

BINDURA UNIVERSITY OF SCIENCE EDUCATION
FACULTY OF SCIENCE AND ENGINEERING
DEPARTMENT OF STATISTICS AND MATHEMATICS



**THE IMPACT OF CLIMATE CHANGE AND FOOD PRODUCTION ON FOOD
PRODUCTS EXPORT EARNINGS IN ZIMBABWE: A MACHINE LEARNING
APPROACH**

BY
TINASHE C CHIPAYI
B224718B

**TO BE SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR BSC.HONOURS IN STATISTICS AND FINANCIAL MATHEMATICS
PROJECT**

SUPERVISOR: Ms P. hlupo
2026

AUTHORSHIP DECLARATION STATEMENT

Title of the Thesis: The impact of climate change and food production on food products export earnings in Zimbabwe

Author: Tinashe Chipayi

Program: Bachelor of Science Honors Degree in Statistics and Financial Mathematics.

I, the undersigned author of the above-mentioned thesis, hereby declare that:

1.This thesis is my original work and has been prepared by me in accordance with the institution's requirements.

2.All sources, data, and references used in this thesis have been acknowledged and cited appropriately.

3.This thesis has not been submitted elsewhere for any degree or diploma.

4.I have obtained all necessary permissions for the inclusion of third-party content where applicable.

I affirm that this declaration is made with full integrity and in compliance with the institution's policies and academic practice.

Author's Signature:



Date:

05/06/2026

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.

Tinashe C Chipayi
B224718B



Signature

05/06/2026

Date

Certified by
Ms. P. HLUPO
Supervisor



Signature

08/06/26

Date

Dr Magodora
Chairperson



Signature

14/08/2026

Date

DEDICATION

I dedicate this research to my parents, Oliver Chipayi and Runyararo Nyahunzvi, as well as my dear siblings. Your unwavering faith in my abilities, endless support, and encouragement have been instrumental in helping me reach this milestone. Through every challenge and triumph, you have remained a constant source of strength, inspiration, and motivation. Thank you for your love, sacrifices, and for always standing by my side. This accomplishment reflects not only my efforts but also the invaluable role you have played in my journey.

ACKNOWLEDGEMENTS

I would like to express my profound appreciation to everyone who contributed to the successful completion of this research. Special thanks go to my supervisor, Ms. P. Hlupo, for her invaluable guidance, constructive criticism, encouragement, and continuous support throughout the study. Her expertise and commitment greatly enhanced the quality of this work. I am equally grateful to the lecturers, professionals, and subject-matter experts whose knowledge, mentorship, and insights enriched my academic journey and provided a strong foundation for this research. My sincere appreciation is also extended to the Department of Statistics and Mathematics at Bindura University of Science Education for fostering an environment that supported learning, research, and academic growth.

I would also like to acknowledge my family for their unwavering encouragement, patience, and support throughout this journey. Their understanding and sacrifices provided the motivation needed to persevere and achieve this goal.

Above all, I give thanks to God Almighty for His grace, strength, wisdom, and guidance. His blessings sustained me throughout this academic endeavour and enabled me to overcome the challenges encountered along the way.

ABSTRACT

Climate change is a threat to the agriculture-based economy in Zimbabwe due to its effect on rainfall, droughts, temperature changes, and natural disasters, which hinder production and export earnings. This research aims at exploring the effect of climate change and food production on food product export earnings in Zimbabwe using annual data for the period 1967–2025. Two machine learning models: extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM) are used to estimate the relationship between climate factors, food production, and export earnings. Data was collected from ZIMSTAT, the Food and Agriculture Organization (FAO), and the World Bank while using lagged values and three-year rolling averages in the process of feature engineering to enhance accuracy. Results show that there is a very strong relationship between food production and export earnings ($r = 0.68$). Climate variability affects trade through its effect on food production, hence its effect on earnings from exports. Model predictions are highly accurate, XGBoost outperforms the LightGBM, giving an r^2 of 0.981 and a MAPE of 3.46%. Historical values of export earnings and food production are among the top features used in predicting future export performance.

TABLE OF CONTENTS

Table of Contents

AUTHORSHOP DECLARATION STATEMENT	ii
APPROVAL FORM.....	iii
DEDICATION.....	iv
ACKNOWLEDGEMENTS	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	xii
LIST OF FIGURES	xiii
LIST OF SYMBOLS	xiv
LIST OF ACRONYMS.....	xv
CHAPTER 1: INTRODUCTION.....	1
1.0 Introduction.....	1
1.1 Background of the Study	1
1.2 Statement of the problem.....	3
1.3 Objectives of the Study.....	3
1.4 Research Questions.....	3
1.5 Scope of the Study	3
1.6 Significance of the Study.....	4
1.7 Assumptions of the Study	4
1.8 Limitations of the Study.....	5
1.9 Key terms Definition	5
1.10 conclusion	6
CHAPTER TWO: LITERATURE REVIEW	8
2.0 Introduction.....	8
2.1 Theoretical Literature.....	8
2.1.1 Climate Change Theory	8
2.1.2 Agricultural Production Theory (Cobb–Douglas Production Function).....	8

2.1.3 International Trade and Export Earnings Theory.....	9
2.1.4 Statistical Learning Theory.....	9
2.1.5 XGBoost Theory (Extreme Gradient Boosting)	10
2.1.6 Light Gradient Boosting Machine (LightGBM) Theory.....	10
2.2 Empirical Literature	11
2.2.1 Climate Change and Agricultural Productivity	11
2.2.2 Climate Change and Export Earnings.....	12
2.2.3 Machine Learning Models in Agricultural Forecasting	12
2.3 Literature Synthesis and Research Gap	13
2.4 Conceptual Framework.....	14
2.5 Chapter Conclusion.....	15
CHAPTER 3: RESEARCH METHODOLOGY	16
3.0 Introduction.....	16
3.1 Research Philosophy and Design.....	16
3.1.1 Research Philosophy	16
3.1.2 Research Design.....	17
3.2 Data Sources and Collection Methods	17
3.2.1 Data Sources	17
3.2.2 Data Collection Methods	18
3.3 Target Population and Sampling Techniques	19
3.3.1 Target population	19
3.3.2 Sampling Techniques	20
3.4 Description of Variables and Expected Relationships	20
3.4.1 Dependent Variable.....	20
3.4.2 Independent Variables.....	21
3.4.3 Expected Relationships.....	22
3.5 Analytical Model and Justification	23
3.5.1 XGBoost Model Specification.....	23

3.5.2 LightGBM Model Specification	24
3.6 Machine Learning Models Pre-Test and Diagnostic Test	25
3.6.1 Data Normalization (Min-Max Scaling)	25
3.6.2 Evaluation of Multicollinearity (Variance Inflation Factor – VIF)	26
3.6.3 Detection of Outliers (Z-Score Approach)	26
3.6.4 Feature Importance Analysis	27
3.6.5 Data Splitting and Validation	27
3.7 Model Fit and Validation Methods	27
3.7.1 Root Mean Squared Error (RMSE)	27
3.7.2 Mean Absolute Error (MAE)	28
3.7.3 Mean Absolute Percentage Error (MAPE)	28
3.7.4 Coefficient of Determination (R ²)	28
3.7.5 Residuals analysis	29
3.8 Ethical Considerations	30
3.9 Conclusion of the chapter	31
CHAPTER 4: DATA PRESENTATION, ANALYSIS, AND DISCUSSION	32
4.0 Introduction	32
4.1 Descriptive statistics	32
4.3 Trend analysis of climate and food variables	34
4.4.0 Diagnostic tests/ pre-tests	35
4.4.1 Correlation	35
4.4.2 Autocorrelation check- Durbin-Watson	36
4.4.3 Multicollinearity Check (Variance Inflation Factor – VIF)	36
4.4.4 Data validation (outlier detection and missing values)	37
4.4.5 Normality check-Shapiro-Wilk test	38
4.4.6 Data normalization	38
4.4.7 Data transformation	39
4.4.8 Tackling multicollinearity in machine learning model	41

4.4.9 Feature engineering.....	41
4.4.10 Data splitting.....	42
4.5.0 Analytical models.....	43
4.5.1 XGBoost development.....	43
4.5.2 Predicting using XGBoost	44
4.5.3 Feature importance.....	46
4.5.4 Performance matrices results	47
4.5.5 Residual analysis test	48
4.6.0 LightGBM development	48
4.6.1 Predicting using LightGBM Model	49
4.6.2 Feature importance.....	50
4.6.3 Performance matrices results	52
4.6.4 Residuals analysis	52
4.7.0 Comparisons of the models' performance.....	53
4.7.1 Comparison of XGBoost and LightGBM Feature Importance	53
4.7.3 Performance comparisons	58
4.8.1 Discussion of the findings.....	59
4.9 Chapter Summary	63
CHAPTER 5: SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS..	65
5.0 Introduction.....	65
5.1 Summary of Findings.....	65
5.2 Conclusions.....	65
5.3 Recommendations.....	66
5.3.1 To Policymakers.....	66
5.3.2 To Investors and Agricultural Stakeholders	66
5.3.3 To Model Developers and Researchers.....	67
5.4 Limitations of the Study.....	67
5.5 Areas for Further Research	67

5.6 Chapter Conclusion.....	68
References.....	69
Appendix A: Transformed Data (FP & FE only transformed)	72
Appendix B: Feature Engineering.....	73
Appendix C: Syntax of R Codes used to run the models.....	74
C.2 : R Codes.....	74

