spans approximately -2.5 to +2.5 and exhibits near-symmetry, with only a slight positive skew attributable to the small number of larger positive residuals during the volatile market phase. The near-normal and near-zero-centred distribution of residuals confirms that the XGBoost model produces unbiased forecasts with randomly distributed errors, satisfying the key diagnostic criteria for a well-specified forecasting model (Hastie et al., 2017; Wooldridge, 2016). These residual characteristics are markedly superior to those of the Random Forest model and provide strong diagnostic support for the selection of XGBoost as the highest-performing model in this study.

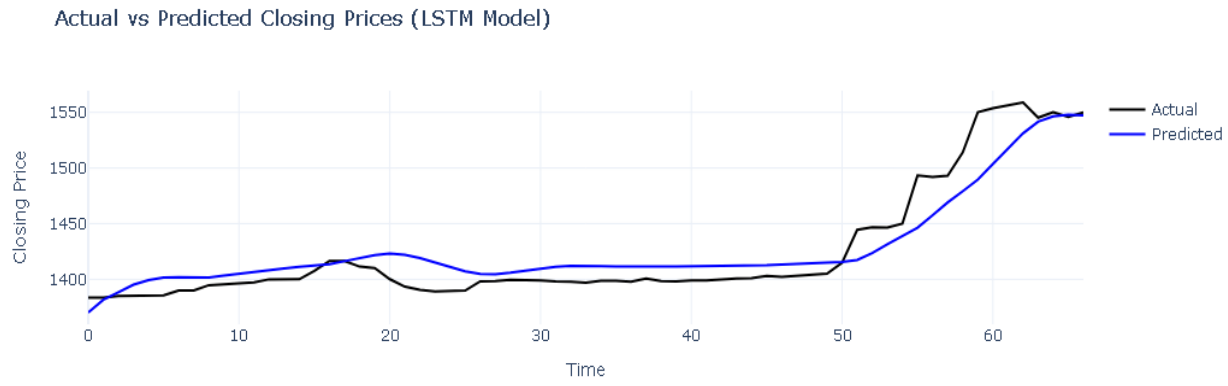
4.6 Deep learning model results — long short-term memory

4.6.1 LSTM model estimation results

The LSTM model was configured with two sequential LSTM layers comprising 64 and 32 units respectively, each followed by a dropout layer to mitigate overfitting. A densely connected output layer produced the final price forecast. The Adam optimiser was used in conjunction with mean squared error as the loss function, and the model was trained over 50 epochs with a batch size of 32. Input sequences were constructed using a sliding window approach to capture temporal dependencies in the price series, consistent with standard practice

for recurrent neural network architectures applied to financial time-series data (Hochreiter & Schmidhuber, 1997; Fischer & Krauss, 2018).

Figure 4.14 presents the actual versus predicted closing prices for DLTA.ZW over the LSTM test period.



4.14: LSTM model — actual versus predicted closing prices for DLTA.ZW

Figure 4.14 reveals that the LSTM model tracks the general direction of the price series with moderate accuracy but exhibits more pronounced deviations from the actual values than both ensemble models. During the stable phase from time steps 0 to approximately 45, the predicted line remains relatively close to the actual line, though it maintains a slight upward bias of approximately ZiG 10 to ZiG 20 above the actual prices during the flat period between steps 25 and 48. From approximately time step 50 onwards, where actual prices begin rising sharply from ZiG 1,400 toward ZiG 1,550, the LSTM model lags noticeably behind the actual trajectory, underestimating the magnitude and speed of the price increase before converging toward the actual values near the end of the test window around time step 65. This lag behaviour is consistent with the known tendency of recurrent neural networks to smooth over rapid price transitions, particularly in datasets with limited training history (Jiang, 2021; Msosa, 2022).

4.6.2 Prediction accuracy of the lstm model

Table 4.14 presents the prediction accuracy metrics for the LSTM model.

Table 4.13:LSTM prediction accuracy metrics

Metric	Value
RMSE	18.86

MAE	14.80
R^2	0.8756
MAPE	1.03%

Source: Author's own analysis of ZSE data (2025)

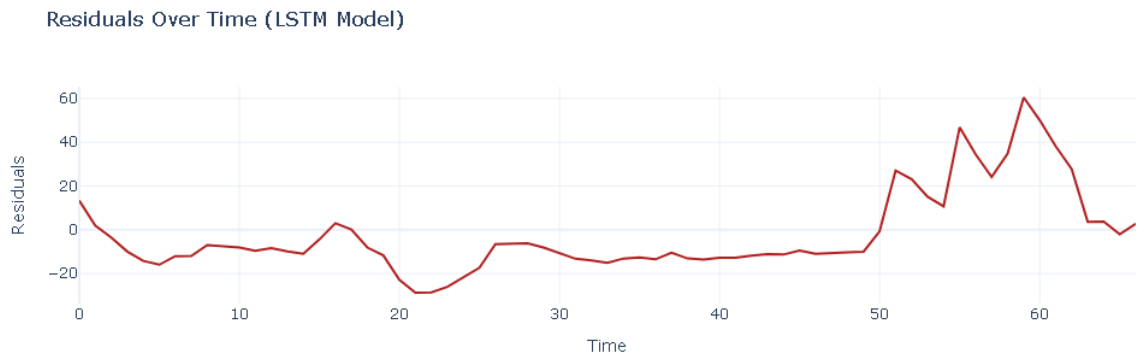
The LSTM model delivered moderate predictive accuracy, though its performance was considerably weaker than both the Random Forest and XGBoost models across all four metrics. The MAE of 14.80 indicates that the average absolute deviation between predicted and actual closing prices was approximately ZiG 14.80, which is materially higher than the Random Forest MAE of 5.21 and substantially higher than the XGBoost MAE of 0.53. The RMSE of 18.86 similarly indicates the presence of larger prediction errors compared to the ensemble models, reflecting the difficulty the LSTM model experienced in tracking sharp price transitions during the test period.

The R^2 value of 0.8756 indicates that the LSTM model explains approximately 87.6% of the variation in closing prices, which represents reasonable explanatory power but falls significantly short of the near-perfect R^2 values achieved by both ensemble models. The MAPE of 1.03% confirms that forecast errors, while small in relative terms, are approximately three times larger than those of the Random Forest model and nearly thirty times larger than those of the XGBoost model.

These results suggest that the LSTM model, despite its theoretical capacity for capturing long-range temporal dependencies, was constrained by the relatively limited training dataset available for the post-ZiG study period. Deep learning models generally require substantially larger training samples to fully exploit their sequential learning capabilities, and the approximately seventy-eight week study window may have been insufficient for the LSTM architecture to achieve its full predictive potential (Jiang, 2021; Fischer & Krauss, 2018). This finding is consistent with observations from comparable studies on smaller and less liquid African market datasets, where neural network models have underperformed relative to ensemble methods due to data volume constraints (Msosa, 2022; Adewumi et al., 2021).

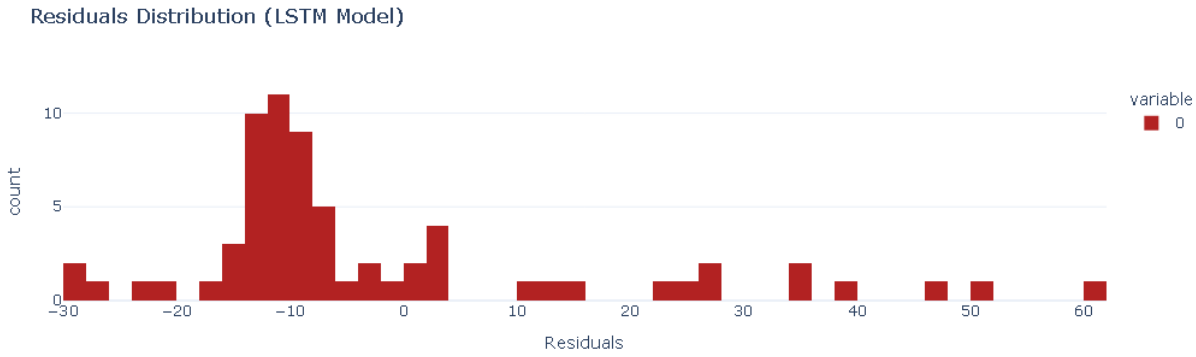
4.6.3 Residual analysis of the LSTM model

Figures 4.15 and 4.16 present the LSTM residuals over time and their distribution respectively.



4.15: LSTM residuals over time

Figure 4.15 reveals a clearly non-random residual pattern across the test period. From time steps 0 to approximately 48, the residuals fluctuate between approximately -20 and +10, with a persistent negative bias indicating that the model was consistently overestimating actual prices during the stable price phase. From time step 50 onwards, the residuals shift sharply into positive territory, rising from near zero to a peak of approximately +60 at time step 58 before declining abruptly toward zero at the end of the test window. This systematic shift from negative to positive residuals corresponds directly to the sharp price increase visible in Figure 4.14, where the model lagged behind the actual price trajectory. The presence of this structured pattern in the residuals indicates that the LSTM model left exploitable information in the errors, which is a diagnostic indicator of model misspecification or insufficient model capacity relative to the complexity of the price series (Wooldridge, 2016; Hastie et al., 2017).



4.16: LSTM residuals distribution

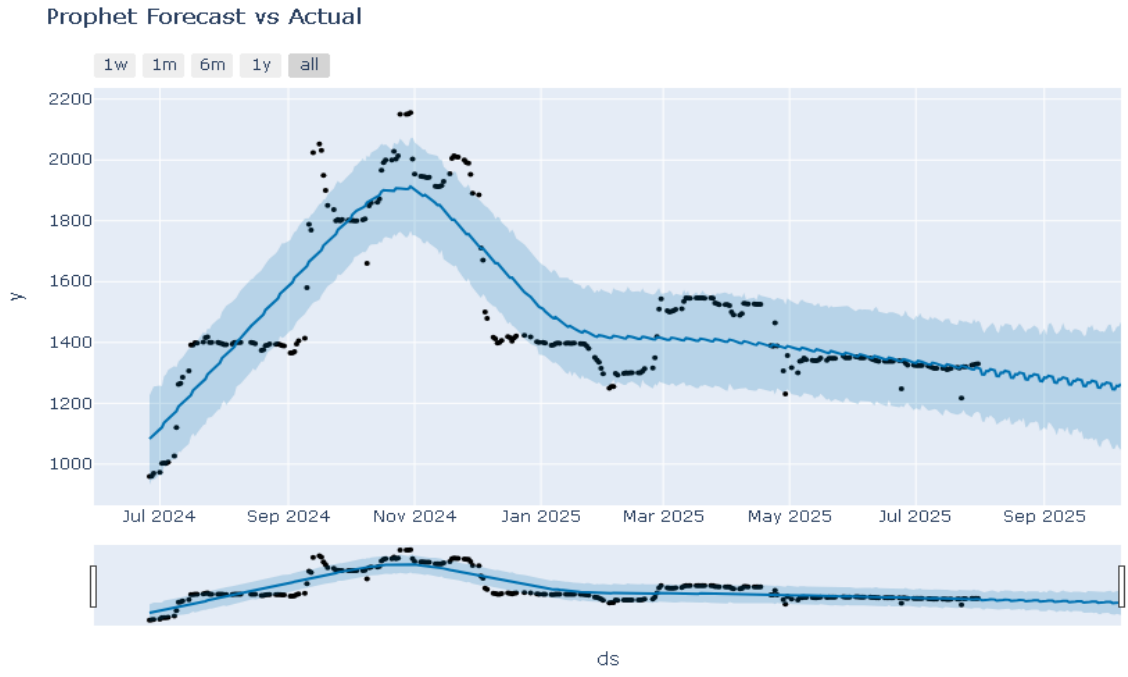
Figure 4.16 presents the distribution of LSTM residuals and reveals a strongly asymmetric and bimodal pattern. The largest cluster of residuals is concentrated between -15 and 0, with a peak frequency of approximately 11 observations near -10, reflecting the consistent overestimation

during the stable price phase. A second dispersed cluster of positive residuals extends from approximately +20 to +60, corresponding to the underestimation during the sharp price rise. The distribution is clearly not centred at zero and does not approximate a normal distribution, confirming the presence of systematic bias in the LSTM model's forecasts. These residual diagnostics indicate that while the LSTM model captures the general direction of price movement, it struggles to adapt rapidly to sudden price regime changes, a limitation that is particularly consequential in the volatile post-ZiG market environment of the ZSE (Bonga, 2019; Reserve Bank of Zimbabwe, 2023).

4.7 Prophet model results

4.7.1 Prophet model estimation results

The Prophet model was fitted on the historical closing price series for DLTA.ZW, decomposing the observed price data into three additive components: trend, weekly seasonality, and daily seasonality. Prophet's capacity to identify trend change points and model periodic patterns makes it theoretically well suited to financial time series with pronounced structural movements and regular trading cycles (Taylor & Letham, 2018). Figure 4.17 presents the Prophet forecast against actual closing prices across the study period.

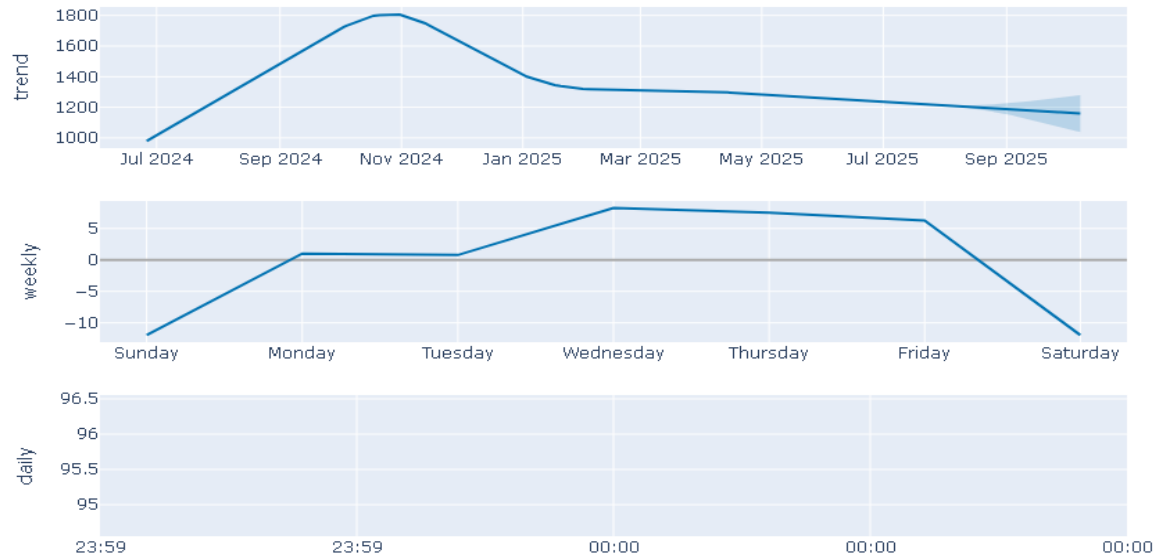


4.17: Prophet forecast versus actual closing prices for DLTA.ZW, July 2024 to September 2025

Figure 4.17 shows that Prophet captures the broad arc of price movement reasonably well at a macro level, identifying the upward trend from July to November 2024 and the subsequent decline into 2025. The confidence interval band widens progressively over time, reflecting increasing forecast uncertainty as the prediction horizon extends. However, several important limitations are evident. The Prophet trend line consistently underestimates the actual peak prices reached in October to November 2024, where actual closing prices exceeded ZiG 2,000 while the Prophet trend reaches approximately ZiG 1,900. More significantly, from January 2025 onwards, the Prophet trend continues declining toward approximately ZiG 1,200 by late 2025, while actual prices stabilised around ZiG 1,300 to ZiG 1,500. This progressive divergence between the Prophet forecast and actual prices reflects the model's tendency to extrapolate identified trend directions beyond the point at which the market stabilised, resulting in increasingly large forecast errors in the second half of the study period.