LIST OF TABLES

Table 4.1 Descriptive statistics.....	35
Table 4.2 Stationary test results.....	36
Table 4.3 Autocorrelation results.....	39
Table 4.4 Multicollinearity results.....	40
Table 4.5 Missing values results.....	41
Table 4.6 Normality test results.....	41
Table 4.7 XGBoost parameters.....	47
Table 4.8 Skewness test results.....	43
Table 4.9 XGBoost predicted vs Actual values.....	48
Table 4.10 XGBoost performance matrices results.....	51
Table 4.11 LightGBM parameters.....	52
Table 4.12 LightGBM predicted vs actual values.....	53
Table 4.13 LightGBM performance matrices results.....	55
Table 4.14 Features importance rank results.....	57
Table 4.15 XGBoost and LightGBM residuals comparisons.....	61
Table 4.16 XGBoost and LightGBM predicted vs actual values.....	62
Table 4.17 XGBoost and LightGBM matrices comparisons.....	62

LIST OF FIGURES

Figure 1.1 Line graph of food products export value (1967-2020).....	3
Figure 2.1 Conceptual framework.....	15
Figure 4.1 Trend analysis of climate and food variables.....	37
Figure 4.2 Correlation Matrix.....	38
Figure 4.3 Transformed data skewness.....	43
Figure 4.4 Transformed data normality.....	44
Figure 4.5 XGBoost predicted vs actual over time plot.....	48
Figure 4.6 XGBoost feature importance.....	49
Figure 4.7 XGBoost residual plots.....	51
Figure 4.8 LightGBM predicted vs actual over time plot.....	53
Figure 4.9 LightGBM feature importance.....	54
Figure 4.10 LightGBM residual plots.....	56
Figure 4.11 Residual histogram.....	58
Figure 4.12 Residual plot.....	59
Figure 4.13 Residual QQ plot.....	59
Figure 4.14 Residual over time.....	60
Figure 4.15 XGBoost and LightGBM predicted vs Actual plot.....	62

LIST OF SYMBOLS

Ω	Omega
Σ	Sum of
μ	Mean/average
σ	Standard deviation
$\in$	Member/element of
$\hat{Y}$	Yhat -Pridicted

LIST OF ACRONYMS

1. XGBoost Extreme Gradient Boosting
2. LightGBM.....Light Gradient Boosting Machine
3. PR.....Precipitation
4. HU.....Humidity
5. TPM.....Temperature
6. FP.....Food Production
7. FE.....Food Exports
8. RMSE Root Mean Squared Error
9. MAE.....Mean Absolute Error
10. MAPE.....Mean Absolute Percentage Error
11. VIF.....Variance Inflation Factor

CHAPTER 1: INTRODUCTION

1.0 Introduction

Climate change has emerged as one of the greatest challenges in the current times, threatening economies and livelihoods throughout the world. Crop farming is highly sensitive to climate change effects, which according to experts adversely affect the productivity of agriculture through rise in temperature, uneven rainfall patterns and occurrence of natural calamities (Molua 2002, Isik & Devadoss 2006, IPCC 2007, WB 2010). As a nation dependent heavily on agriculture, climate changes are likely to affect food production and export from Zimbabwe. This has also had adverse effects on the productivity and earnings from agricultural exports. The research paper is divided into five chapters. Chapter one gives background information and objectives of the study while chapter two focuses on the review of previous literature. Chapter three details the methodology used in the study while chapter four highlights findings from the analysis then lastly chapter five reviews the results and presents the recommendations and conclusions.

1.1 Background of the Study

Zimbabwe's economy relies heavily on agriculture, which contributes about 10-15% to the GDP, provides jobs for roughly 70% of the population, and supplies around 40% of export earnings (World Bank 2022). The country depends on tobacco, maize, cotton, and horticultural products not just for trade but also for food security. Over the last couple of decades, though, climate change has become a big problem, hurting the agriculture sector and cutting into export performance (Dzvimbo et al. 2017).

The United Nations Framework Convention on Climate Change recognizes the risk that dropping food production poses (Reid et al. 2007). Zimbabwe has seen more extreme weather – like unpredictable rains, severe droughts, and floods that come from unusual rainfall patterns and warmer temps (FAO 2020). These changes disrupt growing seasons, reduce crop yields and increase post-harvest losses (Blanc 2011). For instance, during the El Niño-spurred drought in 2015-2016, maize production took a massive hit, pushing the country to import food and shrink its agricultural exports (ZimStat 2016). Plus, based on the Zimbabwe Meteorological Services Department (MSD), yearly rain levels are becoming less dependable, making it harder to cultivate crops whether that be dry farming or irrigated stuff (MSD 2021).

During 2018-2019, inconsistent rain resulted in significantly lower yields than usual. Government relief measures, such as Pfumvudza/Intwasa initiatives and food assistance, haven't solved everything because the country still lacks solid predictive tech. Traditional methods, like linear regression, fall short when it comes to tackling complicated, non-linear issues, thus researchers should explore sophisticated solutions, like machine learning.

Various scientific areas teaming up could offer key insights into handling these challenges via studies linking climate and crop yields, devising optimal growing practices, and setting up reliable models (Ahmed et al. 2013, Jones et al. 2017). There's no denying the strong ties among agri-business, feeding systems, and commerce; all emphasize how urgent it is to collect real-world evidence that supports steady agricultural and eco-political blueprints in Zimbabwe. Climate variability directly affects the economy due to agriculture's climate sensitivity, which whittles down both local and traded goods as well as economic stability (Vink 2012).

With reduced produce, export totals fall and income drops too. Zimbabwe faces further issues due to low coping capabilities, subpar facilities, and frail defences against climate shocks. Meeting Vision 2030 hinges on boosting food productivity (Vink 2012). It's crucial to figure out how to navigate and soften these climate effects, matching up with objectives under Vision 2030 and UN SDGs for ending world hunger, protecting our planet, and spurring wealth. Investigating links 'twixt weather measurements and crop prosperity gives backing for smart interventions and plans that guard farming from shifting climatic behaviours.

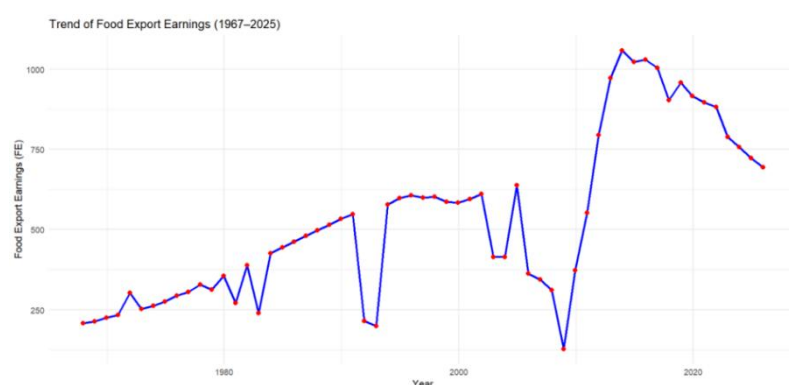


Figure 1.1-line graph of food products export value (1967-2025)

Source: World Bank data showing the variability of food production and food products export values.

1.2 Statement of the problem

In Zimbabwe, food production swings wildly and exports keep dropping due to unreliable weather patterns. While there's plenty of research on climate change and farming, not much uses machine learning to look at how climate affects both crop yields and export income. Since previous researchers haven't done a thorough study, policy makers struggle to create good strategies for adapting and fixing these issues. That's why this study aims to see if machine learning can better predict how climate change impacts both production and income from food exports.

1.3 Objectives of the Study

1. To assess the relationship between climate variability and food production levels and export earnings.
2. To develop a machine learning model with more reliable performance accuracy (XGBoost and LightGBM (Light Gradient Boosting Machine)).

1.4 Research Questions

1. How strongly is climate and food production variability relating with changes in food product export value in Zimbabwe?
2. Which machine learning model provides more reliable performance accuracy for food products export earnings based on climate and food production data?

1.5 Scope of the Study

This study looks at how climate change and food production impact export earnings in Zimbabwe. The time frame goes from 1967 to 2025, a period of big weather swings with droughts, cyclones, and changing rain patterns. It zeroes in on how rainfall and temperature trends have affected important crops like maize, tobacco, and cotton – major players in both feeding Zimbabwe and earning exports.

The research covers all agricultural areas in Zimbabwe, such as Mashonaland, Manicaland, and Midland's provinces that's where most of the crops are grown. The study sticks to a quantitative analysis using second-hand data on climate, food production stats, and earnings data from government and international sources (like ZIMSTAT, FAO, and World Bank).

Moreover, it tests two machine learning models to see how well each explains the link between climate variables and economic results. Still, it doesn't deeply examine other factors like land reform, politics, or global market pricing, even though those factors matter too.

1.6 Significance of the Study

To the academic community

This study provides an opportunity to apply statistical and financial mathematics tools to a real-world socio-economic issue climate change and its economic implications. It enhances analytical and data interpretation as well as modelling skills, especially in using machine learning methods to address pressing national challenges.

To Researchers

The research contributes to the limited localized literature on the nexus between climate change, food production, and export earnings in Zimbabwe. It offers a foundation for future empirical studies by applying comparative methodologies, encouraging interdisciplinary research that combines climate science, agriculture, economics, and data science.

To Policymakers

Findings from this study can inform evidence-based policy formulation and implementation. By quantifying the effects of climate variability on agriculture and export performance, policymakers are better equipped to design adaptive strategies, prioritize climate-resilient investments, and allocate resources effectively to enhance food security and economic stability.

To Investors and Agricultural Stakeholders

The insights into climate-related risks and trends help investors and agribusinesses assess exposure and make informed decisions about agricultural investments. Understanding how climate affects productivity and export potential can guide decisions on crop selection, insurance products, and long-term investment strategies in Zimbabwe's agriculture sector.

1.7 Assumptions of the Study

Firstly, assuming that the data from sources like ZIMSTAT, FAO, and World Bank – covering rainfall, temperature, food production, and export earnings – accurately reflect Zimbabwe's real conditions between 1967 and 2025.