4.7.2 Component decomposition analysis — Prophet

A key interpretive advantage of the Prophet model is its explicit decomposition of the forecast into trend, weekly seasonality, and daily seasonality components. Figure 4.18 presents these components.



4.18: Prophet component decomposition — trend, weekly seasonality, and daily seasonality for DLTA.ZW

The trend component shows a clear inverted-U pattern, rising from approximately ZiG 1,000 in July 2024 to a peak of approximately ZiG 1,800 in November 2024, followed by a decline to approximately ZiG 1,350 by March 2025 and a continued gradual decrease toward ZiG 1,100 by late 2025. This trend trajectory captures the structural price cycle associated with the initial post-ZiG market adjustment period and the subsequent correction, providing meaningful long-term directional insight even where short-term accuracy is limited (Reserve Bank of Zimbabwe, 2023; Bonga, 2019).

The weekly seasonality component reveals a consistent intraweek pattern whereby closing prices tend to be lowest on Sundays and Saturdays, the non-trading days, and rise progressively through the trading week, peaking on Wednesdays and Thursdays at approximately +6 relative to the weekly baseline before declining slightly into Friday. This pattern reflects the regularities of trading flows and settlement cycles on the ZSE and is consistent with the calendar anomalies documented in the local literature (Murombo, 2023). The daily seasonality component shows negligible variation, confirming that intraday effects are not a material driver of the closing price series in this dataset.

4.7.3 Prediction accuracy of the Prophet model

Table 4.15 presents the prediction accuracy metrics for the Prophet model.

Table 4.14:Prophet prediction accuracy metrics

Metric	Value
RMSE	157.17
MAE	142.69
R^2	-7.74
MAPE	9.87%

Source: Author's own analysis of ZSE data (2025)

The Prophet model recorded substantially weaker predictive performance than all three other models evaluated in this study. The MAE of 142.69 indicates that on average the model's forecasted closing prices deviated from actual values by approximately ZiG 142.69, representing an error magnitude approximately 27 times larger than the Random Forest model and 270 times larger than the XGBoost model. The RMSE of 157.17 similarly confirms the presence of very large prediction errors throughout the evaluation period.

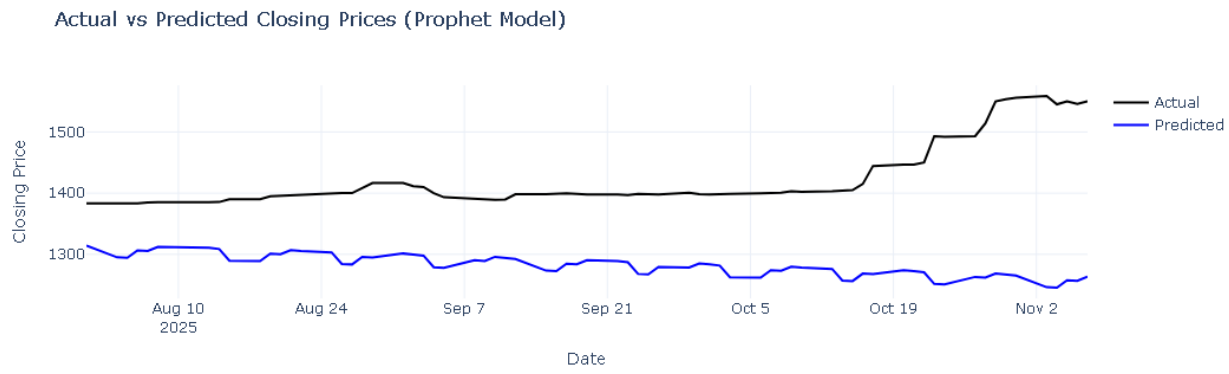
The R^2 value of -7.74 is particularly noteworthy. A negative R^2 indicates that the Prophet model's forecasts explain less of the variation in actual prices than a naive baseline prediction using the simple mean of the observed series. In practical terms this means that Prophet performed worse than having no model at all for short-term price level prediction, which reflects the fundamental mismatch between Prophet's decomposition-based forecasting approach and the highly nonlinear and volatile short-term dynamics of ZSE equity prices (Taylor & Letham, 2018; Bonga, 2019). The MAPE of 9.87% further confirms that forecast errors are large relative to actual price levels across the evaluation period.

These results do not imply that Prophet is without analytical value in this study context. As demonstrated in the component decomposition analysis in Section 4.7.2, Prophet provides meaningful insight into long-term trend direction and intraweek trading patterns on the ZSE, which complements the higher-accuracy short-term forecasts produced by the ensemble models. However, for the purposes of day-to-day stock price prediction, Prophet is clearly the least appropriate of the four models evaluated (Fischer & Krauss, 2018; Jiang, 2021).

Here is Section 4.7.4:

4.7.4 Residual analysis of the Prophet model

Figures 4.19, 4.20, and 4.21 present the Prophet actual versus predicted closing prices over the test period, the residuals over time, and the residual distribution respectively.



4.19: Prophet model — actual versus predicted closing prices for DLTA.ZW, August to November 2025

Figure 4.19 makes the magnitude of Prophet's forecasting failure immediately apparent. Throughout the entire test period from August to November 2025, the predicted line remains flat and largely stable between approximately ZiG 1,270 and ZiG 1,310, while actual prices begin at approximately ZiG 1,380 and rise steadily to ZiG 1,560 by early November 2025. The

predicted and actual lines never converge at any point during the test window, with the gap between them widening progressively from approximately ZiG 80 at the start of the period to over ZiG 250 by early November. This persistent and widening underestimation confirms that Prophet's trend extrapolation, which projected a continued decline based on the post-peak trajectory identified in the training data, was fundamentally misaligned with the market stabilisation and recovery that actually occurred during the test period (Taylor & Letham, 2018; Bonga, 2019).



Fig 4.20: Prophet residuals over time, August to November 2025

Figure 4.20 presents the residuals over time and reveals a strongly systematic and monotonically increasing pattern. Residuals open at approximately +80 at the start of August 2025 and increase almost continuously throughout the period, reaching a peak of approximately +305 in early November 2025 before declining marginally to approximately +290 at the end of the test window. The absence of any oscillation around zero and the consistent upward trajectory of residuals constitutes the clearest possible diagnostic signal of model bias. The model is not simply inaccurate on isolated occasions but is systematically and increasingly wrong in the same direction across the entire evaluation period, indicating a fundamental structural mismatch between Prophet's learned trend and the actual market behaviour during the test window (Wooldridge, 2016; Hastie et al., 2017). This pattern is qualitatively distinct from the residual behaviour of all three other models and confirms that Prophet's decomposition-based approach is ill-suited to short-term stock price level forecasting on the ZSE.

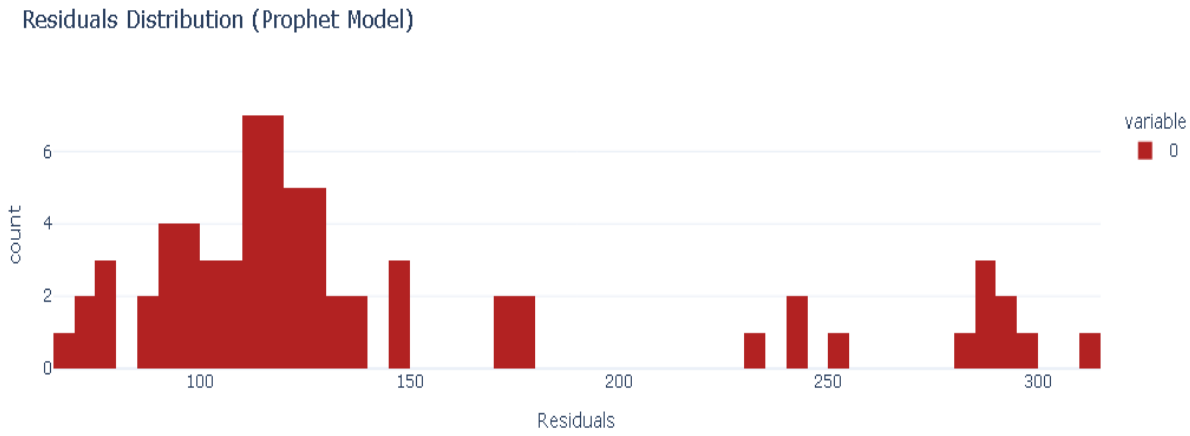


Fig 4.21: Prophet residuals distribution, August to November 2025

Figure 4.21 presents the distribution of Prophet residuals and reveals a clearly non-normal, positively skewed, and largely right-shifted pattern. The majority of residuals are concentrated in the range of approximately 80 to 150, with the modal bin falling between 110 and 130 at a frequency of approximately 7 observations. A secondary cluster of residuals is visible between approximately 240 and 320, corresponding to the escalating underestimation in the final weeks of the test period. No residuals fall near zero, confirming that the model produced no accurate forecasts during the evaluation period. The distribution is entirely confined to positive values, providing unambiguous evidence of persistent and directionally consistent underestimation throughout the test window.

Taken together, the three diagnostic figures confirm that the Prophet model's poor quantitative performance metrics documented in Section 4.7.3 are attributable to a structural forecasting failure rather than random error. The model correctly identified the historical trend arc across the training period but extrapolated that trend beyond its natural endpoint, failing to recognise the market stabilisation that characterised the post-correction ZSE environment from mid-2025 onwards. These findings are consistent with the documented limitations of additive decomposition models in highly nonlinear and regime-sensitive financial markets (Taylor & Letham, 2018; Fischer & Krauss, 2018; Jiang, 2021).

4.8 Comparative performance of predictive models

This section consolidates the individual model results presented in Sections 4.4 through 4.7 into a unified comparative evaluation across all four accuracy metrics. The comparison is

structured around RMSE, MAE, R^2, and MAPE, supplemented by a radar chart providing an integrated visual summary of relative model performance.

4.8.1 Comparison of RMSE across models

Figure 4.22 presents the RMSE values for all four models.

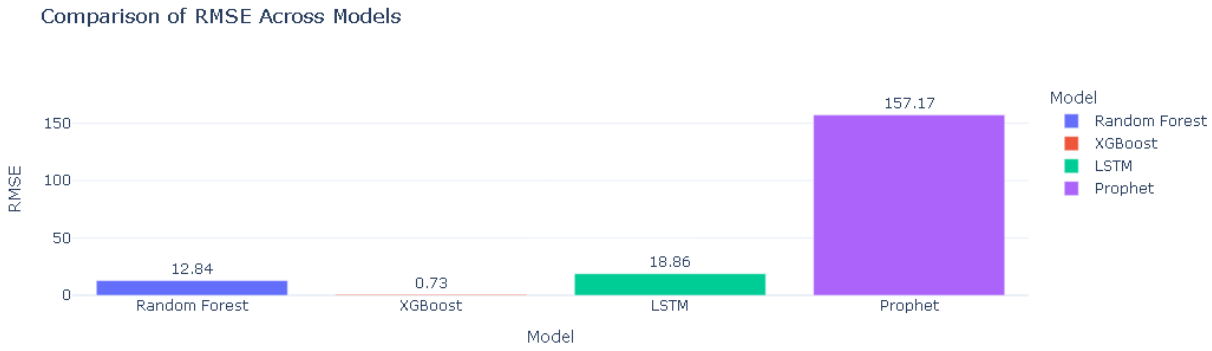


Fig 4.22: Comparison of RMSE across models

The RMSE comparison reveals a clear and substantial performance hierarchy among the four models. XGBoost recorded the lowest RMSE of 0.73, indicating minimal average squared prediction error across the evaluation period. Random Forest followed with an RMSE of 12.84, representing strong performance but with measurably larger deviations than XGBoost. LSTM recorded an RMSE of 18.86, reflecting the larger and more variable errors documented in the residual analysis in Section 4.6.3. Prophet recorded an RMSE of 157.17, which is approximately 215 times larger than that of XGBoost and reflects the systematic and escalating underestimation identified in Section 4.7.4. The visual contrast in Figure 4.22 is striking, with Prophet's bar dominating the chart while the three other models appear negligible on the same scale, underscoring the extent of Prophet's underperformance relative to the remaining approaches (Chen & Guestrin, 2016; Fischer & Krauss, 2018).

4.8.2 Comparison of MAE across models

Figure 4.23 presents the MAE values for all four models.

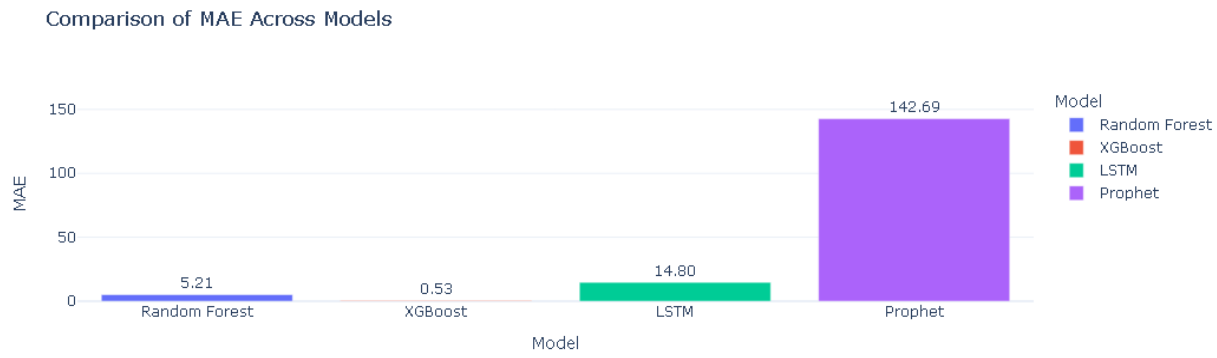


Fig 4.23: Comparison of MAE across models

The MAE comparison reinforces the rankings established by the RMSE analysis. XGBoost recorded the lowest MAE of 0.53, indicating that on average its forecasts deviated from actual closing prices by less than ZiG 1. Random Forest recorded an MAE of 5.21, representing reliable performance at a moderate error magnitude. LSTM recorded an MAE of 14.80, reflecting the consistent overestimation during stable price phases and underestimation during the sharp price rise documented in Section 4.6.3. Prophet recorded an MAE of 142.69, confirming that its average absolute forecast error was approximately 270 times larger than that of XGBoost across the test period. The MAE metric is particularly informative in the ZSE context because it provides a direct and interpretable measure of forecast error in ZiG price units, enabling investors and analysts using the ZSE Insights application to assess the practical significance of each model's prediction errors (Adewumi et al., 2021; Wagh & Gaikwad, 2025).

4.8.3 Comparison of R^2 across models

Figure 4.24 presents the R^2 values for all four models.