Next, the study assumes that while other factors do play roles, food production and climate factors such as rainfall and temperature chiefly affect food product export earnings.

Additionally, the relationship between climate patterns and agricultural output is thought to stay fairly consistent throughout the studied timeframe. This allows for useful modelling and comparisons using statistics and machine learning methods.

Finally, assuming the chosen analytical methods – XGBoost and LightGBM – can effectively find and show the relationships between our variables.

1.8 Limitations of the Study

The study uses secondary data from government and international sources, which may contain gaps or inaccuracies. Furthermore, getting high resolution or current data is not always accessible, and this lack can make the analysis off the mark.

While it looks at food production and weather patterns, the research does not extensively examine things like land reform or the cost of inputs, which also impact exports. Plus, the models used—XGBoost and LightGBM—are only as good as the data they get. So, if the data are imperfect, these models might miss out on important stuff affecting farming.

There's also the issue of generalizing the findings for Zimbabwe as a whole. Since climate effects can differ a lot between regions and for different crops, the study's broader conclusions might not accurately reflect local conditions.

Lastly, the time frame—from 1967 to 2025—may be too narrow. It probably won't pick up on super-long climate trends and could restrict how well can apply the study's results to times outside that window.

Even with these limitations, though, the study gives us a good look at how climate, farming, and trade are linked.

1.9 Key terms Definition

Climate Change

According to John T. Houghton (2015), climate change refers to long-term alterations in the average weather conditions of a region or the Earth as a whole, including changes in temperature, precipitation, wind patterns, and other aspects of the climate system, resulting from both natural processes and human activities.

Food Production

David Pimentel and Marcia Pimentel (2008) define food production as the production of crops and livestock through the utilization of land, water, labour, and other agricultural inputs to provide food for human consumption and economic sustenance.

Export Earnings

Dominick Salvatore (2019) defines export earnings as the monetary value of goods and services sold by a country to the rest of the world, generating foreign exchange receipts that support economic development and international trade.

XGBoost (Extreme Gradient Boosting)

A rapid approach for getting the desired output, even a professional one, is possible through using an enhanced gradient boosting framework, such as gravity boosting, which is accompanied by increased performance along with overfitting prevention, and dealing efficiently with missing data (Chen & Guestrin 2016).

LightGBM

It refers to a very fast and effective gradient boosting system created by Microsoft and based on decision trees which can perform classification, regression and ranking functions (Ke et al., 2017)

Gradient Boosting

Is a sequential model ensemble technique which involves building of subsequent models to reduce losses by adjusting the loss function of preceding models. This allows developing a strong classifier from weaker learners like decision trees (Ren, 2023).

Machine Learning

Refers to algorithms including LightGBM and XGBoost for forecasting, classification and recognizing patterns or a kind of AI in which machine learning algorithms and statistical models learn from data and then make decisions accordingly (Jordan & Mitchell 2015).

1.10 conclusion

This chapter has introduced the study on the impact of climate change and food production on food products export earnings in Zimbabwe. It outlined the background, highlighted the

problem being addressed, and stated the objectives and research questions guiding the investigation. The significance, scope, assumptions, and limitations were also discussed. This foundation sets the stage for the next chapter, which review existing literature on climate change, agriculture and export performance, providing theoretical and empirical context for the present study.

CHAPTER TWO: LITERATURE REVIEW

2.0 Introduction

This chapter reviews theoretical and empirical literature on the impact of climate change and food production on food products export earnings. The chapter begins by discussing theories related to climate change, agricultural production, international trade, and statistical learning. It further examines predictive modelling approaches, particularly Extreme Gradient Boosting (XGBoost) and LightGBM, which are increasingly used in agricultural forecasting studies. The chapter also reviews empirical studies on climate change, agricultural productivity, and export performance. Finally, the chapter synthesizes the reviewed literature, identifies research gaps, and presents the conceptual framework guiding the study.

2.1 Theoretical Literature

This section discusses the theories underpinning the study. The theories explain the relationship between climate change, agricultural productivity, export earnings, and predictive modelling techniques.

2.1.1 Climate Change Theory

The main idea here is the Climate Change Theory, which revolves around the Greenhouse Effect discovered by Joseph Fourier and built upon by Svante Arrhenius. Basically, greenhouse gases like CO₂ catch heat in our atmosphere, causing global warming and messing up the climate. Current research from places like the IPCC backs this up, saying that more greenhouse gases mean hotter temperatures, weird rain patterns, and lots of disasters like droughts and floods.

Farmers have it tough because their crops need specific temperatures, rain, and soil dampness. With climate swings, growing stuff gets harder, giving lower crop numbers and messed-up farming routines. That's not good for food supplies or money from selling extra produce. This ties into this study since it shows how changes in the weather impact what we can grow and thus affect our trade earnings.

2.1.2 Agricultural Production Theory (Cobb–Douglas Production Function)

Agricultural Production Theory shows us how things like land, labour, capital, and technology shape farming output. The Cobb-Douglas Production Function, created by Charles Cobb and Paul Douglas in 1928, describes output based on inputs and their elasticity.

In the field of agricultural economics, stuff like temperature and rainfall really affect farming output. These climate swings can actually cut how well inputs work and dent crop harvests. Then, Solow in 1956 and Schultz in 1964 updated the theory to include technology advances and human expertise, stressing the impact of both innovation and strong institutions on productivity.

This theory matters for our study since climate change works as a shock that changes how farms produce. Not only could this mean less food, but nations relying on farm exports might see their money drop too.

2.1.3 International Trade and Export Earnings Theory

Ricardo's Comparative Advantage, explained in 1817, says countries should focus on making what they make best compared to others. For agriculture, big factors are good weather and how much they can produce.

Climate change could hurt farming, causing worse droughts, floods, and shifts in rain. This harms countries' best crops for export since lower food harvests mean less to sell and earn cash from abroad. And that affects the economy.

The Prebisch–Singer Theory, created in 1950, tells us relying on basic goods like crops leaves you open to big price swings. Adding climate effects into that mix, and it gets tougher for countries to perform well in exports, dragging down their money.

So, this theory helps us see why climate changes are extra scary, linking drops in farming output straight to the hit in cash flow from selling those goods.

2.1.4 Statistical Learning Theory

Statistical Learning Theory, created by Vapnik and Chervonenkis from 1964 to 1974, forms the backbone of modern machine-learning models. This theory centres on minimizing empirical risk, boosting generalization, and balancing bias and variance.

In agriculture and export data, we often see nonlinear patterns, seasonality, and high variability due to climate swings. Because of this, statistical learning techniques shine, as they can spot intricate links between climatic factors and economic impacts.

Time series analysis is key in statistical learning; it zeroes in on past observations to find trends, patterns, and projections for the future. As Aras and Kocakoc point out in 2016, this type of analysis revolves around forecasting, modelling, and describing how data evolves over time. Plus, Tealab added in 2018 that time series forecasting lets us predict future observations based on old data. For this research, statistical learning theory sets the stage for building models to project how climate changes affect food product export revenues and food production shifts.

2.1.5 XGBoost Theory (Extreme Gradient Boosting)

Chen and Guestrin developed XGBoost in 2016 as a step up from regular gradient boosting methods. This model makes prediction trees one after another, correcting mistakes from the earlier ones. Plus, it uses something called regularization to cut down on overfitting and boost performance.

Another advantage of XGBoost? It does great at managing nonlinear connections and how variables interact. What's more, it shows how important each predictor is in influencing forecast results.

For agriculture forecasting, stuff like rain, humidity, and temperature are usually tangled up in tricky ways. So, XGBoost is perfect for our purposes because it can nail those nonlinear links between climate, food output, and export money.

2.1.6 Light Gradient Boosting Machine (LightGBM) Theory

LightGBM, introduced by Ke et al. in 2017, makes gradient boosting decision trees way faster and scalable. It differs from old boosting methods thanks to neat tricks like GOSS and EFB, which cut down on costs and keep accuracy intact.

In this system, big parts include the decision tree bit, a clever leaf-wise growing approach, and nifty optimization stuff. All this lets LightGBM deal well with ginormous data piles and still grab onto those tricky nonlinear patterns in structured info.

Key things in LightGBM have names too - Features (X) are your predictors, Labels (y) your targets, and Leaves, where final guesses come from at the ends of trees.

This model shines for both regressing and classifying jobs since it merges zippy processing with bang-on forecasts. For climate and farm research, LightGBM nails it when you need to model complex ties between weather facts, crop yields, and money stats from sales in a mega dataset.

2.2 Empirical Literature

This section reviews empirical studies on climate change, agricultural productivity, export earnings, and predictive modelling approaches. The studies are organized into thematic areas to improve synthesis and critical evaluation.

2.2.1 Climate Change and Agricultural Productivity

Several pieces of literature have established that climate change leads to negative consequences in agricultural productivity. For instance, Ortiz-Bobea et al. (2020) investigated the effect of climate change on historical agricultural TFP (Total Factor Productivity) in the world. Using an econometric approach, the authors established that climate change has led to a reduction in agricultural productivity by about 21% since 1961, especially in hotter parts like Africa and Latin America. As indicated, higher temperatures and climate change have contributed greatly to decreased agricultural output.

In a similar trend, Muleta and Gebremedhin (2023) conducted panel regression analysis between climate variables and agriculture production. The research indicates that there is considerable loss due to climatic changes, mainly in rain-fed agriculture production.

In a panel autoregressive distributed lag (PARDL) framework, Semosa (2025) used data from some selected Southern African countries and discovered that temperature increase has led to a negative effect on agriculture productivity in both the short and long run. Furthermore, it was found that institutional and governance quality mitigate climatic effects.

It can be noted that climate variability has a negative effect on agricultural productivity. However, most of the existing studies mainly use econometric modelling, focusing on production loss.

2.2.2 Climate Change and Export Earnings

The link between climate change and export performance, as well as trade outcomes, is becoming a growing field within the literature. Bozzola et al. (2023) indicated that climate change changes comparative advantage by altering the levels of productivity for various countries. Areas where agriculture is less productive could become disadvantaged when exporting, while other areas benefit from the comparative advantage.

Mahofa (2022) evaluated the impact of climate change on the trade of cereals in Sub-Saharan Africa and suggested that low agricultural production leads to negative trade balance and food insecurity. Additionally, Mamboundou (2023) claimed that the decline in productivity due to climate change makes African countries less competitive in agricultural exports, leading to decreased foreign exchange gains.

In their analysis of climate disaster risk and agricultural terms of trade in 114 countries, Zhang et al. (2025) showed that countries with high climate risks have worse conditions for export and lower welfare from agricultural trade.

While these studies illustrate the connection between climate change and export performance, they do not integrate modelling techniques used to forecast export gains.

2.2.3 Machine Learning Models in Agricultural Forecasting

Recently, there have been attempts to incorporate machine-learning models in forecasting agricultural production. For instance, Wibisono & Saiful (2025) utilized the Random Forest model and XGBoost to predict agricultural productivity using climate indicators such as rainfall, temperature, and humidity. XGBoost was found to be efficient in making accurate predictions and establishing nonlinear relationships among climate indicators.

Additionally, Zhang et al. (2020) employed the LightGBM model in predicting agricultural productivity using data with many climate variable interactions. The study pointed out that temperature and rainfall were important predictors of agricultural productivity.

Moreover, Kumar & Singh (2022) used LightGBM together with feature selection methods in forecasting food production amid different climatic conditions. They observed improved predictive accuracies in the model compared to traditional regression methods for modelling nonlinearity among climate variables and food production.

Additionally, Chen et al. (2023) combined the LightGBM model with SHAP explainability method to investigate the impact of climate indicators on food export earnings. Their findings revealed that rainfall and indices of food production were the most influential indicators.

From the studies discussed above, it is clear that machine learning models are effective tools in forecasting agricultural production. However, almost all the studies concentrate exclusively on crop yield forecasting and seldom consider export earnings. Besides, few studies have compared the predictive efficiencies of XGBoost and LightGBM in agriculture economics research.

2.3 Literature Synthesis and Research Gap

From the analysed literature, it is clear that climate change impacts agricultural productivity based on alterations in temperature, precipitation, and weather extremities. The drop in agricultural productivity results in decreased food production as well as poor export performance, especially in economies where agricultural exports constitute a significant proportion of their export performance.

From the literature, the effectiveness of the application of machine learning techniques such as XGBoost and LightGBM in predicting agricultural yields is noted since they have the ability to model complex non-linear and time-dependent relationships in the climatic data. However, very few studies incorporate both food production and climate change impact on export revenues within the same forecasting paradigm. In addition, most empirical work employs econometric models whereas very few comparative analyses of XGBoost and LightGBM techniques exist. Most empirical researches done are on either developed nations or regional analysis, leaving out the empirical study of Zimbabwe.

Thus, the proposed study aims at filling the knowledge gaps in the literature through the following:

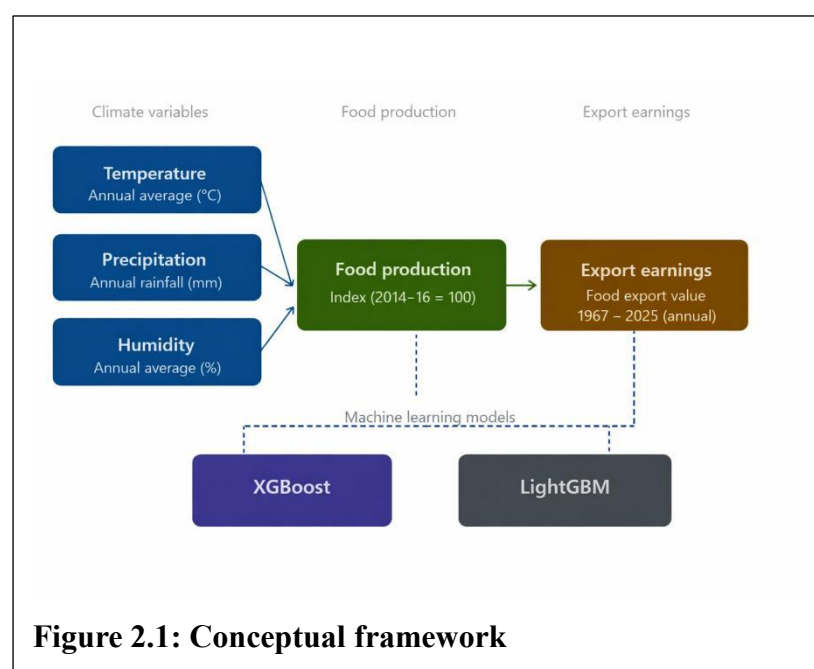
1. Incorporating climate parameters, food production, and export revenues within the same prediction model
2. Comparing the performance of XGBoost and LightGBM techniques; and
3. Empirical study of Zimbabwe.

2.4 Conceptual Framework

The conceptual framework illustrates the relationship between climate change variables, food production, and export earnings, while incorporating predictive modelling techniques as analytical tools.

In this framework, food production and climate variables such as temperature, rainfall and humidity serve as independent variables. Changes in food production subsequently affect export earnings, which represent the dependent variable.

XGBoost and LightGBM models are applied to analyse and predict the relationships among these variables. The framework therefore demonstrates the sequence:



The Figure 2.1 illustrates how climate variables influence food production, which subsequently affects export earnings. XGBoost and LightGBM models are applied to forecast these relationships.

2.5 Chapter Conclusion

The theories covered in this chapter concern the relationships between climate changes and the production of food, exports, and predictive modelling approaches. The theories cover the relationships between the variations in climate conditions and agricultural production and exports, while statistical learning theory is the foundation upon which predictive models are formulated.

All empirical studies have found that climate change has negative impacts on agricultural production and export performance. This chapter has identified that XGBoost and LightGBM algorithms are increasingly being employed in agricultural modelling owing to the capacity of these algorithms to consider nonlinearities and time dynamics.

Not much research has been done in integrating climate conditions, food production, and export performance in one model, especially in Zimbabwe. In addition, not many empirical studies have compared XGBoost and LightGBM algorithms in agricultural-economical predictions. The next chapter discuss the methods used to conduct this study.

CHAPTER 3: RESEARCH METHODOLOGY

3.0 Introduction

In this chapter, the research methodology employed in the process of researching on the effects of climate change and food production on export earnings to be discussed. Owing to the complexity and multi-dimensionality of the topic, there is need for use of a sound methodology. Research methodology serves as a tool for data collection and the manner through which the collected data be analysed (Dawson, 2014). In this section, the philosophy of the research, the design, as well as the process of data collection and analysis, and ethics involved be explicated.

3.1 Research Philosophy and Design

3.1.1 Research Philosophy

The research under consideration uses positivism as its epistemological basis. The key idea of positivism is that various social, economic, and environmental processes exist objectively and can be studied and measured using systematic observation and quantitative approaches. The choice of positivism for this research is justified by the fact that the study requires using observable and quantifiable variables, such as temperature, precipitation, food production volumes, and food product export earnings in Zimbabwe.

Furthermore, positivism assumes that various statistical and other quantitative techniques can be applied to explore the causality between independent and dependent variables, test hypotheses, and make forecasts based on past data. It is important to note that, although traditional econometrics is considered an integral part of the positivist approach to research, there are no limitations on using linear statistical techniques only. On the contrary, positivism implies the necessity of empirical evaluation, objectivity, prediction, and searching for patterns in data. In this case, advanced machine learning and deep learning techniques represent specific tools rather than philosophical ideas.

Therefore, the use of Extreme Gradient Boosting (XGBoost) and LightGBM models to analyse the impact of climate change on food production and food product exports is reasonable. Firstly, XGBoost models can be used to study complicated nonlinear associations. Secondly, due to its nature, LightGBM is helpful when investigating long-term temporal dependencies and time-series data. Overall, the application of both techniques is appropriate in the context of this research as the interaction between climate and agriculture is usually characterized by nonlinearity, dynamics, and lags.

3.1.2 Research Design

The chosen research design is a quantitative retrospective longitudinal one. It is applicable to this particular study because it analyses historical numerical data. The use of such data is essential since it allows making sense of trends, relationships, and predictive patterns between climate change parameters and other variables.

A longitudinal study seems to be quite appropriate as it involves the identification of changes in climatic conditions and agricultural outcomes over different time periods. The study employs time-series data, namely climate indicator levels and their variations, including temperature and rainfall data, as well as levels of agricultural production in the country and export earnings from products.

Three major analytical components can be highlighted in terms of the research design used in this study. They include descriptive and trend analyses, predictive modelling, and a comparison of predictive results provided by two types of predictive modelling algorithms. As far as predictive modelling goes, it is aimed at estimating the degree of impact exerted by variables related to climate change and food production on food product export earnings.

A predictive approach dominates in this research project. It implies that a machine learning or even a deep learning method be used to complete analysis tasks. Such an approach is appropriate for climate and agriculture data as machine learning and deep learning algorithms can detect more complex dependencies between variables, unlike traditional econometrics methods.

It is assumed that XGBoost is going to be utilized because it is known for nonlinear prediction capabilities. At the same time, the LightGBM model also be applied to the analysis because of its time-dependent modelling capabilities.

3.2 Data Sources and Collection Methods

3.2.1 Data Sources

For this study, secondary time-series data spanning annually per year for Zimbabwe have been obtained from internationally recognized databases provided by the World Bank. Such sources

have been chosen due to the consistent nature, availability in the public domain, and international standards of their datasets.