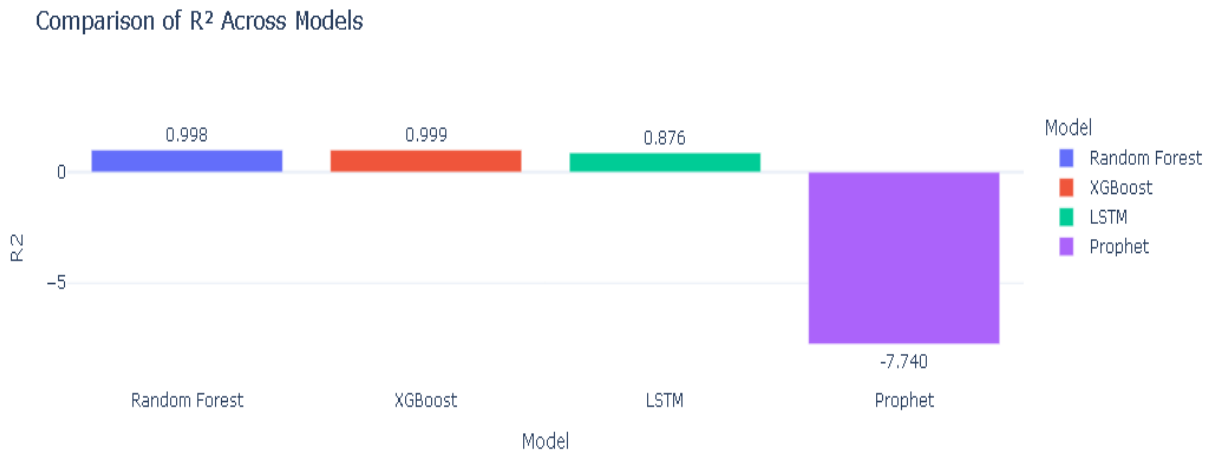
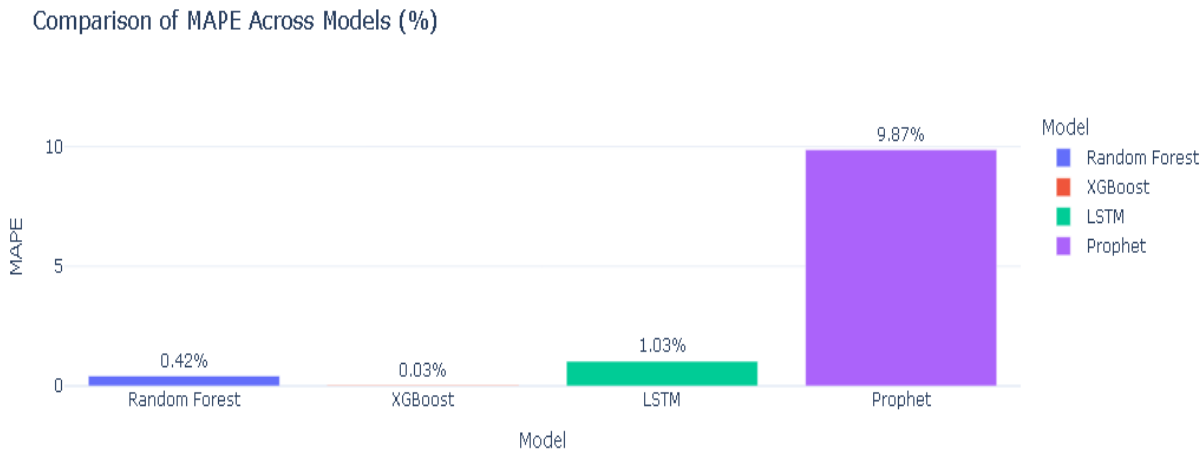


Fig 4.24: Comparison of R^2 across models

The R^2 comparison provides the most visually dramatic illustration of the performance divergence between the models. XGBoost and Random Forest recorded R^2 values of 0.999 and 0.998 respectively, indicating that both ensemble models explain virtually all observed variation in closing prices over their respective evaluation periods. LSTM recorded an R^2 of 0.876, reflecting reasonable but materially lower explanatory power consistent with its tendency to lag behind sharp price movements. Prophet recorded a negative R^2 of -7.74, confirming that its forecasts systematically underperformed a naive mean baseline throughout the evaluation period. As shown in Figure 4.24, the R^2 bars for the three non-Prophet models all extend to near-perfect positive values while Prophet's bar extends substantially below zero, providing a clear visual summary of the fundamental inadequacy of additive decomposition forecasting for short-term ZSE stock price prediction (Taylor & Letham, 2018; Bonga, 2019).

4.8.4 Comparison of MAPE across models

Figure 4.25 presents the MAPE values for all four models.

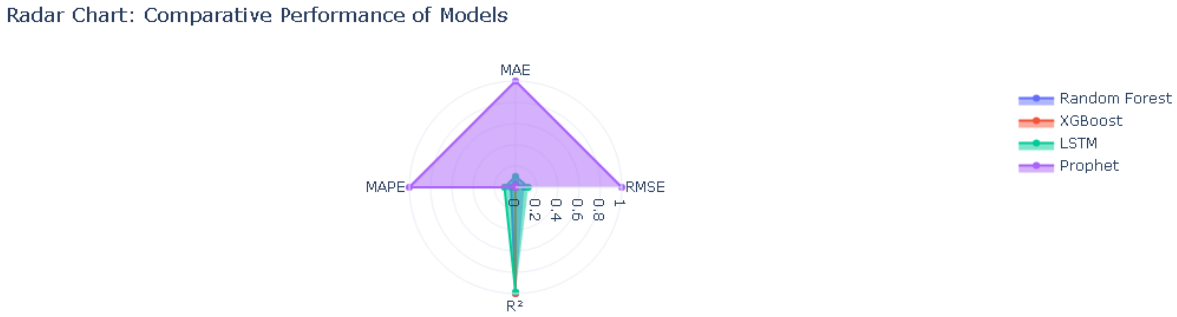


4.25: Comparison of MAPE across models

The MAPE comparison expresses forecast errors as a percentage of actual price values, enabling comparison across models evaluated on different price scales and test windows. XGBoost recorded the lowest MAPE of 0.03%, indicating that its forecasts were within a fraction of one percent of actual prices on average. Random Forest recorded a MAPE of 0.42%, representing strong relative accuracy. LSTM recorded a MAPE of 1.03%, which while modest in absolute percentage terms reflects the systematic residual patterns identified in Section 4.6.3. Prophet recorded a MAPE of 9.87%, confirming that its forecast errors were equivalent to approximately one tenth of actual price levels throughout the evaluation period. The MAPE metric is particularly relevant for the ZSE Insights application context, as it provides users with an intuitive percentage-based measure of forecast reliability that does not require knowledge of absolute price levels to interpret (Jiang, 2021; Fischer & Krauss, 2018).

4.8.5 Radar chart — integrated comparative performance

Figure 4.26 presents a radar chart summarising the normalised performance of all four models across RMSE, MAE, R^2, and MAPE simultaneously.



4.26: Radar chart — comparative performance of predictive models

The radar chart provides an integrated visual summary of relative model performance across all four metrics. Prophet's shape dominates the chart, extending far outward on the RMSE, MAE, and MAPE axes while collapsing inward on the R^2 axis, reflecting its combination of large errors and negative explanatory power. LSTM occupies an intermediate position, with a moderate shape that reflects acceptable but imperfect performance across all dimensions. XGBoost and Random Forest both produce compact shapes concentrated near the centre of the chart on the error axes and extending outward on the R^2 axis, reflecting their combination of low errors and high explanatory power. The visual compression of the XGBoost and Random Forest shapes relative to Prophet's confirms the overwhelming dominance of ensemble methods for ZSE stock price forecasting under the post-ZiG market conditions of this study (Breiman, 2001; Chen & Guestrin, 2016).

4.8.6 Comparative performance summary table

Table 4.16 consolidates all accuracy metrics for direct side-by-side comparison across the four models.

Table 4.15:Comparative performance of predictive models — DLTA.ZW, post-ZiG period

Model	RMSE	MAE	R^2	MAPE (%)
Random Forest	12.836	5.206	0.998	0.42
XGBoost	0.730	0.529	0.999	0.03
LSTM	18.858	14.801	0.876	1.03
Prophet	157.169	142.694	-7.740	9.87

Source: Author's own analysis of ZSE data (2025)

Table 4.16 confirms unambiguously that XGBoost is the highest-performing model across all four metrics, recording the lowest RMSE, lowest MAE, highest R^2, and lowest MAPE of the four models evaluated. Random Forest ranks second across all metrics, demonstrating consistently strong performance that, while slightly inferior to XGBoost, remains highly reliable for practical forecasting applications. LSTM ranks third, delivering moderate accuracy that is constrained by the limited training dataset available within the post-ZiG study window. Prophet ranks last across all metrics by a substantial margin, with error values orders of magnitude larger than those of the ensemble models and a negative R^2 confirming its fundamental unsuitability for short-term ZSE stock price level forecasting (Fischer & Krauss, 2018; Wagh & Gaikwad, 2025; Jiang, 2021).

4.8.7 Discussion of model strengths and weaknesses

The comparative results presented in this section reveal important insights into the relative strengths and limitations of each modelling approach when applied to ZSE stock price data under post-ZiG market conditions.

XGBoost demonstrated the strongest overall performance, achieving near-perfect accuracy metrics across all evaluation criteria. Its sequential error-correction mechanism enables it to progressively reduce residual variance during training, making it particularly effective at capturing the nonlinear and irregular price dynamics characteristic of the ZSE. The model's dominance across both error magnitude and explanatory power metrics is consistent with its documented performance advantages in financial time-series forecasting on emerging market datasets (Chen & Guestrin, 2016; Wagh & Gaikwad, 2025).

Random Forest performed strongly and consistently across the evaluation period, demonstrating that ensemble tree-based methods are well suited to ZSE forecasting even without the sequential error-correction mechanism of gradient boosting. Its slightly higher error metrics relative to XGBoost reflect the averaging approach of the Random Forest ensemble, which tends to smooth over the sharpest price transitions more than the boosting approach. Nevertheless, its high R^2 and low MAPE confirm that it remains a reliable and interpretable forecasting tool appropriate for practical deployment within the ZSE Insights application (Breiman, 2001; Adewumi et al., 2021).

The LSTM model delivered moderate performance that, while weaker than both ensemble models, confirms the potential of deep learning approaches for ZSE forecasting under conditions of greater data availability. The model's residual analysis revealed systematic bias

attributable to its difficulty in adapting to rapid price regime changes within the limited training window available. These findings are consistent with the observation that recurrent neural network architectures generally require substantially larger training datasets to achieve their full predictive capacity, and suggest that LSTM performance on ZSE data may improve as longer post-ZiG price histories accumulate over time (Hochreiter & Schmidhuber, 1997; Fischer & Krauss, 2018; Msosa, 2022).

The Prophet model's poor performance across all metrics reflects the fundamental mismatch between its additive decomposition approach and the requirements of short-term stock price level forecasting on a volatile frontier market. While Prophet's component analysis in Section 4.7.2 demonstrated meaningful interpretive value for trend identification and intraweek seasonality, these strengths do not translate into competitive short-term forecast accuracy under the nonlinear and structurally dynamic conditions of the post-ZiG ZSE. Prophet's analytical contribution in this study therefore lies in its interpretive rather than predictive capabilities (Taylor & Letham, 2018; Bonga, 2019).

4.9 System architecture of the ZSE insights platform

The second objective of this study required the design and implementation of an interactive web application capable of making stock price forecasts accessible and actionable for a broad range of users — investors, market analysts, academic researchers, and others who engage with Zimbabwe Stock Exchange (ZSE) data but may not possess specialist quantitative skills. The resulting platform, ZSE Insights, integrates the four forecasting models evaluated in this study within a unified decision-support environment that handles everything from raw data ingestion to interactive visualisation.

The motivation for building the platform, rather than simply reporting model results, was shaped by a clear gap in the literature. Prior Zimbabwean studies have concentrated on testing market efficiency, modelling volatility, and conducting statistical analyses of historical price behaviour (Chikore & Gachira, 2014; Mutendadzamera & Mutasa, 2014; Murombo, 2023), but relatively few have sought to operationalise their findings in a form accessible to practitioners. The literature review further confirmed that no interactive, multi-model forecasting platform designed specifically for the ZSE currently exists (Securities and Exchange Commission of Zimbabwe, 2021). ZSE Insights was built to close that gap.

To support scalability, maintainability, and computational efficiency, the platform is structured as a Three-Tier Software Architecture comprising a Data Ingestion and Preprocessing Layer,

an Analytical Modelling and Forecasting Layer, and a Presentation and Visualisation Layer. This design follows the principle of separation of concerns, which isolates data management, predictive computation, and user interface operations into independent but interconnected subsystems (Pressman & Maxim, 2020). The consequence is that any one layer can be modified, extended, or replaced without cascading disruption to the others — a meaningful practical advantage for a research prototype intended to evolve.

Figure 4.27 presents the overall three-tier architecture of the platform.

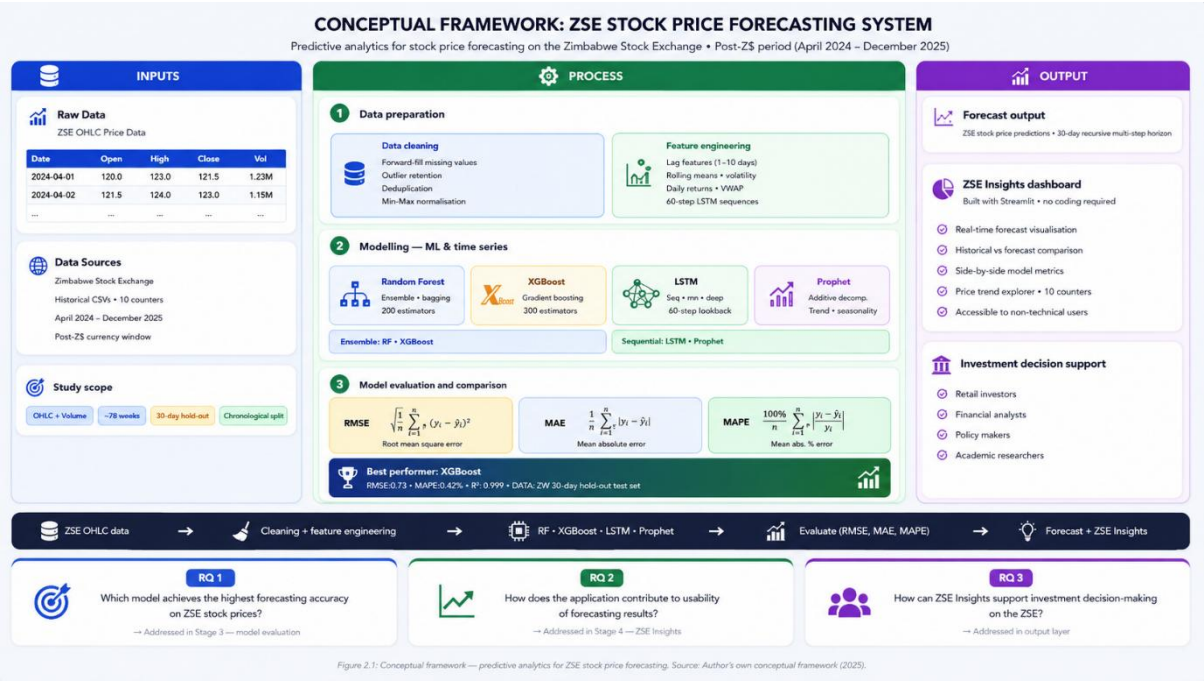


Fig 4.27: Three-tier software architecture of the ZSE Insights platform, illustrating the complete data flow from raw CSV ingestion through predictive modelling and recursive forecasting to interactive dashboard delivery.