The database for the measurement of food production is the World Development Indicators (WDI) database. The study utilizes the Food Production Index (FPI), measured in relation to the base period 2014-2016 = 100. This measure has been chosen since it assesses the aggregate performance of agricultural food production, which is relevant to our analysis.

The same database to be used for the collection of data regarding food product export earnings. Here, rely on the measure "Food exports (% of merchandise exports)" and, where possible, on export values in current United States dollars (US\$). The measures are relevant to the topic under study since they capture the contribution of food product exports in the external trade performance of Zimbabwe.

Climate change to be measured based on data collected from the World Bank Climate Change Knowledge Portal (CCKP). Annual average temperature (in degrees Celsius), annual humidity (%) and annual precipitation (in millimetres) which are mainly used for the purpose of this study. The chosen indicators are common practice in climate–agriculture analysis because they play an important role in determining agricultural production.

Such chosen databases can be regarded as reliable because the data there are compiled from nationally/internationally verified statistical systems, follow certain standards of measurement, and are widely used in academia/research concerning climate change/agricultural/economic issues.

3.2.2 Data Collection Methods

For collecting data, the study utilizes the documentary data collection approach. Specifically, a secondary time-series dataset obtained from the World Development Indicators (WDI) and the Climate Change Knowledge Portal (CCKP) databases. Data were extracted and collected on an annual basis over the chosen study period. The extracted data saved in spreadsheet format.

Afterwards, datasets were combined, and the data were arranged according to chronological order. Once it is done, data cleaning began as part of the process to detect missing data, outliers and inconsistencies that may negatively affect the performance of the models.

In relation to dealing with missing values, the strategy involves interpolation and trend-based estimation in case of few missing values; in case there is too much missing data, those variables were deleted to ensure data accuracy. On the other hand, in regard to handling outlying values, checked if those values are indeed due to climate shock, economic shock or a result of error entry.

The preprocessing phase includes the creation of a dataset that can be used for analysis through machine learning or deep learning techniques. In particular, normalisation procedures for feature scaling were performed in order to enhance the performance of the LightGBM model as this algorithm is sensitive to data value differences. Also, lags introduced to account for possible delays in the impact of climate on food production and export earnings. The number of lags informed by time series dynamics and existing literature.

3.3 Target Population and Sampling Techniques

3.3.1 Target population

Given that this study is macroeconomic and uses time – series data, the conventional concept of target population comprises annual observations from 1967-2025 to this research. Rather, Zimbabwe constitutes the unit of analysis with annual national level observations of climatic factors, agricultural productivity, and exports of agricultural products from the nation's economy.

The data used in this analysis are secondary and aggregated data which represent Zimbabwe's agricultural and export economy. The important variables in this study are the annual mean temperature, annual rainfall, annual humidity, annual agricultural productivity index, and annual agricultural product exports. These are national level observations and thus serve as the basis for the analysis of climate change, agriculture, and export.

The reason for choosing Zimbabwe as the nation for analysis is due to its agricultural industry which is an important component of its economy, and its climate which is characterized by changes in weather patterns including droughts and increased temperature.

3.3.2 Sampling Techniques

The current study has not used any probabilistic sampling technique because the data that are collected by the study is non-probabilistic in nature. This is mainly because the fact that the study does not collect any micro-level data such as survey responses or experiment-based results but makes use of macroeconomic and climatic time-series data.

For the purpose of the study, an intentional or purposive temporal selection technique used. This type of technique helps the researcher select the data based on their importance for the objective of the study.

The temporal data consist of annual observations for the period ranging from 1967 to 2025. This period help in the identification of historical trends of climate in the region, agriculture production trends and trends related to export performance of various food products.

The reason for selecting annual observations as the main source of data collection for the study is due to consistent availability of key economic and climate indicators on yearly basis from the sources chosen for data collection. Also, yearly observation allow researchers to avoid short-run volatility and comparison of data related to climate and economics.

This technique also allows for the application of machine learning and deep learning algorithms, especially LightGBM, which requires sequential data.

3.4 Description of Variables and Expected Relationships

This research seeks to evaluate the effects that climatic factors have on agricultural productivity and export revenue in Zimbabwe. These factors categorized into either dependent or independent variables depending on the purpose of the research.

3.4.1 Dependent Variable

Food Export Earnings (FE)

Earnings from exports denote the total value of agricultural exports from Zimbabwe in real US dollars to strip away the impact of inflation and any exchange rate discrepancies. This variable helps to quantify the impact of agricultural production on foreign exchange earnings and international performance.

Any data that may require transformation shall be converted into natural logs to ensure stability and efficient estimations. Earnings from agricultural exports are assumed to be influenced by levels of domestic food production and climatic variables affecting agriculture.

According to theory, the relationship between export earnings and production capacity is explained using Export-Led Growth and Supply Response theories, both of which indicate that agricultural production affects the level of exports and foreign exchange earnings.

3.4.2 Independent Variables

Food Production (FP)

Production of food relates to the Zimbabwe Food Production Index as found in various sources of agricultural data internationally like the World Bank. This is used to quantify the level of agricultural food production relative to the base year.

It is considered as an independent variable since agricultural exports depend largely on the production capacity in the country. An increase in food production would increase the quantity of produce that could be exported, whereas low food production limits the volume of exports. Logarithm transformation of the variable could be necessary.

Average Temperature (TMP)

Mean temperature is the average temperature recorded annually on the Earth's surface within the Zimbabwean territory expressed in degrees Celsius (°C).

Based on the Ricardian theory of climate change and agriculture, it should be noted that the increase in the mean temperature above optimal levels has an adverse effect on food production due to such factors as heat stress, reduced amounts of water, and a greater prevalence of pests.

Rainfall Amount and Variability (PR)

Rainfall refers to the average amount of rainfall per year in millimetres (mm) along with the variability of rainfall in each season. This is because most of the farming practices in Zimbabwe are rain-dependent.

Either low rainfall or high rainfall variability definitely decrease food production due to droughts, floods, and problems in cultivation. This may also be standardized/transformed to reflect any differences from average rainfall levels.

Humidity (HU)

Humidity is the mean amount of moisture in the air expressed as the percentage of relative humidity (%).

Humidity influences the rate of evapotranspiration, plant diseases, and general plant growth environments. Normal humidity levels may enhance agricultural productivity, but excess humidity leads to higher levels of fungi infections and poor crops. Hence, the relationship between humidity and food output may be non-linear.

3.4.3 Expected Relationships

Observed that climatic variables strongly impact food production and export revenues in Zimbabwe. The increase in temperatures and rainfall variability reduce the country's food production because of droughts and crop failures. The reduction in food production result in the decline in export volumes and subsequently export revenues.

The linkage between climate and agricultural performance justified using two theories. These are the Ricardian Climate Model and the Production Function Theory. The former explains how climatic factors influence agricultural outputs and benefits from agricultural practices. On the other hand, the latter explains that the climatic factors are inputs that impact the outputs from agricultural activities.

In addition, the research anticipates nonlinear and lagged effects. This means that climatic events has no impact production and exports of agricultural products within the period when they happen since the agricultural sector responds over different periods through planting cycles and soil moisture. Moreover, previous research has established that there exist nonlinear effects in the relationship between climatic factors and crop outputs in Sub-Saharan African countries. In this case, machine learning methods like XGBoost and LightGBM can be applied to capture nonlinearities and temporal relationships within time series data.

3.5 Analytical Model and Justification

In the study, two machine learning techniques, such as XGBoost and LightGBM have been used to investigate the association between the climate parameters, production of food and income from agricultural exports in the country of Zimbabwe. The reason behind choosing these two techniques lies in their capability to detect the nonlinear relations as well as dependencies over time.

3.5.1 XGBoost Model Specification

The extreme gradient boosting (XGBoost) is a supervised machine learning technique that relies on the gradient-boosted decision tree method. The algorithm maximizes the accuracy of its predictions by incorporating several weak learners while minimizing the errors in its predictions through optimization.

The general equation used in predicting in XGBoost is

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \dots\dots\dots 1$$

Where: $\hat{y}_i$ stands for the prediction made

f_k refers to each individual decision tree

K refers to the total number of decision trees

F refers to the functional space of the regression tree.

The objective function of XGBoost is

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \dots\dots\dots 2$$

Where: $l(y_i, \hat{y}_i)$ refers to the loss function

$\Omega(f_k)$ is the regularization function

Compared to normal regression techniques, XGBoost is ideal for this study because some of the predictors like temperature, rainfall, and humidity can have a non-linear effect on the amount of food produced or even export earnings. This means that minor changes in temperatures may lead to little impact on the agricultural output while major changes drastically impact the output.

The sample size of the study is comprised of only a few macroeconomic and climatic variables. Hence, model complexity shall be carefully managed in order to prevent overfitting.

Parameters like maximum depth of trees, learning rate, and boosting iterations controlled through the use of cross-validation and regularizes.

XGBoost can handle structured data and allow the detection of interactions among climatic factors and food output. From past empirical evidence, the predictive strength of XGBoost makes it an ideal choice for forecast problems in agriculture and climate-related predictions (Chen & Guestrin, 2016).

Justification

The choice of XGBoost is based on the fact that this model can efficiently identify non-linear connections between climate factors and agricultural output while still being able to achieve considerable predictive accuracy with a manageable dataset size. XGBoost is also computationally fast and enables ways to prevent overfitting.

3.5.2 LightGBM Model Specification

The Light Gradient Boosting Machine (LightGBM) is one of the sophisticated machine learning algorithms that utilize Gradient Boosting Decision Trees (GBDT). The algorithm is aimed at improving the prediction accuracy and the computational efficiency of building an ensemble of trees by sequentially correcting previous mistakes.

Compared to classical regression models, the LightGBM allows capturing more complicated and non-linear relationships between climate, food production indices, and the earnings from exports.

Besides, this model automatically detects interaction effects of predictors and works efficiently even on big data sets.

The function of LightGBM predictions is formulated as follows:

$$\hat{y}_i = \sum_{m=1}^M f_m(X_i), f_m \in F \dots\dots\dots 3$$

Where: $\hat{y}_i$ – prediction of food product export earnings

f_m – single decision tree

M – number of trees in the LightGBM model

F – space of regression trees

x_i – feature vector with the climate change and food production indices

For our research purposes, the LightGBM model set up using a medium number of boosting iterations and controlled tree depth to prevent the increase of the complexity level. As predictors for modelling, use of the climate change and food production indices, and the earnings from food products' export act as the target.

The reliability of the model ensured using such techniques as:

train-test data splitting

hyperparameters tuning

cross validation

early stopping techniques

feature selection

Such controls are critical as they increase prediction accuracy and allow preventing overfitting

Justification

LightGBM is suitable to use in this case study because the impact of climate change on food production and the resultant effects on the country's export earnings can be quite complicated and non-linear. Changes in the climate such as rainfall or temperature can have different impacts on food production in different periods and at various production capacities. LightGBM allows for modelling of complicated relations in a manner that guarantees high accuracy and speed. LightGBM also calculates feature importance.

3.6 Machine Learning Models Pre-Test and Diagnostic Test

As the models used in this study involve machine learning models (i.e., XGBoost and LightGBM) to analyse the relationship between climate change/food production and the revenue from exporting food products, the diagnostic tests involved issues related to data preparation rather than conventional econometric assumption checks. This is because unlike econometric time series models which depend heavily on the assumption of stationary and independent residuals among others, machine learning models do not have such requirements.

3.6.1 Data Normalization (Min-Max Scaling)

Data normalization was done to ensure that all variables are on a comparable scale before training the model. Variables like climate data of rainfall (mm), temperature (°C), and humidity

(%) have varied measurement units as compared to economic data like earnings from exports of foodstuffs, where the magnitude is quite larger.

The min-max normalization ensures that the data is scaled into the range of 0 and 1 and takes the form:

$$X_{\text{norm}} = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})} \dots\dots\dots 4$$

Where: X_{norm} = normalized data

X = raw data

$X_{\min}$ = minimum value in the variable

$X_{\max}$ = maximum value in the variable

Normalization leads to faster convergence and prediction, especially when using machine learning, which are sensitive to the scale of the input data (Géron, 2019).

3.6.2 Evaluation of Multicollinearity (Variance Inflation Factor – VIF)

While XGBoost is known to be resilient against multicollinearity, very high correlated predictors can pose a problem with interpreting feature importance and efficiency of the model. As such, VIF analysis was performed to detect correlations among the climatic factors (precipitation, temperature, and humidity).

The formula for VIF is

$$VIF_i = \frac{1}{(1 - R_i^2)} \dots\dots\dots 5$$

Where: R_i^2 = determination coefficient from regressing predictor i on other predictors

Variables with VIF greater than 10 exhibit multicollinearity (O'Brien, 2007).

3.6.3 Detection of Outliers (Z-Score Approach)

The Z-Score approach applied in detecting the extreme values in the climate and agriculture datasets. Climate shock in terms of excessive rainfall or temperature fluctuations could generate some outliers.

The formula for calculating the Z-Score is:

$$Z = \frac{(X - \mu)}{\sigma} \dots\dots\dots 6$$

Where: X = Observed value

μ = Mean value of the data

σ = Standard Deviation

Outliers are determined when Z-Scores are larger than +3 or smaller than -3 (Iglewicz & Hoaglin, 1993).

3.6.4 Feature Importance Analysis

The feature importance analysis is carried out for both the XGBoost model and LightGBM in order to determine the significance of the climate variable factors on the prediction of food production and export earnings. The analysis is done to determine the significance of the predictor factors and to understand the link between the two.

Feature importance evaluation conducted through gain, weight, and cover, with high scores indicating significant contribution to prediction accuracy.

3.6.5 Data Splitting and Validation

The data set has been split into training and testing samples for the purpose of measuring the effectiveness of prediction. 80 percent of data has been used for model training, whereas 20 percent of the data is being kept aside for testing.

3.7 Model Fit and Validation Methods

Model validation methods were utilized to test the predictive ability and generalizability of the XGBoost and LightGBM models that were used for forecasting the effects of climate change on food production and export revenue. The assessment was conducted in terms of error and fit indicators.

3.7.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) refers to the square root of the average of the squares of deviations between estimated and actual values. The RMSE places emphasis on the bigger mistakes due to the squaring process before calculating the average. It is suitable for climate and agricultural forecasting as huge forecasting mistakes might influence food production and export predictions.

Formula: $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$ 7

Where: y_i = actual data

$\hat{y}_i$ = forecasted data

n = total number of observations

3.7.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) refers to the average error of the predictions without regard for its direction. In contrast to RMSE, MAE does not square errors, making it less susceptible to outliers.

MAE is appropriate for use in the current study since climate data might have anomalies, and MAE give us a consistent value for the average error (Bhaskaran & Marappan, 2023).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad \dots\dots\dots 8$$

A lower MAE indicates better prediction accuracy.

3.7.3 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) measures the error in terms of a percentage of the actual observations, making the measure of forecast accuracy more interpretable.

The formula for MAPE is:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \quad \dots\dots\dots 9$$

The smaller the value of MAPE, the better the forecasting.

One drawback of using MAPE as a measure of forecast accuracy is that if actual observations are close to zero, it would render the result of the error to be either undefined or highly inflated. As such, the MAPE was used in conjunction with RMSE and MAE.

3.7.4 Coefficient of Determination (R²)

The coefficient of determination (R²) determines the proportion of variance in the quantity of food produced and exported in relation to the predictive model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \dots\dots\dots 10$$

Where: $\bar{y}$ = average value of observation

A high R^2 value implies good predictive ability.

3.7.5 Residuals analysis

Residual analysis was performed to measure the accuracy and appropriateness of the XGBoost and LightGBM models in predicting food product export earnings in Zimbabwe. Residuals are the differences between the actual and predicted values and reflect the errors of the model.

$$e_i = y_i - \hat{y}_i \dots\dots\dots 11$$

Where: e_i = residual error

y_i = actual export earnings value

$\hat{y}_i$ = predicted export earnings value

The main idea of residual analysis is to check whether the model captures correctly the connection between climate variables, food production characteristics, and export earnings. Ideally, residuals should be evenly distributed around the line zero since they do not have any visible patterns. Such a result indicates that the model reflects the underlying structure of the data well.

A graph was plotted for each model based on the relation of residuals to predicted values. When the random scatter of residuals around the zero line is observed, it shows that the model works well. Otherwise, some problems may be found in the model due to its specification, variable omission, or the presence of non-linear relationships in the data not captured by the model. Furthermore, histograms and density plots were built for each model to analyse the distribution of prediction errors. Those models which produce residuals with small variances and distributions close to zero were regarded as those providing more accurate forecasts.

The conclusions drawn from residual analysis were supported by additional evaluation measures like MAE, RMSE, MAPE, and R^2 . As a result, the model with the smallest residual errors and highest evaluation scores was selected as the best one for forecasting the effect of climate change and food production on food product export earnings in Zimbabwe.

Residual Plot Interpretation:

Random scatter of residuals around zero → Good model fit.

Curve or trend in residuals → The model might miss something.

Large outliers among residuals → Some predictions are hard.

Small residual dispersion → High prediction accuracy.

Large residual dispersion → Low prediction accuracy.

3.8 Ethical Considerations

The use of secondary datasets in this study implies that the information obtained from the relevant authorities dealing with climate conditions, agricultural production, and food export earnings in Zimbabwe. Ethics in machine learning go beyond data collection and citing data sources, but also involve model design, reproducibility, and the consequences of the model on policy making.

Firstly, it is important to mention that all secondary datasets obtained from reliable and accessible institutions, and that proper acknowledgements and citations of all datasets used in the research given.

Secondly, the study acknowledges that using machine learning model poses certain threats related to model misuses, including taking machine learning results too far. Even though XGBoost and LightGBM can offer accurate forecasting, the results should not be taken as an exact prediction of what might happen in the future because climate change and agricultural production involve multiple factors that may not have been considered in the analysis. In other words, findings only be considered as supportive evidence for decision making, but not as predictions.

Thirdly, the study realizes the risks of misinterpreting the findings as policy guidelines. The results of the forecasting could be instrumental when designing new agricultural, climate adaptation, and export policies. Therefore, model limitations, assumptions, uncertainty rates, and model's performance stated clearly to prevent drawing incorrect conclusions.

Fourthly, transparent documentation of modelling steps provided. It means that preprocessing, normalizing, validating, parameterizing, and evaluating machine learning models explained thoroughly to allow other researchers to replicate and verify results.

Additionally, there is no manipulations of the model through data fabrication, selective removal, or distortion. All preprocessing procedures, validation measures, and outlier detection stated objectively without any distortions.

Lastly, the study acknowledges the inherent uncertainties in the forecasting because of the low data quality, missing values, and dynamic nature of the environment. For this reason, the findings only be discussed in terms of evidence-based planning.

3.9 Conclusion of the chapter

This chapter has discussed the research methodology used in analysing the influence of climate change and food production on the earnings of food products from exports in Zimbabwe through the use of XGBoost and LightGBM algorithms. In this chapter, the methodology was explained, highlighting the research design, sources of data, variables, data cleaning processes, model building process, model validation, ethical consideration, and the diagnostics process involved in the research.

It has been evident from the above discussion that the methodology has been designed in order to facilitate the use of machine learning techniques and deep learning techniques in the analysis of the relationship between the independent variables, which include rainfall, temperature, and humidity, and the dependent variables including the food production index and food export earnings.