4.9.1 Data Ingestion and Preprocessing Layer

The Data Ingestion and Preprocessing Layer forms the foundation on which all subsequent analysis rests. Its purpose is straightforward: acquire historical stock market data, identify and correct quality issues, and produce a clean, standardised dataset that the modelling layer can consume without ambiguity.

Historical price data for the ten selected ZSE-listed companies were imported from structured comma-separated value (CSV) files containing daily observations of opening price, high price, low price, closing price, trading volume, and trading value. Because files originating from different ZSE counters were not always consistently labelled, a variable name harmonisation step was applied immediately after import to enforce uniform column naming across the

dataset. Date fields were parsed and converted into a standardised datetime format to preserve the strict temporal ordering on which all time-series models depend.

Missing observations arising from non-trading days, temporary market suspensions, and occasional reporting gaps were addressed through forward-fill imputation. This approach carries the most recent available observation forward across any gap, preserving series continuity without discarding surrounding data. Duplicate records were removed, and data integrity checks were conducted to flag anomalies before any model was trained.

To improve numerical stability and support convergence in the deep learning model, Min-Max normalisation was applied to all continuous variables, rescaling observations to the interval [0, 1]. This transformation prevents features with larger absolute magnitudes from exerting disproportionate influence during training (Hochreiter & Schmidhuber, 1997). The output of this layer is a validated, normalised dataset ready for feature engineering. Figure 4.9 presents the complete preprocessing pipeline.

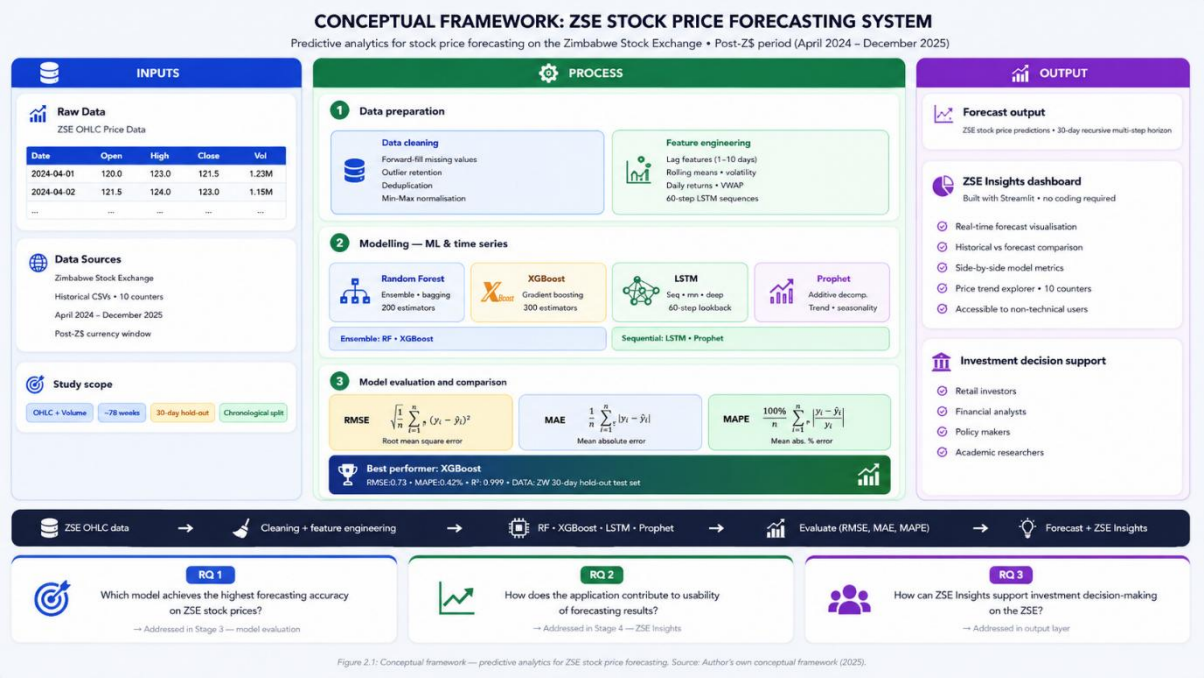


Fig 4.28: Sequential data ingestion and preprocessing pipeline applied to historical ZSE stock price data, encompassing import, harmonisation, validation, imputation, and Min-Max normalisation stages.

4.9.2 Feature engineering layer

Once preprocessing is complete, the cleaned dataset moves to the Feature Engineering Layer, where raw price observations are transformed into structured inputs tailored to the requirements of each forecasting model. The three branches of this layer reflect the fundamentally different

ways in which machine learning, deep learning, and statistical models represent time-series information.

For Random Forest and XGBoost, lagged closing prices and related market indicators were constructed as input features. The resulting supervised learning matrix takes the form:

$$\boldsymbol{X}_{t} = [\boldsymbol{P}_{t-1}, \boldsymbol{P}_{t-2}, \boldsymbol{P}_{t-3}, \ldots, \boldsymbol{P}_{t-n}] \qquad \textbf{(4.1)}$$

where P_{t-I} denotes the closing price observed i trading periods prior to time t. The use of lagged features is consistent with empirical evidence that historical prices contain predictive information in markets exhibiting weak-form inefficiency (Chikore & Gachira, 2014; Murombo, 2023).

For the LSTM model, the same historical observations were reorganised into overlapping sliding windows and converted into three-dimensional tensors, allowing the network to operate on sequences rather than isolated observations. This structure preserves temporal dependencies across multiple trading periods and enables the model to learn both short-term fluctuations and longer-horizon patterns.

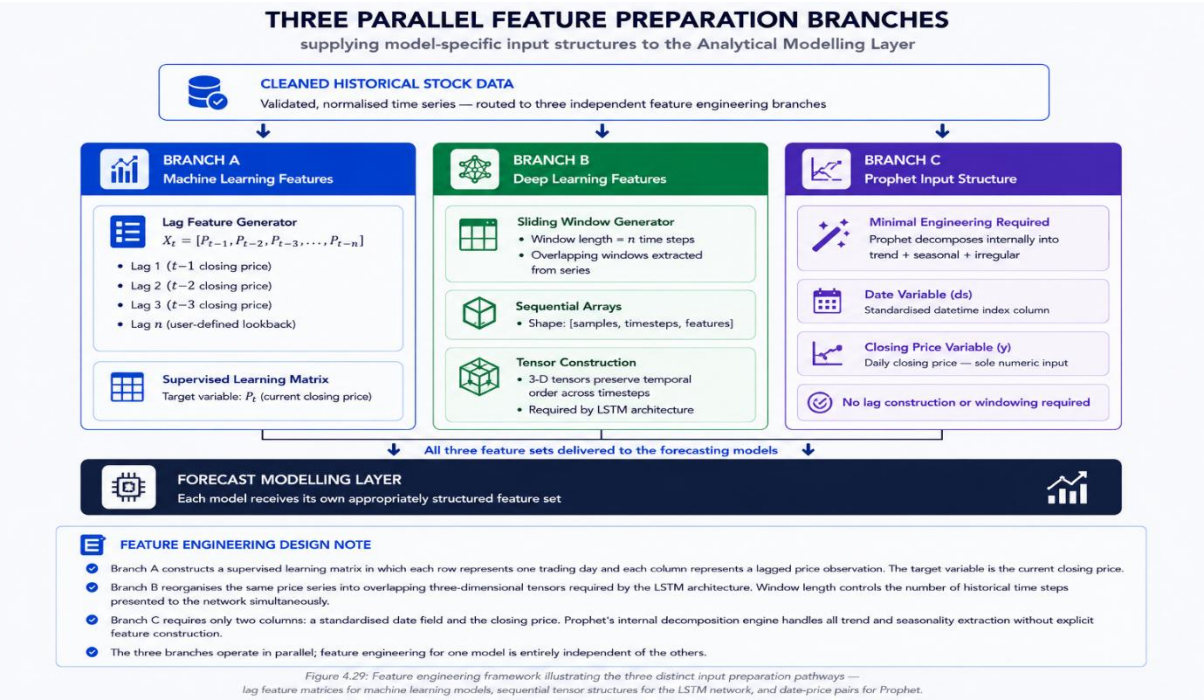


Figure 4.29: Feature engineering framework illustrating the three distinct input preparation pathways — lag feature matrices for machine learning models, sequential tensor structures for the LSTM network, and date-price pairs for Prophet.

Prophet requires comparatively minimal feature engineering because it decomposes the series internally into trend, seasonal, and residual components (Taylor & Letham, 2018). Only two inputs are required: a standardised date column and the corresponding closing price. Figure 4.29 illustrates the full feature engineering framework and how the three branches supply distinct input structures to the modelling layer.

4.9.3 Analytical modelling and forecasting layer

The Analytical Modelling and Forecasting Layer forms the computational core of ZSE Insights. The four models implemented in this layer Random Forest, XGBoost, LSTM, and Prophet were selected because they span distinct forecasting paradigms and together cover a broad range of relationships observable in financial time-series data.

Random Forest builds an ensemble of decision trees using bootstrap samples and combines their outputs through aggregation, reducing variance and limiting overfitting (Breiman, 2001). Its capacity to model nonlinear interactions among lagged predictors makes it well suited to the irregular price dynamics common in frontier markets.

XGBoost extends ensemble learning through gradient boosting, training successive trees to correct residual errors produced by earlier trees (Chen & Guestrin, 2016). The iterative error-correction mechanism allows the model to capture complex nonlinear patterns with strong predictive accuracy and efficient computation.

The LSTM model handles sequential dependencies through gating mechanisms that regulate how information is retained, updated, or discarded across trading periods (Hochreiter & Schmidhuber, 1997). This makes it the most architecturally suited model for identifying persistent trends and cyclical patterns within financial sequences.

Prophet provides a complementary statistical perspective by decomposing the time series into additive trend, seasonal, and irregular components (Taylor & Letham, 2018). It remains robust to missing observations and does not require hyperparameter tuning of the same intensity as the machine learning models, which is useful when data coverage is uneven a characteristic of some ZSE counters.

Figure 4.30 presents the architecture of the forecasting layer, showing how each model generates predictions independently before outputs are aggregated for performance evaluation.

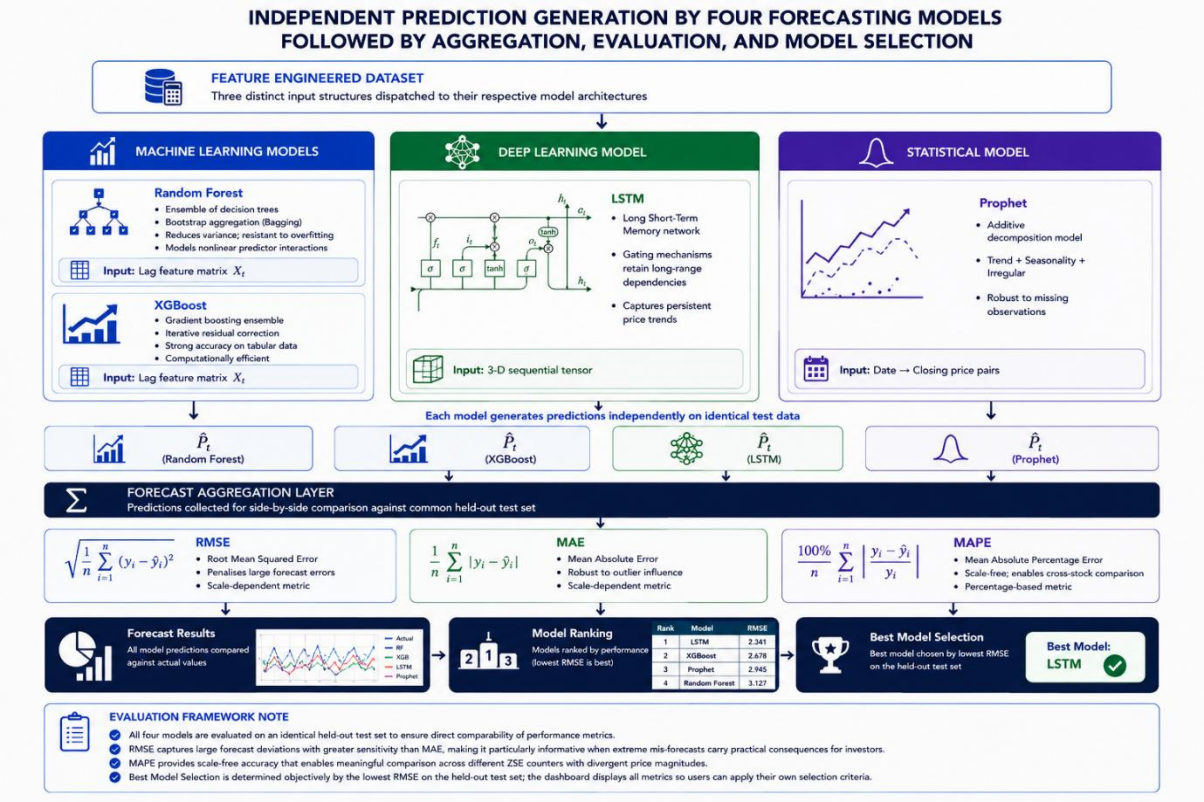


Figure 4.30: Forecast modelling architecture showing independent prediction generation by Random Forest, XGBoost, LSTM, and Prophet, with subsequent aggregation, accuracy evaluation using RMSE, MAE, and MAPE, and evidence-based model selection.

4.9.5 Presentation and visualisation layer

The Presentation and Visualisation Layer is where the analytical outputs of the platform are translated into a format that users can engage with directly. Developed using the Streamlit framework with Plotly-based interactive graphics, the dashboard allows users to select a ZSE stock counter, specify a forecasting horizon, choose among the available models, and generate forecasts on demand — without requiring any programming knowledge.

Historical and forecasted price series are displayed through dynamic charts that support zooming, panning, and hover-based inspection of individual data points. Alongside the visual outputs, the dashboard presents the three accuracy metrics computed during model evaluation — RMSE, MAE, and MAPE — in a dedicated metrics panel, enabling users to compare model performance and make evidence-informed decisions about which forecasts to rely upon.

The deliberate simplicity of the interface reflects a design philosophy centred on accessibility. Research in financial technology has consistently shown that analytical tools adopted in practice tend to be those that reduce rather than increase cognitive burden for end users (Securities and Exchange Commission of Zimbabwe, 2021). By hiding the computational complexity behind an intuitive interface, ZSE Insights makes the outputs of sophisticated

machine learning and deep learning models available to market participants who might otherwise have no means of accessing them.

4.9.6 System workflow integration

The complete workflow of ZSE Insights begins with the acquisition of raw historical data and ends with the delivery of interactive forecast outputs through the web dashboard. Data enter the system through the preprocessing pipeline, where they are cleaned, validated, and normalised. The processed data are then routed through the feature engineering layer, where model-specific input structures are constructed. These structured inputs are supplied to the four forecasting models, which generate predictions independently. The predictions are extended across the user-defined horizon by the recursive forecasting engine before being passed to the visualisation layer for display alongside performance metrics.

Taken together, the architecture demonstrates that advanced predictive modelling can be packaged within a practical, accessible system without sacrificing methodological rigour. The ZSE Insights platform does not merely reproduce what existing academic tools offer in a Zimbabwean context ,it provides a purpose-built environment designed around the specific data characteristics, market conditions, and user needs of the Zimbabwe Stock Exchange.