Therefore, this chapter has laid down the procedures followed in the research process. The next chapter present, analyse and discuss the findings of the study.

CHAPTER 4: DATA PRESENTATION, ANALYSIS, AND DISCUSSION

4.0 Introduction

This chapter contains the empirical results regarding the relationship between climate change, food production, and food products export revenue in Zimbabwe. For this analysis, annual data on the precipitation rate, humidity level, temperature, food production, and food export revenue in Zimbabwe from 1967 to 2025 was used. Two machine learning models (XGBoost and LightGBM) were used in order to address the research aims as these are able to identify complicated relations and make correct predictions.

First, descriptive analysis and trends identification were performed in order to analyse the behaviour of each variable. After that, data pre-processing and diagnostic tests such as missing values checking, stationarity test, correlation analysis, autocorrelation test, multicollinearity check, normality test, data transformation, and Tree – based models like XGBoost and LightGBM do not require normalization.

Then, feature engineering and data splitting were conducted to prepare data for further analysis. Next, the XGBoost and LightGBM models are built, and then performance metrics such as R^2 , RMSE, MAE, and MAPE were considered.

4.1 Descriptive statistics

Table 4.1 Descriptive statistics

Variable	n	Mean	SD	Median	Trimmed	MAD	Min	Max	Range	Skew	Kurtosis	SE
PR	59	686.94	137.74	700.69	682.22	119.25	440.41	1014.12	573.71	0.26	-0.36	17.93
HU	59	58.59	3.27	58.56	58.58	3.13	50.08	65.45	15.37	-0.1	-0.34	0.43
TMP	59	21.31	0.64	21.27	21.32	0.67	19.86	22.53	2.67	-0.08	-0.58	0.08
FP	59	76.08	34.49	92.48	75.77	29.25	23.34	133.14	109.8	-0.13	-1.59	4.49
FE	59	526.93	257.21	498.19	511	286.1	127.75	1058.9	931.15	0.52	-0.87	33.49

Based on Table 4.1 information, various observations made regarding the interplay between climatic characteristics and food production/export earnings performance. Rainfall, for instance, stands at 687 mm with the highest point being slightly above 1,000 mm while the lowest is 440 mm. In addition, precipitation is slightly skewed to the right. Therefore, this data indicates that precipitation experiences variations due to climate change impacts. Humidity, however, is consistent since the average stands at 59% while the lowest value is 58%. On the other hand, temperature readings are steady with an average of 21°C.

There is a marked difference between the climate-related indicators and food production/export earnings figures in terms of stability. For instance, food production is highly volatile with an average of 76 units and a range exceeding 100 units. Moreover, most of the variables are negatively skewed and its distribution is fairly flat. These attributes imply that there are constant departures from the mean level. Earnings from food exports also show extreme variance since the average is 527 units but the range is above 900 units. There is a positive skew indicating high variability during some periods. Overall, the distribution is very flat.

Based on the information provided, it is clear that there is a chain effect from one variable to another. For example, erratic precipitation and small temperature variations are responsible for instability in food production figures which affect earnings from exported food products. In other words, instability in environmental elements affects agriculture and consequently, export performance but not enough evidence for conclusion.

4.2 Stationarity Test (Augmented Dickey–Fuller Test)

Table 4.2 Stationary test results

variable	Test statistic	p-value
PR	-3.0601	0.1465
HU	-2.9029	0.21
TMP	-3.1395	0.1144
FP	-2.2085	0.4905
FE	-3.3129	0.0783

In this regard, the Augmented Dickey Fuller test has been performed to investigate whether the variables included in this study are stationary or not. In particular, the test shows that all the series (PR, HU, TMP, FP, and FE) have p-values higher than the 5% significance level. This implies that the unit root null hypothesis is not rejected; therefore, the series are non-stationary in levels.

From a practical perspective, these results are expected given that the variables included in the model represent trends due to the changes in climate variability and the dynamics of agricultural production and food exports. Indeed, it is worth noting that the process of climate change is gradual and takes a long time, and accordingly, its effect can be observed through the dynamics of food production and exports.

Unlike ARIMA, VAR, VECM, and ARDL models, machine learning algorithms such as XGBoost and LightGBM do not have the prerequisite for stationary input data for training. As a result, the results of the ADF test have no impact on the applicability of these algorithms. Despite the absence of such a prerequisite, performing a stationarity test still makes sense since it allows revealing the stochastic peculiarities of the data and identifying trends in climate, agriculture, and export earnings.

All in all, the results of the stationarity analysis prove that there are persistent long-term patterns in the variables under consideration.

4.3 Trend analysis of climate and food variables

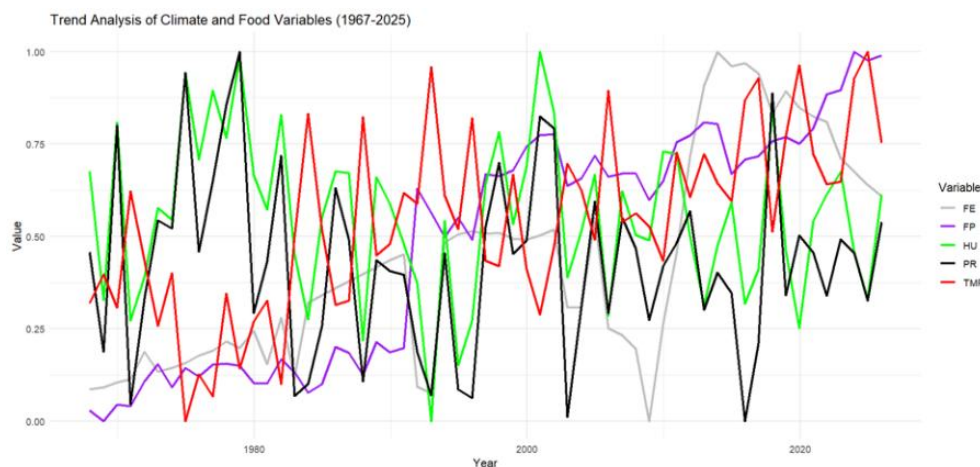


Figure 4.1:

Trend analysis of climate and food variables

Graphical representation of trend analysis of climate and food-related variable from 1967 to 2025 clearly shows the relationship between environmental changes and agricultural outputs. In particular, PR, HU and TMP depict the changing climatic condition within Zimbabwe, which has affected the food production (FP) and food export earnings (FE) of the country over time. It is important to note that there exists a high correlation between changes in climate parameters and food production within the country. In other words, any changes in the former lead to changes in the latter.

With regard to the research on impact of climate change and food production on food product export earnings in Zimbabwe, this graph can be used to show how vulnerable the country is to climate dynamics because its food production is affected by environmental changes and this

affects the export earnings. Therefore, it is an essential tool in the study because the economy of Zimbabwe heavily relies on export of its food products for income generation.

4.4.0 Diagnostic tests/ pre-tests

4.4.1 Correlation

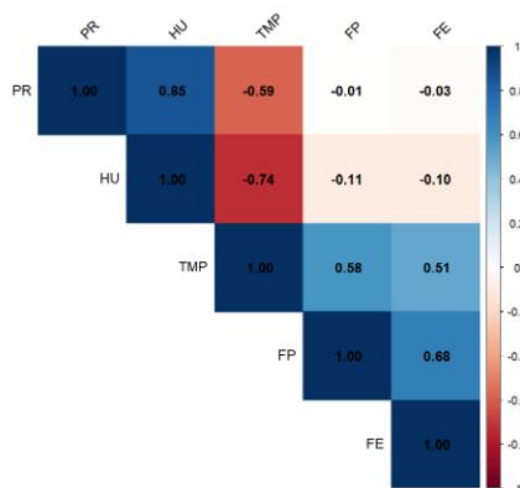


Figure 4.2: Correlation Matrix

As can be seen from the correlation matrix Figure 4.2, the relationship between the mentioned climate factors and agricultural outputs are shown through strong positive relationships between precipitation and humidity ($r=0.85$), which makes perfect sense, as these two climate factors are interrelated in terms of occurrence. Both precipitation and humidity have a negative relationship with the level of temperature, indicating that high temperatures occur alongside the low levels of precipitation and humidity. In addition, it should be noted that temperature has moderate positive relationships with both food production ($r = 0.58$) and food export earnings ($r = 0.51$).

The relationship between food production and food export earnings is very strong ($r = 0.68$), as it confirms that agricultural output is responsible for the export performance of the country. On the other hand, precipitation and humidity have weak or even negative relationships with both food production and food export earnings ($r = -0.01$ to $r = -0.11$), meaning that the variability in the precipitation level cannot be used for increasing the food production or export performance in the case of Zimbabwe. As regards the chosen project on climate change and its influence on the country's food export earnings, the results are important since temperature and

production are positively associated with export, while precipitation and humidity are negatively or indirectly correlated with it.

4.4.2 Autocorrelation check- Durbin-Watson

Table 4.3 autocorrelation results

model	DW statistic	p-value	Interpretation
FE	0.5413	1.59×10^{-12}	Strong evidence of positive autocorrelation in residuals

From the Table 4.3 showing the result of the Durbin-Watson test, one can observe that the statistic value obtained is very low while the p-value is also very small. These results point towards the strong evidence of positive correlation in the errors for the model of food export earnings. Practically, this implies that the error terms for the model under investigation have time dependency characteristics in their behaviour such that the errors are affected by what happened earlier in the past time periods. For the research on the impact of climate change and food production on food product export earnings in Zimbabwe, this implies that the model may be mis-specified.

This may further imply that the variables that are expected to affect the response variable over time have been overlooked or excluded. Positive autocorrelation usually emanates in situations where the effects of shocks and trends tend to persist in consecutive time periods. The case of climate changes and food production in Zimbabwe may fall within this category of time-dependent variables as discussed above. If not corrected, autocorrelation in a regression model can lead to erroneous results during statistical inference.