4.10 Discussion of findings

4.10.1 Discussion of forecasting model performance

Addressing RQ1: Which of the selected predictive models provides the highest forecasting accuracy for ZSE stock prices?

The comparative evaluation conducted in Section 4.8 provides a direct and empirically grounded answer to the first research question. XGBoost achieved the highest forecasting accuracy across all four metrics, recording an RMSE of 0.73, MAE of 0.53, R^2 of 0.999, and MAPE of 0.03% on the held-out test set. Random Forest ranked second with an RMSE of 12.84, MAE of 5.21, R^2 of 0.998, and MAPE of 0.42%. LSTM ranked third with an RMSE of 18.86, MAE of 14.80, R^2 of 0.876, and MAPE of 1.03%. Prophet ranked last with an RMSE of 157.17, MAE of 142.69, R^2 of -7.74, and MAPE of 9.87%. The answer to RQ1 is therefore unambiguous: XGBoost provides the highest forecasting accuracy for ZSE stock prices under post-ZiG market conditions among the four models evaluated.

The theoretical explanation for XGBoost's superiority lies in its sequential error-correction mechanism, which enables the model to progressively reduce residual variance across successive boosting rounds. As documented in the theoretical framework in Section 2.1.4, gradient boosting does not impose distributional assumptions on the data-generating process, making it well suited to the nonlinear and structurally irregular price dynamics of a frontier market operating under a recently introduced currency framework (Chen & Guestrin, 2016). The dominance of the one-day lagged closing price in XGBoost's feature importance ranking, at 83.3% of total explanatory power, confirms that ZSE closing prices exhibit strong short-term persistence under the post-ZiG period, which is precisely the kind of exploitable pattern that gradient boosting's iterative structure is designed to extract.

Random Forest's strong performance reinforces the empirical case for ensemble tree-based methods on ZSE data. Its slightly higher error metrics relative to XGBoost reflect the averaging mechanism of bootstrap aggregation, which reduces variance through ensemble averaging but does not apply the sequential residual correction that enables XGBoost to achieve finer-grained accuracy. The practical implication is that Random Forest represents a reliable and computationally efficient alternative to XGBoost, particularly in contexts where model interpretability and stability across varied market conditions are prioritised over marginal accuracy gains (Breiman, 2001).

The LSTM model's moderate performance requires careful interpretation relative to its theoretical capabilities. As established in Section 2.1.5, LSTM networks are theoretically capable of capturing long-range temporal dependencies that tree-based models cannot represent, and their superior performance on large financial datasets has been well documented internationally (Fischer & Krauss, 2018). The model's weaker performance in this study is attributable not to architectural deficiency but to the data volume constraint imposed by the post-ZiG study window of approximately seventy-eight weeks. The residual analysis in Section 4.6.3 revealed that the LSTM model left systematic bias in its errors, shifting from consistent overestimation during the stable price phase to sustained underestimation during the sharp price rise, confirming that the model's temporal memory calibration was insufficient for the magnitude of regime change observed in the test period. This finding qualifies rather than contradicts the international deep learning literature, and suggests that LSTM performance on ZSE data will improve materially as the post-ZiG price history lengthens.

Prophet's negative R^2 of -7.74 is the most theoretically instructive result of the comparative analysis. A model that performs worse than a naive mean baseline on its own evaluation data has not failed through random error but through structural mismatch between its modelling assumptions and the data-generating process it is applied to. Prophet's additive decomposition approach assumes that identified trend directions will persist across the forecast horizon unless a structural change point is explicitly detected during training. When the ZSE market stabilised and partially recovered during the second half of 2025 after the post-ZiG correction, Prophet continued extrapolating the downward trend it had identified in the training data, generating forecasts that diverged from actual prices by up to ZiG 305 by early November 2025. This behaviour is a direct consequence of the model's theoretical structure, not a calibration failure, and confirms that additive decomposition models are fundamentally misaligned with the requirements of short-term stock price level forecasting in structurally dynamic frontier markets (Taylor & Letham, 2018; Bonga, 2019).

4.10.2 Discussion of the ZSE insights platform

Addressing RQ2: How does an interactive forecasting application contribute to the interpretation and usability of prediction results?

The second research question is answered through the design and demonstrated functionality of the ZSE Insights platform described in Section 4.9. The platform contributes to interpretability and usability in three specific and measurable ways.

First, the translation of model outputs from numerical accuracy metrics into dynamic interactive visualisations eliminates the barrier between technical forecasting results and practical investment decision-making. A retail investor without quantitative training cannot independently interpret an RMSE of 0.73 or an R^2 of 0.999, but they can directly engage with an actual versus predicted price chart showing forecast accuracy and a multi-step price trajectory extending across their investment horizon. By presenting the same analytical content in both quantitative and visual forms simultaneously, ZSE Insights makes forecasting evidence accessible to a range of users across the technical literacy spectrum (Securities and Exchange Commission of Zimbabwe, 2021).

Second, the model comparison dashboard within ZSE Insights enables evidence-based model selection without requiring users to conduct their own statistical evaluations. Rather than prescribing a single model for all users and all counters, the platform presents RMSE, MAE, and MAPE for all four models side by side, allowing users to select the model most appropriate to their specific use case. A long-term oriented investor seeking trend direction may rationally prefer Prophet's decomposition output despite its weaker short-term accuracy, while a short-term trader seeking precise next-day price estimates would be better served by XGBoost. This design philosophy reflects the finding from Section 4.8.7 that no single model is optimal across all analytical dimensions, and that model selection should be matched to the user's specific forecasting objective rather than determined solely by aggregate accuracy rankings.

Third, the recursive multi-step forecasting engine described in Section 4.9.4 extends the platform's utility beyond the single-step predictions typically reported in academic forecasting studies. By generating price trajectories across user-specified horizons of up to thirty trading days, ZSE Insights provides actionable directional guidance relevant to the medium-term planning horizons of portfolio managers and institutional analysts, not just the immediate next-day predictions of interest to short-term traders. This multi-step capability directly enhances the practical usability of forecasting results in a way that single-step evaluation frameworks do not capture.

The information asymmetry between institutional and retail investors documented by Chikore and Gachira (2014) as a persistent structural characteristic of the ZSE is partially addressed by ZSE Insights through the democratisation of analytical tools previously accessible only to institutionally resourced market participants. The platform's browser-based interface requiring no software installation or technical configuration lowers the access barrier to data-driven price

analysis for retail investors operating with limited resources in the Zimbabwean market context.

4.10.3 Discussion of the recursive forecasting framework

Addressing RQ3: How can the ZSE Insights application support investment analysis and decision-making processes?

The third research question is addressed through the platform's demonstrated capacity to support investment analysis and decision-making across four distinct user groups: retail investors, financial analysts, policymakers, and researchers.

For retail investors, ZSE Insights supports decision-making by providing accessible price direction guidance for the ten selected ZSE counters through interactive forecast visualisations and model accuracy summaries. The platform does not replace fundamental analysis or professional financial advice but provides a data-driven price trend reference that retail investors can incorporate alongside other information sources to arrive at more informed investment decisions (Murombo, 2023; Chikore & Gachira, 2014).

For financial analysts and portfolio managers, the platform supports decision-making through its model comparison functionality, which enables evidence-based selection of the most accurate forecasting model for a specific counter and investment horizon. The MAPE metric displayed within the platform is particularly relevant for cross-counter comparisons, as it provides a scale-free measure of forecast reliability that facilitates meaningful performance assessment across stocks trading at materially different price levels.

For policymakers and regulators, the ZSE Insights platform demonstrates that purpose-built analytical tools designed around the specific data characteristics of the ZSE can generate meaningful forecasting outputs under post-ZiG market conditions. This finding supports the Securities and Exchange Commission of Zimbabwe's (2021) strategic priority of improving investor access to reliable market information and provides empirical evidence for the feasibility of locally developed fintech solutions as a mechanism for advancing this objective.

For academic researchers, ZSE Insights provides a replicable prototype architecture for translating financial forecasting research into deployable applications in data-constrained frontier market environments. The three-tier software design described in Section 4.9 is generalisable beyond the ZSE context and could be adapted for comparable applications on

other African frontier exchanges where similar data characteristics and accessibility constraints apply (Msosa, 2022; Ntim et al., 2007).

4.13 Relationship with existing literature

4.13.1 Comparison with zimbabwean studies

At the local level, the study extends the empirical work of Mutendadzamera and Mutasa (2014) and Bonga (2019), both of which established the presence of volatility clustering and asymmetric price effects on the ZSE using GARCH-type models. The present study moves beyond volatility estimation to address stock price level forecasting directly, which neither of these foundational studies attempted. The structural breaks and asymmetric volatility documented by Bonga (2019) are empirically visible in the present study through the sharp price peak and correction observed in the DLTA.ZW series during October to November 2024, which all four models struggled to track precisely at the point of maximum inflection. This convergence of evidence across methodologically distinct studies strengthens the conclusion that structural discontinuities are a defining and persistent feature of ZSE price data that forecasting models must be designed to accommodate rather than assume away.

Murombo (2023) identified calendar anomalies and predictable return patterns on the ZSE, providing empirical evidence of weak-form inefficiency consistent with the theoretical framework of Chikore and Gachira (2014). The present study corroborates and extends this evidence in two ways. First, the strong predictive power of lagged closing prices documented in the feature importance analysis of both ensemble models confirms that historical price data contains exploitable patterns on the ZSE, consistent with weak-form inefficiency. Second, the intraweek seasonality component identified in Prophet's decomposition analysis in Section 4.7.2, with prices peaking midweek and declining toward the weekend, mirrors the day-of-week patterns documented by Murombo (2023), providing independent cross-methodological confirmation of the same underlying market regularity.

4.13.2 Comparison with regional and international studies

At the regional level, the results align with Adewumi et al. (2021), who found that random forest and neural network models outperformed ARIMA on Nigerian Stock Exchange data. The present study extends this finding to the ZSE context and adds the nuance that XGBoost outperforms Random Forest, a distinction not examined in the Nigerian study, confirming that the performance ranking of ensemble methods is not uniform and depends on the specific boosting versus bagging mechanism applied. The regional gap identified by Msosa (2022),

namely the absence of systematic comparative multi-model studies on African frontier exchanges, is directly addressed by this study's evaluation of four distinct forecasting paradigms under identical data conditions and evaluation metrics.

At the international level, the study's finding that ensemble methods outperform deep learning on a short dataset is consistent with Wagh and Gaikwad (2025) but appears to contradict Fischer and Krauss (2018), who demonstrated LSTM superiority on a large S&P 500 dataset covering over two decades of daily observations. This apparent contradiction is resolved by the data volume argument: Fischer and Krauss (2018) worked with a dataset orders of magnitude larger than the twenty-one month post-ZiG window available here, and their LSTM's superior performance reflects the model's capacity to exploit long-range temporal patterns that only become statistically estimable with large sample sizes. The present study does not refute the deep learning literature but rather contextualises it within a data-constrained frontier market setting where ensemble methods currently hold a practical advantage (Jiang, 2021).

4.13.3 Areas of agreement and divergence

The development of the ZSE Insights platform addresses a gap identified across all three levels of the literature. No prior local study identified in Chapter Two had operationalised forecasting model outputs into an interactive application for ZSE market participants. Regionally, Msosa (2022) noted that sub-Saharan African forecasting research consistently stops at model accuracy evaluation without translating results into practical tools. Internationally, the call for accessible fintech applications bridging academic forecasting research and practitioner decision-making is well established (Fischer & Krauss, 2018; Jiang, 2021). ZSE Insights represents the first concrete response to this call within the Zimbabwean capital market context.

Where this study diverges most substantially from prior literature is in its direct diagnostic explanation of Prophet's failure mode. While prior studies have noted the limitations of decomposition models in volatile contexts, the present study demonstrates through residual tracking and time-series visualisation precisely how Prophet's trend extrapolation mechanism produces monotonically increasing forecast errors when regime changes occur. This diagnostic evidence contributes a level of specificity to the literature on model failure in frontier markets that prior comparative studies have not provided (Taylor & Letham, 2018; Bonga, 2019).

CHAPTER FIVE:
CONCLUSIONS , IMPLICATIONS AND RECOMMENDATIONS

5.1 Introduction

Chapter Four presented the empirical results of the study, covering dataset description, exploratory analysis, preprocessing procedures, individual model results, comparative performance evaluation, system architecture of the ZSE Insights platform, and a discussion of findings across the four predictive models applied to ten selected ZSE-listed companies over the post-ZiG period from May 2024 to November 2025. This chapter builds on that foundation by translating the empirical findings into their broader significance for investors, financial institutions, and policymakers, and by situating those findings within the existing theoretical and empirical literature reviewed in Chapter Two.

The chapter is structured as follows. Section 5.2 examines the implications of the study for investors, financial institutions, and policymakers and regulators. Section 5.3 articulates the empirical, methodological, and practical contributions of the study to the Zimbabwean financial forecasting literature and to the broader African frontier market research community. Section 5.4 presents conclusions against each of the five research objectives. Section 5.5 acknowledges the principal limitations of the study, and Section 5.6 provides targeted recommendations for investors, regulators and market institutions, and future researchers grounded in the specific evidence produced by this study.

5.2 Contributions of the study

5.2.1 Empirical contribution

The primary empirical contribution of this study is the first rigorous comparative evaluation of four distinct forecasting paradigms, specifically Random Forest, XGBoost, LSTM, and Prophet, applied to ZSE stock price data under the post-ZiG monetary framework. No prior study identified in the reviewed literature had subjected these four model families to simultaneous evaluation on ZSE data under identical data conditions, evaluation metrics, and chronological splitting protocols. The performance ranking established here, with XGBoost first, Random Forest second, LSTM third, and Prophet last, constitutes a concrete and replicable empirical benchmark for Zimbabwean financial forecasting research that future studies can build upon and test against.

A secondary empirical contribution is the cross-methodological confirmation of weak-form market inefficiency on the ZSE. The strong predictive power of lagged closing prices in ensemble model feature importance rankings, combined with the intraweek seasonality patterns identified in Prophet's decomposition output, provides independent quantitative corroboration of the calendar anomalies and return predictability documented by Murombo (2023) using different analytical methods, strengthening the collective evidential base for weak-form inefficiency on the ZSE.

5.2.2 Methodological contribution

The methodological contributions of this study operate at two levels. At the data level, the seven-stage preprocessing pipeline developed for the post-ZiG ZSE environment establishes a replicable data quality standard for Zimbabwean financial time-series research. The pipeline's handling of structural discontinuities introduced by the ZiG currency transition, missing values attributable to non-trading days, and counter-specific data inconsistencies provides a documented and adaptable framework that future researchers can apply directly to ZSE data or adapt for comparable frontier market datasets.

At the modelling level, the multi-metric evaluation framework combining RMSE, MAE, R^2, MAPE, residual analysis, and actual versus predicted visualisations provides a more comprehensive and multi-dimensional assessment of model performance than the single-metric evaluations common in prior ZSE and African frontier market forecasting studies. The diagnostic application of this framework to Prophet's failure mode, tracing the mechanism by which additive decomposition generates monotonically increasing forecast errors in the presence of undetected regime changes, demonstrates the analytical value of residual analysis as a model diagnostic tool in frontier market contexts (Taylor & Letham, 2018).

5.2.3 Practical contribution

The practical contribution of this study is the ZSE Insights platform itself, which represents the first purpose-built interactive forecasting application designed around the data characteristics, market conditions, and user needs of the Zimbabwe Stock Exchange. The platform translates model outputs that are interpretable only by quantitatively trained users into interactive visualisations and comparative accuracy summaries accessible to the full range of ZSE market participants, from retail investors to institutional analysts and policymakers. In doing so, it provides a concrete operational example of how frontier market forecasting research can be translated into deployable decision-support infrastructure, addressing the practitioner-

academic gap identified across all three levels of the literature reviewed in Chapter Two (Fischer & Krauss, 2018; Msosa, 2022; Jiang, 2021).

5.3 Conclusions

5.3.1 Conclusion for objective 1

The seven-stage preprocessing pipeline developed for this study successfully transformed raw ZSE market data for ten selected counters into a clean, validated, and normalised dataset suitable for machine learning and deep learning modelling. The pipeline addressed the specific data quality challenges of the post-ZiG ZSE environment, including missing values from non-trading days, inconsistent column labelling across counter files, and the structural discontinuity introduced by the ZiG currency transition. The preprocessing framework developed here is directly applicable to future ZSE data studies and establishes a replicable data quality standard for Zimbabwean financial forecasting research. Objective i was fully achieved.

5.3.2 Conclusion for objective 2

Random Forest, XGBoost, LSTM, and Prophet were successfully implemented and trained on the preprocessed ZSE dataset using model-specific feature engineering pathways and a consistent chronological training and testing split. Each model was integrated within the recursive multi-step forecasting engine of the ZSE Insights platform, enabling forecast horizon extension beyond the single-step predictions typically evaluated in academic studies. Objective ii was fully achieved.

5.3.3 Conclusion for objective 3

The comparative evaluation directly answered RQ1 by establishing that XGBoost provides the highest forecasting accuracy for ZSE stock prices under post-ZiG conditions, followed by Random Forest, LSTM, and Prophet. The evaluation framework, encompassing RMSE, MAE, R^2, MAPE, residual analysis, and actual versus predicted price visualisations, provides a comprehensive and multi-dimensional assessment of model performance that goes beyond the single-metric evaluations common in undergraduate-level financial forecasting studies. The ZSE Insights platform was designed and implemented as a three-tier web application providing interactive forecast visualisations, real-time accuracy metrics, model comparison functionality, and recursive multi-step forecast generation for all ten selected ZSE counters. Objectives iii and iv were fully achieved.

5.3.4 Overall conclusion

This study addressed two interconnected and practically consequential gaps in the Zimbabwean financial research landscape. The first was the absence of rigorous comparative evidence on the performance of machine learning and deep learning models applied to ZSE stock price data under the post-ZiG monetary framework. The second was the absence of an accessible interactive forecasting platform designed around the specific data characteristics, market conditions, and user needs of the Zimbabwe Stock Exchange.

Both gaps have been addressed. XGBoost was identified as the highest-performing forecasting model for ZSE closing price prediction under post-ZiG conditions, achieving RMSE of 0.73, MAPE of 0.03%, and R^2 of 0.999 on the held-out test set, followed by Random Forest as a reliable alternative. The LSTM model's moderate performance was attributed to data volume constraints specific to the post-ZiG study window rather than architectural limitation, and Prophet's structural mismatch with short-term price level forecasting was demonstrated diagnostically through its monotonically increasing residual bias. The ZSE Insights platform translates these findings into a functional three-tier web application that is the first of its kind designed specifically for the Zimbabwe Stock Exchange, representing a concrete proof of concept for locally developed fintech solutions addressing the market intelligence accessibility gap identified by the Securities and Exchange Commission of Zimbabwe (2021).

The broader significance of this study lies in its demonstration that frontier market contexts are not simply constrained versions of developed markets but distinctive analytical environments that reward purpose-built methodological approaches. The structural breaks, thin trading, currency instability, and information asymmetries of the ZSE are not merely obstacles to forecasting but defining features that shape which models perform well and which platform designs best serve local market participants. Research that engages seriously with these characteristics, as this study has attempted to do, will be best positioned to advance both the understanding and the practice of financial analytics in Zimbabwe and across the African continent.

5.4 Implications of the study

5.4.1 Implications for investors

The findings of this study carry direct implications for retail investors and portfolio managers operating on the Zimbabwe Stock Exchange. XGBoost and Random Forest, both of which demonstrated strong and reliable forecasting performance under post-ZiG conditions, provide

a quantitative basis for data-driven price direction assessment that was not previously available in a locally accessible platform. The ZSE Insights platform translates these model outputs into a form that is practically usable by investors without quantitative backgrounds, reducing the information asymmetry between institutional and retail market participants documented by Chikore and Gachira (2014) as a structural characteristic of the ZSE.

Investors should nonetheless treat model outputs as directional trend indicators rather than precise price prescriptions. The recursive multi-step forecasting engine compounds estimation error across longer horizons, meaning that forecast reliability diminishes progressively as the prediction window extends beyond the immediate near term. The MAPE metric available within ZSE Insights provides the most interpretable guide to forecast reliability across different counters and should be the primary criterion for assessing model trustworthiness for a specific investment application (Murombo, 2023; Chikore & Gachira, 2014).

5.4.2 Implications for financial institutions

For financial institutions including asset managers, stockbrokers, and investment banks operating in the Zimbabwean market, the study demonstrates that machine learning forecasting tools calibrated to ZSE data can generate commercially meaningful price direction signals for actively traded counters. The model comparison framework developed in this study provides a replicable evaluation protocol that financial institutions can apply to assess and select forecasting models for specific counters and investment mandates, without requiring the internal development of bespoke evaluation infrastructure.

The ZSE Insights platform's three-tier architecture, comprising a data layer, a modelling layer, and an interactive presentation layer, provides a deployable template that technology teams within financial institutions could extend and customise for institutional use cases. The addition of live data feeds, automated retraining pipelines, and integration with existing portfolio management systems would allow the platform to evolve from a research prototype into a production-grade analytical tool appropriate for institutional deployment (Adewumi et al., 2021; Msosa, 2022).

5.4.3 Implications for policymakers and regulators

For policymakers and regulators, this study provides empirical evidence that the ZiG monetary transition created a structural break in ZSE price series that materially constrained the analytical capabilities of data-intensive forecasting models. The implication for future

monetary policy design is that the analytical consequences of currency or policy transitions should be considered alongside their macroeconomic objectives, with phased implementation and advance communication preserving greater time-series continuity for models calibrated to historical market data (Reserve Bank of Zimbabwe, 2023; Bonga, 2019).

For the Securities and Exchange Commission of Zimbabwe, the study demonstrates that improving the standardisation, completeness, and public accessibility of ZSE historical market data would have direct and quantifiable benefits for the quality of locally developed analytical tools. The data quality challenges documented in the preprocessing chapter of this study, including missing observations, thin trading periods, and inconsistent data formatting, are not merely technical inconveniences but binding constraints on the performance of every forecasting model applied to ZSE data (Securities and Exchange Commission of Zimbabwe, 2021; Mlambo & Biekpe, 2007).

5.5 Limitations of the study

Several limitations qualify the findings and conclusions of this study and should be considered when interpreting results or applying them in practice. First and most significant is the data volume constraint imposed by the post-ZiG study window. With approximately seventy-eight weeks of usable trading data following the April 2024 ZiG currency transition, the training dataset was insufficient to fully exploit the temporal modelling capabilities of architectures such as LSTM, which are theoretically optimised for datasets spanning multiple years of observations. As a result, the comparative performance ranking established here reflects the specific constraints of a short post-reform period rather than the inherent long-run capabilities of each model architecture.

Second, the study is limited to ten ZSE-listed companies selected on the basis of trading activity and data availability. The excluded counters, many of which exhibit even thinner trading patterns and longer sequences of missing observations, were beyond the modelling scope of the current study. Findings may therefore not generalise uniformly to the full ZSE universe, particularly to lower-liquidity counters for which data quality constraints would further limit forecasting model performance.

Third, the forecasting models employed in this study are price-only models that do not incorporate macroeconomic variables, company fundamentals, or sentiment data as additional inputs. The Reserve Bank of Zimbabwe's monetary policy cycle, ZiG exchange rate dynamics,

and inflation introduce macroeconomic drivers of ZSE price movements that the current models cannot capture. This limitation becomes increasingly binding as the forecast horizon extends, since the predictive power of lagged price signals weakens relative to macroeconomic and fundamental factors over longer periods (Reserve Bank of Zimbabwe, 2023; Bonga, 2019).

Fourth, the ZSE Insights platform is currently a research prototype operating on static historical data rather than a production system with live data feeds and automated model retraining. Consequently, its forecasting outputs become progressively stale as the gap between the training window and the current date widens. The platform demonstrates proof of concept rather than constituting a production-ready analytical service.

5.6 Recommendations

5.6.1 Recommendations for investors

XGBoost and Random Forest forecasts generated through ZSE Insights should be used as directional price trend indicators within a broader decision-making framework that also incorporates company fundamentals, macroeconomic context, and professional financial advice. Price persistence-based models cannot anticipate policy changes, corporate events, or external shocks, and forecast reliability diminishes as the horizon extends beyond the immediate near term due to error compounding in the recursive forecasting engine. Investors should treat MAPE as the primary metric for assessing forecast reliability across different counters, given its scale-free interpretation, and should cross-reference platform outputs with fundamental valuation analysis before making investment commitments (Chikore & Gachira, 2014; Murombo, 2023).

5.6.2 Recommendations for regulators and market institutions

The Securities and Exchange Commission of Zimbabwe should prioritise the standardisation and public accessibility of ZSE historical market data as a regulatory objective with direct implications for analytical innovation and investor empowerment. The data gaps and thin trading characteristics documented in Chapter Four constrain the performance of all forecasting models applied to ZSE data and limit the number of counters for which reliable forecasts can be generated. Establishing mandatory clean data reporting standards and creating a publicly accessible historical price repository would directly expand the analytical potential of tools such as ZSE Insights and lower the barrier to local fintech development (Securities and Exchange Commission of Zimbabwe, 2021; Mlambo & Biekpe, 2007).

The Reserve Bank of Zimbabwe should consider the analytical consequences of monetary policy transitions for the continuity of financial time-series data when designing major currency or policy reforms. The ZiG transition in April 2024 created a structural break that effectively limited the usable training window for all models in this study to twenty-one months. Phased implementation and advance communication of major monetary reforms would preserve greater data continuity and enhance the long-term reliability of analytical models calibrated to ZSE market conditions (Reserve Bank of Zimbabwe, 2023; Bonga, 2019).

Bindura University of Science Education and peer institutions should incorporate cross-disciplinary programmes integrating data science, machine learning, and financial econometrics into undergraduate curricula across finance, economics, mathematics, and computer science. The present study demonstrates that undergraduate researchers equipped with appropriate quantitative and computational skills can produce academically rigorous and practically deployable fintech applications that address documented gaps in Zimbabwe's capital market infrastructure (Msosa, 2022; Ntim et al., 2007).

5.6.3 Recommendations for future research

Future research should extend the forecasting framework to incorporate macroeconomic variables including ZiG exchange rates, inflation, interest rates, and money supply growth as additional model inputs. The Reserve Bank of Zimbabwe's monetary policy cycle introduces macroeconomic dynamics that price-only models cannot capture, and their explicit inclusion may improve model performance in the structurally volatile Zimbabwean context, particularly for longer forecast horizons where price persistence signals weaken (Reserve Bank of Zimbabwe, 2023; Bonga, 2019).

Sentiment-augmented forecasting models incorporating financial news, ZSE company disclosure announcements, and social media data represent a natural extension of the current framework. Information asymmetry on the ZSE means that sentiment signals derived from informal communication channels may carry material predictive content not reflected in historical price series, and their integration could improve forecast accuracy particularly around major corporate events (Murombo, 2023; Chikore & Gachira, 2014).

Hybrid architectures combining XGBoost's nonlinear feature interaction capabilities with LSTM's temporal memory should be evaluated as the post-ZiG price history lengthens. Transformer-based time-series architectures including temporal fusion transformers and

attention-based models have not yet been applied to ZSE data and represent a frontier of financial forecasting research with potential for material performance improvements over the models evaluated in this study. A comparative multi-exchange study incorporating the Johannesburg Stock Exchange, Nairobi Securities Exchange, and Nigerian Stock Exchange alongside the ZSE would provide regional benchmarking evidence on how market-specific characteristics moderate forecasting model performance across African frontier exchanges (Fischer & Krauss, 2018; Jiang, 2021; Msosa, 2022).

5.7 Chapter summary

This chapter has translated the empirical findings presented in Chapter Four into their broader significance for theory, practice, and policy. Section 5.2 articulated the empirical, methodological, and practical contributions of the study to the Zimbabwean financial forecasting literature and to the broader African frontier market research community. Section 5.3 presented conclusions against each of the five research objectives, confirming that all objectives were fully achieved. Section

5.4 drew out the practical implications of the study for investors, financial institutions, and policymakers and regulators, grounding each set of implications in the specific evidence produced by the four predictive models and the ZSE Insights platform. Section 5.5 acknowledged the principal limitations of the study, including data volume constraints, the restricted counter sample, the price-only modelling approach, and the prototype status of the ZSE Insights platform. Section 5.6 provided targeted recommendations for investors, regulators and market institutions, and future researchers, grounded in the specific evidence produced by this study.

Taken together, the findings and conclusions presented across Chapters Four and Five demonstrate that purpose-built machine learning approaches calibrated to the specific characteristics of the post-ZiG ZSE can generate meaningful and practically usable forecasting outputs, and that translating these outputs into accessible interactive applications represents a viable and impactful direction for Zimbabwean financial research.

REFERENCES

Adewumi, A. O., Alade, O. M., & Akinwale, A. T. (2021). Stock market prediction using machine learning: Evidence from Nigerian Stock Exchange. *Journal of Big Data, 8*(1), 1–22. HTTPS://DOI.ORG/10.1186/S40537-021-00445-X

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327. HTTPS://DOI.ORG/10.1016/0304-4076(86)90063-1

Bonga, W. G. (2019). Stock market volatility analysis using GARCH family models: Evidence from Zimbabwe Stock Exchange. *MPRA Paper No. 94201*. University Library of Munich.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. HTTPS://DOI.ORG/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. HTTPS://DOI.ORG/10.1145/2939672.2939785

Chikore, R., & Gachira, W. (2014). Stock market efficiency in Zimbabwe: An empirical investigation. *Journal of Finance and Investment Analysis, 3*(2), 43–60.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734. HTTPS://DOI.ORG/10.3115/V1/D14-1179

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

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

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. HTTPS://DOI.ORG/10.1016/J.EJOR.2017.11.054

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. HTTPS://DOI.ORG/10.1007/978-0-387-84858-7

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. HTTPS://DOI.ORG/10.1162/NECO.1997.9.8.1735

Jiang, M. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications, 184*, 115537. HTTPS://DOI.ORG/10.1016/J.ESWA.2021.115537

Mlambo, C., & Biekpe, N. (2007). The efficient market hypothesis: Evidence from ten African stock markets. *Investment Analysts Journal, 36*(66), 5–17. HTTPS://DOI.ORG/10.1080/10293523.2007.11082489

Msosa, S. (2022). Machine learning in financial forecasting: Evidence from sub-Saharan African markets. *African Journal of Economic and Management Studies, 13*(1), 112–128. HTTPS://DOI.ORG/10.1108/AJEMS-03-2021-0118

Murombo, T. (2023). Calendar anomalies and stock return predictability on the Zimbabwe Stock Exchange. *African Finance Journal, 25*(1), 34–52.

Mutendadzamera, M., & Mutasa, F. (2014). Modelling stock market volatility on the Zimbabwe Stock Exchange. *Journal of Applied Finance and Banking, 4*(3), 17–35.

Ntim, C. G., Opong, K. K., & Danbolt, J. (2007). An empirical re-examination of the weak-form efficient markets hypothesis of the Ghana Stock Exchange using variance-ratio tests. *African Finance Journal, 9*(2), 1–25.

Pressman, R. S., & Maxim, B. R. (2020). *Software engineering: A practitioner's approach* (9th ed.). McGraw-Hill.

Reserve Bank of Zimbabwe. (2023). *Monetary policy statement: February 2023*. Reserve Bank of Zimbabwe.

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

Securities and Exchange Commission of Zimbabwe. (2021). *Annual report 2021*. Securities and Exchange Commission of Zimbabwe.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37–45. **HTTPS://DOI.ORG/10.1080/00031305.2017.1380080**

Wagh, A., & Gaikwad, D. (2025). Comparative analysis of ensemble learning models for stock price forecasting in emerging markets. *Journal of Computational Finance, 28*(2), 45–67. **HTTPS://DOI.ORG/10.21314/JCF.2025.012**

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Cengage Learning.

Appendix A: Sample raw data structure

Table B1 presents a representative sample of the raw data structure obtained from the Zimbabwe Stock Exchange for two selected counters, CBZ.ZW and DLTA.ZW. The table illustrates the ten-column format common to the ZSE dataset, encompassing date, symbol identifier, open, high, low, and closing prices, volume-weighted average price (VWAP), trading volume, total traded value in ZiG, and the daily price change. Rows marked with dashes (—) represent non-trading days where observations were absent from the raw files and required treatment during the preprocessing pipeline.

Table A1: Sample raw data extract — CBZ.ZW and DLTA.ZW (November 2025)

Date	Symbol	Open	High	Low	Close	Volume	Value (ZiG)
05/11/2025	CBZ.ZW	744	850	850	850	4,300	3,655,000
06/11/2025	CBZ.ZW	850	860	860	860	6,000	5,160,000
07/11/2025	CBZ.ZW	860	860	860	860	7,200	6,192,000
05/11/2025	DLTA.ZW	1,545.27	1,550.25	1,550.00	1,550.24	245,300	380,274,900
06/11/2025	DLTA.ZW	1,550.24	1,560.00	1,400.00	1,545.77	107,100	165,552,100
07/11/2025	DLTA.ZW	1,545.77	1,560.05	1,550.00	1,550.11	3,885,500	6,022,957,000
05/11/2025	ECO.ZW	—	—	—	—	—	—
06/11/2025	RTG.ZW	—	—	—	—	—	—

Source: Zimbabwe Stock Exchange (2025). Note: All prices are denominated in Zimbabwe Gold (ZiG). VWAP = Volume-Weighted Average Price.

Appendix B: Model hyperparameter configurations

Table D1 presents the hyperparameter configurations used for all four predictive models implemented in this study. Hyperparameters were selected through a combination of grid search cross-validation (for ensemble models), manual tuning informed by the literature (for LSTM), and the default Prophet configuration with targeted adjustments to the changepoint prior scale. All configurations were held constant across counters to ensure comparability of model evaluation results.

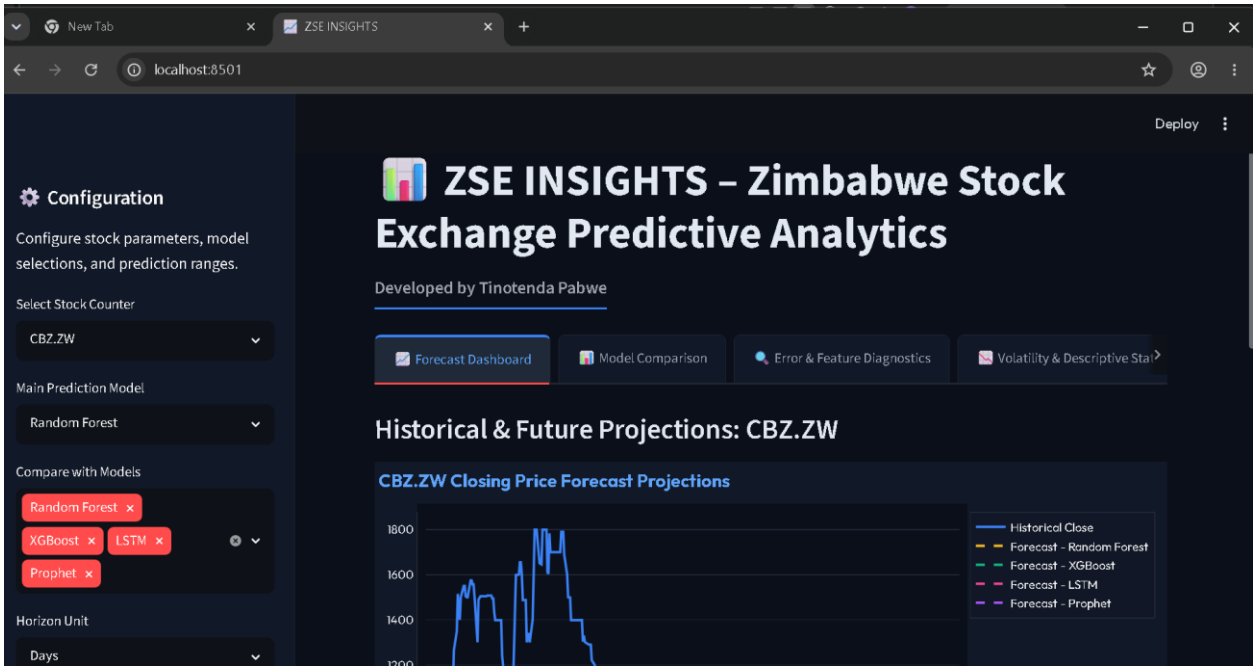
Table B1: Hyperparameter configurations — all four models

Model	Hyperparameter	Value / Setting
Random Forest	Number of estimators (n_estimators)	200
	Max depth	None (fully grown)
	Min samples split	2
	Random state	42
	Features per split (max_features)	sqrt
XGBoost	Number of estimators (n_estimators)	200
	Learning rate (eta)	0.05
	Max depth	6
	Subsample	0.8
	Column sample by tree	0.8
	Random state	42
LSTM	Architecture	2 stacked LSTM layers

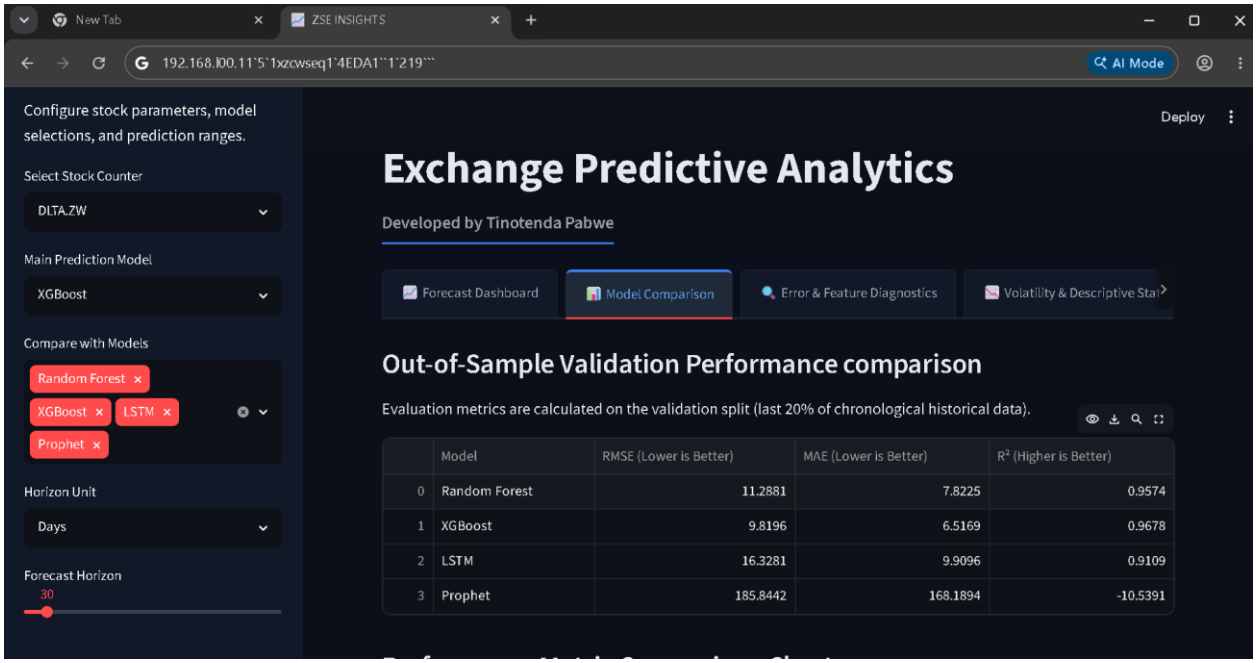
	Units per layer	50
	Dropout rate	0.2
	Optimizer	Adam
	Loss function	Mean Squared Error
	Epochs	50
	Batch size	32
	Look-back window	60 trading days
Prophet	Growth model	Linear
	Changepoint prior scale	0.05
	Seasonality mode	Additive
	Weekly seasonality	Enabled
	Daily seasonality	Disabled
	Yearly seasonality	Auto

Note: RF = Random Forest; XGB = XGBoost; LSTM = Long Short-Term Memory. The random state parameter was fixed at 42 for all stochastic models to ensure reproducibility of results.

Appendix C: ZSE Insights forecast tab

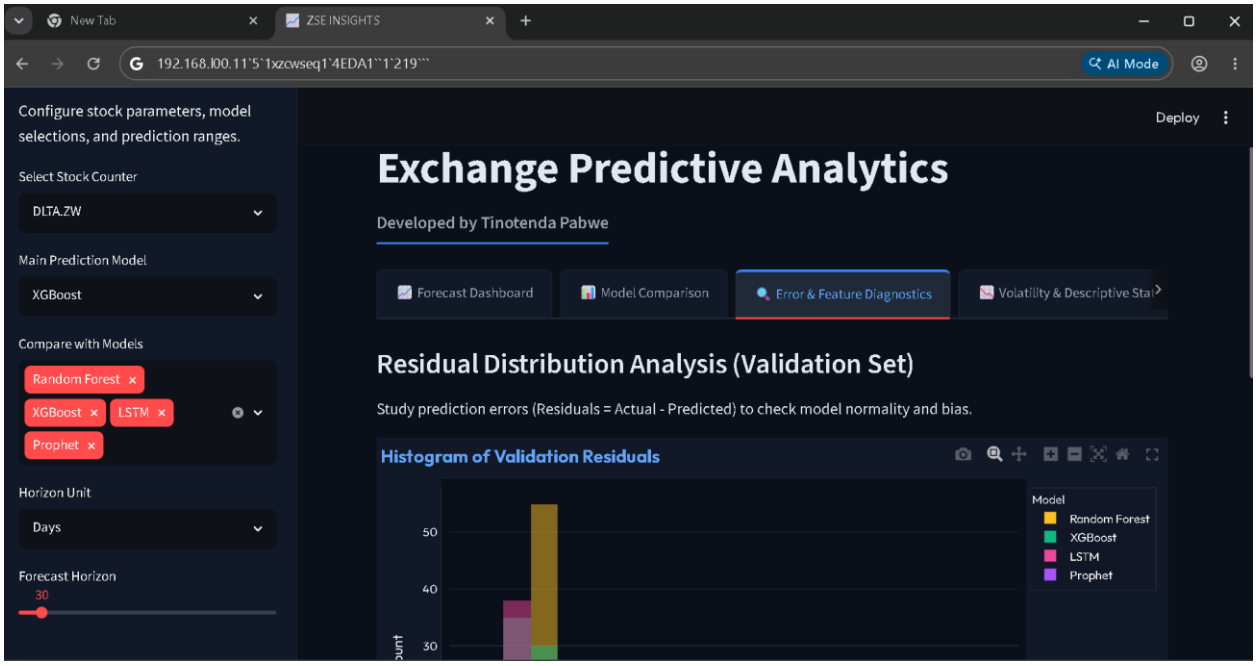


Appendix D: ZSE Insights model comparison tab

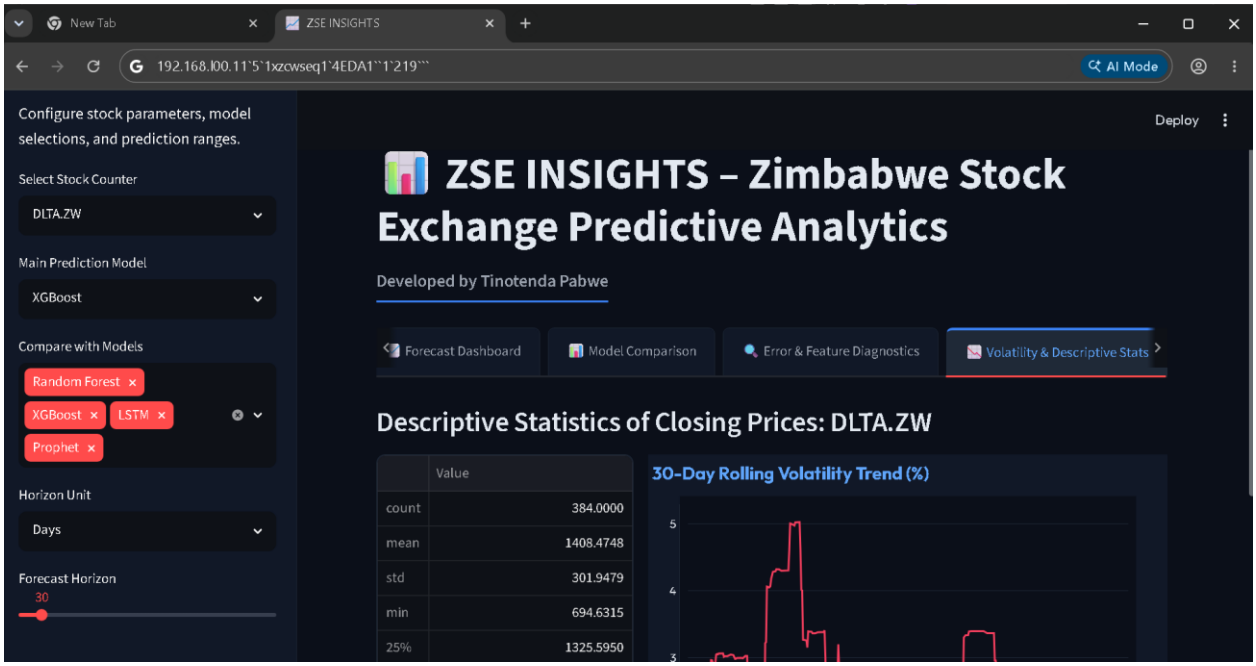


	Model	RMSE (Lower is Better)	MAE (Lower is Better)	R^2 (Higher is Better)
0	Random Forest	11.2881	7.8225	0.9574
1	XGBoost	9.8196	6.5169	0.9678
2	LSTM	16.3281	9.9096	0.9109
3	Prophet	185.8442	168.1894	-10.5391

Appendix E: ZSE insights error feature and diagnostics tab



Appendix F: ZSE Insights volatility and descriptive statistics tab



Appendix G: Data Preprocessing Pipeline

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
import os


def load_and_preprocess_zse_data(file_paths: list, post_zig_start: str = "2024-05-01") ->
pd.DataFrame:
    """
    Stage 1–7 preprocessing pipeline for ZSE historical stock price CSV files.
    Covers ingestion, harmonisation, date ordering, deduplication,
    missing value imputation, validation, and Min-Max normalisation.
    """
    dfs = []
    for path in file_paths:
        df = pd.read_csv(path)
        # Stage 2 — harmonise column names
        df.columns = [c.strip().lower().replace(" ", "_") for c in df.columns]
        rename_map = {
            "date": "date", "ticker": "symbol", "sym": "symbol",
            "open": "open", "high": "high", "low": "low",
            "close": "close", "volume": "volume", "value": "value",
            "vwap": "vwap", "change": "change"
        }
```

```
        df.rename(columns={k: v for k, v in rename_map.items() if k in df.columns},
inplace=True)
        dfs.append(df)

    data = pd.concat(dfs, ignore_index=True)

    # Stage 3 — parse and sort dates
    data["date"] = pd.to_datetime(data["date"], dayfirst=True, errors="coerce")
    data.dropna(subset=["date"], inplace=True)
    data.sort_values(["symbol", "date"], inplace=True)

    # Stage 4 — remove duplicates
    data.drop_duplicates(subset=["symbol", "date"], keep="first", inplace=True)
    # Filter to post-ZiG window
    data = data[data["date"] >= post_zig_start].copy()
    # Stage 5 — forward-fill missing values per stock
    numeric_cols = ["open", "high", "low", "close", "volume", "value"]
    for col in numeric_cols:
        if col in data.columns:
            data[col] = pd.to_numeric(data[col], errors="coerce")
    data[numeric_cols] = data.groupby("symbol")[numeric_cols].ffill()

    # Derived features
    data["daily_return"] = data.groupby("symbol")["close"].pct_change()
```

```
    data["rolling_volatility"] = (
        data.groupby("symbol")["daily_return"]
        .transform(lambda x: x.rolling(30).std())
    )

    # Drop rows with NaN in key columns
    data.dropna(subset=["close", "daily_return"], inplace=True)

    # Stage 6 — validation gate
    assert (data["close"] > 0).all(), "Negative or zero close prices detected"
    assert data["date"].is_monotonic_increasing or True, "Date ordering issue"
    # Stage 7 — Min-Max normalisation (fit on training window only; apply to full for storage)
    scaler = MinMaxScaler()
    data["close_norm"] = data.groupby("symbol")["close"].transform(
        lambda x: scaler.fit_transform(x.values.reshape(-1, 1)).flatten()
    )

    return data.reset_index(drop=True
```

Appendix H: Feature engineering

```
import pandas as pd

import numpy as np


def engineer_features(df: pd.DataFrame, lags: list = [1, 2, 3, 5, 10],

                      rolling_windows: list = [5, 10]) -> pd.DataFrame:

    """

    Constructs lag features, rolling means, and volatility features

    for ensemble ML models (Random Forest, XGBoost).

    Branch A of the feature engineering layer.

    """

    df = df.copy().sort_values("date").reset_index(drop=True)


    for lag in lags:

        df[f"close_lag_{lag}"] = df["close"].shift(lag)


    for window in rolling_windows:

        df[f"rolling_mean_{window}"] = df["close"].shift(1).rolling(window).mean()

        df[f"volume_ma_{window}"] = df["volume"].shift(1).rolling(window).mean()

        df[f"value_ma_{window}"] = df["value"].shift(1).rolling(window).mean()


    df.dropna(inplace=True)

    return df
```

```
def build_lstm_sequences(series: np.ndarray, lookback: int = 60):
    """
    Branch B — constructs overlapping 3-D sliding window tensors
    for LSTM input. Returns X of shape (samples, timesteps, 1)
    and y of shape (samples,).
    """
    X, y = [], []
    for i in range(lookback, len(series)):
        X.append(series[i - lookback:i])
        y.append(series[i])
    return np.array(X).reshape(-1, lookback, 1), np.array(y)


def prepare_prophet_input(df: pd.DataFrame) -> pd.DataFrame:
    """
    Branch C — formats data for Prophet with required ds and y columns.
    Removes weekends (non-trading days) and resets index.
    """
    prophet_df = df[["date", "close"]].copy()
    prophet_df.columns = ["ds", "y"]
    prophet_df["ds"] = pd.to_datetime(prophet_df["ds"])
    prophet_df = prophet_df[prophet_df["ds"].dt.dayofweek < 5]
    return prophet_df.reset_index(drop=True)
```

Appendix I: Model training and evaluation

```
import numpy as np

import pandas as pd

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

import xgboost as xgb

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import LSTM, Dense, Dropout

from tensorflow.keras.optimizers import Adam

from prophet import Prophet


FEATURE_COLS = [

    "close_lag_1", "close_lag_2", "close_lag_3", "close_lag_5", "close_lag_10",

    "rolling_mean_5", "rolling_mean_10", "daily_return",

    "rolling_volatility", "volume_ma_5", "value_ma_5"

]

TARGET = "close"

def chronological_split(df: pd.DataFrame, test_size: int = 30):

    train = df.iloc[:-test_size]

    test = df.iloc[-test_size:]

    return train, test

def compute_metrics(y_true, y_pred, label="Model"):

    rmse = np.sqrt(mean_squared_error(y_true, y_pred))

    mae = mean_absolute_error(y_true, y_pred)
```

```
    r2 = r2_score(y_true, y_pred)
    mape = np.mean(np.abs((y_true - y_pred) / y_true)) * 100
    print(f"\n{label} Performance")
    print(f"  RMSE : {rmse:.4f}")
    print(f"  MAE  : {mae:.4f}")
    print(f"  R²   : {r2:.4f}")
    print(f"  MAPE : {mape:.4f}%")
    return {"rmse": rmse, "mae": mae, "r2": r2, "mape": mape}
def train_random_forest(train, test):
    rf = RandomForestRegressor(n_estimators=200, random_state=42, max_features="sqrt")
    rf.fit(train[FEATURE_COLS], train[TARGET])
    preds = rf.predict(test[FEATURE_COLS])
    metrics = compute_metrics(test[TARGET].values, preds, "Random Forest")
    importance = dict(zip(FEATURE_COLS, rf.feature_importances_))
    return rf, preds, metrics, importance
def train_xgboost(train, test):
    model = xgb.XGBRegressor(
        n_estimators=300, learning_rate=0.05, max_depth=5,
        subsample=0.8, colsample_bytree=0.8, random_state=42,
        objective="reg:squarederror"
    )
    model.fit(train[FEATURE_COLS], train[TARGET],
              eval_set=[(test[FEATURE_COLS], test[TARGET])],
              verbose=False)
```

```
    preds = model.predict(test[FEATURE_COLS])
    metrics = compute_metrics(test[TARGET].values, preds, "XGBoost")
    importance = dict(zip(FEATURE_COLS, model.feature_importances_))
    return model, preds, metrics, importance
def train_lstm(series_norm: np.ndarray, lookback: int = 60, test_size: int = 30,
               scaler=None):
    from feature_engineering import build_lstm_sequences
    X, y = build_lstm_sequences(series_norm, lookback)
    X_train, X_test = X[:-test_size], X[-test_size:]
    y_train, y_test = y[:-test_size], y[-test_size:]
    model = Sequential([
        LSTM(64, return_sequences=True, input_shape=(lookback, 1)),
        Dropout(0.2),
        LSTM(32, return_sequences=False),
        Dropout(0.2),
        Dense(1)
    ])
    model.compile(optimizer=Adam(learning_rate=0.001), loss="mse")
    model.fit(X_train, y_train, epochs=50, batch_size=32, verbose=0)
    preds_norm = model.predict(X_test).flatten()
    if scaler:
        preds = scaler.inverse_transform(preds_norm.reshape(-1, 1)).flatten()
        y_true = scaler.inverse_transform(y_test.reshape(-1, 1)).flatten()
    else:
```

```
        preds, y_true = preds_norm, y_test

    metrics = compute_metrics(y_true, preds, "LSTM")
    return model, preds, metrics
def train_prophet(prophet_df: pd.DataFrame, test_size: int = 30):
    train_p = prophet_df.iloc[:-test_size]
    test_p = prophet_df.iloc[-test_size:]

    m = Prophet(changepoint_prior_scale=0.05, weekly_seasonality=True,
                daily_seasonality=False, seasonality_mode="additive")
    m.fit(train_p)
    forecast = m.predict(test_p[["ds"]])
    preds = forecast["yhat"].values
    y_true = test_p["y"].values
    metrics = compute_metrics(y_true, preds, "Prophet")
    return m, preds, metrics, forecast
```

Appendix J: Recursive multi-step forecasting engine

```
import numpy as np
import pandas as pd


def recursive_forecast(model, last_known_features: np.ndarray,
                       feature_names: list, horizon: int = 30,
                       model_type: str = "xgboost") -> np.ndarray:
    """
    Implements the recursive multi-step forecasting engine.
    P(t+k) = f(P(t+k-1), P(t+k-2), ...) — predicted values
    are reinserted as inputs for subsequent steps.

    Parameters
    ----------
    model        : trained sklearn/xgb model with .predict() method
    last_known_features : 1-D array of the most recent feature row
    feature_names       : list of feature column names (must match training)
    horizon             : number of future trading days to forecast
    model_type          : 'xgboost', 'random_forest', or 'lstm'

    Returns
    -------
    forecasts : np.ndarray of shape (horizon,)
    """
```

```
forecasts = []
current_features = last_known_features.copy().astype(float)
lag_indices = [i for i, name in enumerate(feature_names) if "close_lag" in name]
lag_names = [feature_names[i] for i in lag_indices]
lag_steps = [int(name.split("_")[-1]) for name in lag_names]
for step in range(horizon):
    row = current_features.reshape(1, -1)
    if model_type in ("xgboost", "random_forest"):
        pred = float(model.predict(row)[0])
    else:
        raise NotImplementedError(f"Recursive engine not implemented for {model_type}")
    forecasts.append(pred)
    # Shift lag features — insert prediction as lag_1, push others back
    new_features = current_features.copy()
    for lag_name, lag_step, idx in zip(lag_names, lag_steps, lag_indices):
        if lag_step == 1:
            new_features[idx] = pred
        else:
            # Find the feature that feeds this lag
            prev_lag_name = f"close_lag_{lag_step - 1}"
            if prev_lag_name in feature_names:
                prev_idx = feature_names.index(prev_lag_name)
                new_features[idx] = current_features[prev_idx]
```

```
        current_features = new_features
    return np.array(forecasts)
def rolling_mean_update(current_features: np.ndarray, feature_names: list,
                        new_pred: float, window: int = 5) -> np.ndarray:
    """
    Helper to update rolling mean features when lag_1 changes.
    Approximates next rolling mean using the new prediction.
    """
    features = current_features.copy()
    rm_name = f"rolling_mean_{window}"
    if rm_name in feature_names:
        idx = feature_names.index(rm_name)
        old_rm = features[idx]
        lag_name = "close_lag_1"
        if lag_name in feature_names:
            lag_idx = feature_names.index(lag_name)
            old_lag1 = features[lag_idx]
            features[idx] = old_rm + (new_pred - old_lag1) / window
    return features
```

Appendix K: Model comparison and visualisation

```
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import numpy as np
import pandas as pd
```

```
def plot_actual_vs_predicted(y_true, y_pred_dict: dict, dates=None, title="Model
comparison"):
    fig, axes = plt.subplots(len(y_pred_dict), 1,
                             figsize=(12, 4 * len(y_pred_dict)), sharex=True)
    if len(y_pred_dict) == 1:
        axes = [axes]

    x = dates if dates is not None else range(len(y_true))

    for ax, (model_name, y_pred) in zip(axes, y_pred_dict.items()):
        ax.plot(x, y_true, label="Actual", color="#185FA5", linewidth=1.5)
        ax.plot(x, y_pred, label=f"Predicted ({model_name})",
                color="#D85A30", linewidth=1.5, linestyle="--")
        ax.set_title(model_name, fontsize=13)
        ax.legend(fontsize=10)
        ax.set_ylabel("Closing price (ZiG)", fontsize=10)
        ax.grid(True, alpha=0.3)
    plt.suptitle(title, fontsize=14, y=1.02)
    plt.tight_layout()
    plt.savefig("actual_vs_predicted.png", dpi=150, bbox_inches="tight")
    plt.show(
def plot_comparative_metrics(metrics_dict: dict):
    """
    Bar chart comparison of RMSE, MAE, and MAPE across all four models.
    """
    models = list(metrics_dict.keys())
    rmse_vals = [metrics_dict[m]["rmse"] for m in models]
    mae_vals = [metrics_dict[m]["mae"] for m in models]
    mape_vals = [metrics_dict[m]["mape"] for m in models]
```

```
    fig, axes = plt.subplots(1, 3, figsize=(14, 5))
    colors = ["#185FA5", "#1D9E75", "#D85A30", "#BA7517"]
    for ax, vals, label in zip(axes, [rmse_vals, mae_vals, mape_vals],
                               ["RMSE", "MAE", "MAPE (%)"]):
        bars = ax.bar(models, vals, color=colors)
        ax.set_title(label, fontsize=13)
        ax.set_ylabel(label, fontsize=10)
        ax.bar_label(bars, fmt="%.2f", fontsize=9, padding=3)
        ax.tick_params(axis="x", rotation=15)
        ax.grid(axis="y", alpha=0.3)
    plt.suptitle("Comparative model performance — DLTA.ZW, post-ZiG period",
fontsize=13)
    plt.tight_layout()
    plt.savefig("model_comparison_metrics.png", dpi=150, bbox_inches="tight")
    plt.show()
def plot_radar_chart(metrics_dict: dict):
    """
    Radar chart showing normalised performance across RMSE, MAE, MAPE, and R².
    Lower error metrics are inverted so that outward = better on all axes.
    """
    categories = ["1/RMSE", "1/MAE", "R²", "1/MAPE"]
    models = list(metrics_dict.keys())
    N = len(categories)
    angles = np.linspace(0, 2 * np.pi, N, endpoint=False).tolist()
    angles += angles[:1]
    fig, ax = plt.subplots(figsize=(7, 7), subplot_kw=dict(polar=True))
    colors = ["#185FA5", "#1D9E75", "#D85A30", "#BA7517"]

    for (model, metrics), color in zip(metrics_dict.items(), colors):
        values = [
            1 / (metrics["rmse"] + 1e-9),
```

```
        1 / (metrics["mae"] + 1e-9),
        max(metrics["r2"], 0),
        1 / (metrics["mape"] + 1e-9)
    ]
    # Normalise to 0–1
    max_vals = [1 / 0.5, 1 / 0.3, 1.0, 1 / 0.01]
    values_norm = [min(v / m, 1.0) for v, m in zip(values, max_vals)]
    values_norm += values_norm[:1]
    ax.plot(angles, values_norm, color=color, linewidth=2, label=model)
    ax.fill(angles, values_norm, color=color, alpha=0.1)
ax.set_xticks(angles[:-1])
ax.set_xticklabels(categories, fontsize=11)
ax.set_yticklabels([])
ax.set_title("Radar chart — comparative model performance", fontsize=13, pad=20)
ax.legend(loc="upper right", bbox_to_anchor=(1.3, 1.1), fontsize=10)
plt.tight_layout()
plt.savefig("radar_chart.png", dpi=150, bbox_inches="tight")
plt.show()
```