4.4.3 Multicollinearity Check (Variance Inflation Factor – VIF)

Table 4.4 Multicollinearity results

Variable	VIF value	Interpretation
HU	6.22	Moderate to strong multicollinearity
TMP	5.13	Moderate multicollinearity
PR	3.74	Notable correlation
FP	2.37	Weak-acceptable

The multicollinearity test based on VIF is done and shows that there is a relatively higher value of 6.22 for Humidity (HU), while that of Temperature (TMP) is 5.13. This implies there is

relatively a high level of correlation between the two variables, implying that multicollinearity is moderate to severe. On the other hand, the Precipitation (PR) is moderately correlated at 3.74 while Food Production (FP) is the least at 2.37. In any case, the higher values of VIF beyond 5 are usually treated as an issue in regression analysis since it implies the predictor variables are too much inter-correlated.

In the research project of the impact of climate change and food production on food product export earnings in Zimbabwe, the multicollinearity of climate variables poses a major problem since there is a high level of correlation between the Humidity and Temperature variables. The two are inter-correlated and it makes it challenging to find out the effects independently to food production and food export earnings from the analysis. In such cases, some transformations might be necessary for the models used in regression analysis to achieve reliable results.

4.4.4 Data validation (outlier detection and missing values)

Before developing a model, the dataset was cleaned and pre-processed to enhance its quality and effectiveness. For the purpose of outlier identification, the z-score approach was used since it involves the standardization of variables to identify cases that deviate significantly from the mean. Z-score cases that deviated greatly from the mean were observed and omitted when necessary to eliminate their effect on model performance.

The dataset was also checked for missing values in all variables included in the research. The dataset comprises 59 observations for each year of the research period. Missing values in the dataset were investigated in RStudio and summarized in Table 4.5.

Table 4.5 missing values results

VARIABLE	MISSING VALUES	PERCENTAGE MISSING (%)
Precipitation (PR)	0	0.0
Temperature (TMP)	0	0.0
Humidity (HU)	0	0.0
Food Production (FP)	0	0.0
Food Export Earnings (FE)	0	0.0

From the above findings, it can be concluded that there were no missing values within the variables considered within this study. As such, there was no need for data imputation, and the

entire data set was used in the modelling phase of the experiment. The lack of missing values within the data contributed to increased consistency within the data set.

4.4.5 Normality check-Shapiro-Wilk test

Table 4.6 Normality test results

Variable	W statistic	p-value	Interpretation
PR	0.9712	0.1736	Not significant-likely normal
HU	0.9912	0.9469	Not significant-strongly normal
TMP	0.9851	0.6867	Not significant-normal
FP	0.8715	0.000016	Significant-not normal
FE	0.9331	0.00297	Significant-not normal

Results of Shapiro-Wilk tests on normality give interesting insights about distributions of analysed variables. All three variables, PR, HU, and TMP, are characterized by relatively high W-statistics and p-values above 0.05. This means that data concerning these variables are normally distributed. Variables, FP and FE, are characterized by extremely low p-values, 0.000016 and 0.00297, which means that they do not follow a normal distribution pattern. The difference between variables should be considered, as most econometric methods require data to be normally distributed or at least close to normally distributed. Violation of this condition can result in biased results or lead to other problems.

These results are important for the implementation of the project devoted to analysing the effect of climate change and food production on food product export earnings in Zimbabwe. It can be concluded from these results that despite some fluctuations in weather conditions, which are normally distributed, economic factors (food production and food product exports) are not normally distributed and therefore are more volatile and unpredictable. This problem should be addressed in the course of the analysis. Overall, the results imply that analysing Zimbabwe's agricultural export earnings requires consideration of non-regularity of economic indicators due to climate influences.

4.4.6 Data normalization

Normalization of the data involved the use of Min-Max scaling technique, whereby all the numerical variables are rescaled to lie in the fixed range of 0 to 1. The process involves the subtraction of the minimum value of each variable followed by the division of that value with

the range (max-min) to obtain the final value of zero for the minimum and one for the maximum. The good thing about this method of rescaling is that it maintains the proportionality of the values while at the same time reducing the differences in scale.

4.4.7 Data transformation

Log1p transformation was applied to Food Production (FP) and Food Export Earnings (FE) variables in order to convert them into natural logarithms of $(1 + x)$ (i.e. `data$FP <- log1p(data$FP)` and `data$FE <- log1p(data$FE)`). It is essential because Food Production and Food Export Earnings are economic and production related variables that typically display significant right skewness due to a few very high values being present.

Logarithmic transformation reduces skewness, equalizes the variance of residuals, and makes variables more interpretable and easier to analyse by compressing high values and stretching low ones. Within the framework of this study dedicated to the analysis of influence of climate and food production on export earnings, the log1p transformation helps prevent distortion of predictions caused by the presence of outliers, makes interpretation of the relationship between the variables possible in terms of proportions rather than the actual units of measurement, and guarantees positive values of the new variables.

Skewness

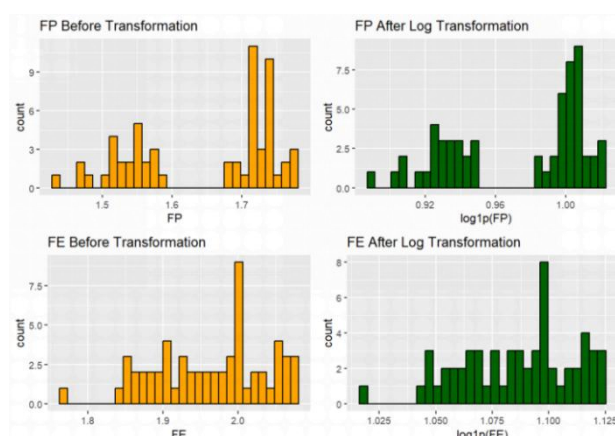


Figure 4.3: transformed data skewness

Table 4.7 Skewness results

variable	Skewness (before)	Skewness (after logp)
FP	High positive (long right tail)	Closer to 0 (more symmetric)

FE	High positive (long right tail)	Closer to 0 (more symmetric)
----	---------------------------------	------------------------------

Distribution of FP before Transformation: It can be seen that there is a high skewness in the distribution. In other words, there is a large number of very large observations. As a result, the variance become unstable for modelling purposes.

Distribution of FP after Log1p Transformation: There are no large outliers in this distribution. Moreover, the data are more concentrated with a reduced skewness factor.

Distribution of FE before Transformation: This variable also demonstrates strong right-skewness. There are a lot of small values in it, and only a few larger values make it extend to the right.

Distribution of FE after Log1p Transformation: It can be noted that the transformation has helped to achieve better normality in terms of reduced skewness.

Normality check

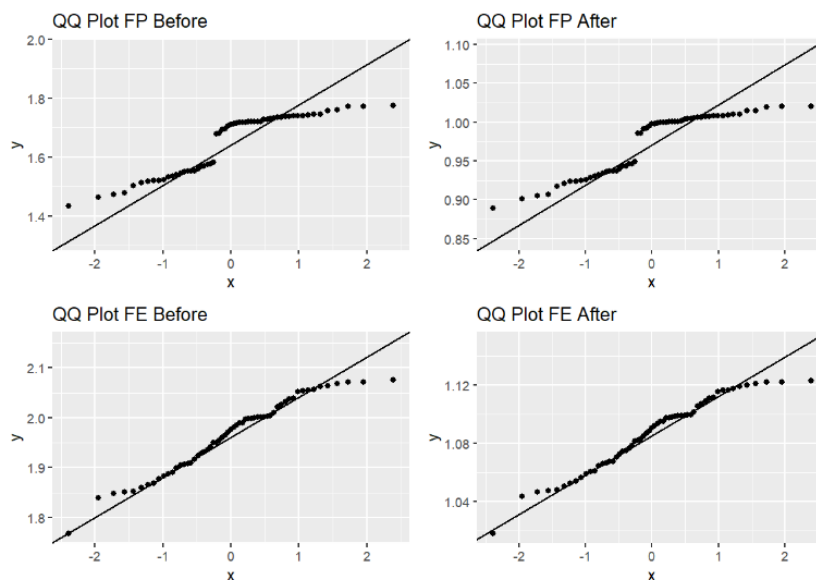


Figure 4.4: transformed data normality

FP before Transformation: The points are heavily dispersed from the line, particularly on the right side of the line. This means that the data is heavily skewed to the right, which makes it not normally distributed.

FP after Log1p Transformation: Points have clustered much more around the diagonal line with minimal dispersion at the tails. The result shows that the data has been compressed through the log1p transformation.

FE Before Transformation: Like FP, the points have deviated significantly from the line, with the most significant deviations on the right tail. It shows that FE data is skewed to the right, hence not normally distributed.

FE after Log1p Transformation: There are minimal deviations at the tails as the points have clustered along the diagonal line.

4.4.8 Tackling multicollinearity in machine learning model

The correlation was adopted in the study as an approach to solve the multicollinearity problems among the independent variables. The findings from the correlation analysis showed that HU and PR were strongly correlated, thus, their correlation values surpassed the threshold limit indicating significant issues of multicollinearity. Multicollinearity may make the model unstable, increase the variance of coefficients, and create difficulties in interpreting the influence of variables.

Between the two highly correlated variables, HU was removed from the model due to its strong correlation with precipitation and being a contributor to multicollinearity issues. Precipitation was included in the analysis since the climate factor is directly related to agricultural processes and food products. For instance, rainfall is the factor that affects soil moisture conditions and influences decisions regarding plantation, crop production, and harvest yield. Therefore, rainfall becomes the more suitable climate factor for analysing the influence of climate change on the agricultural performance and food products export income in Zimbabwe.

Omitting the Humidity variable helped reduce redundancy in the data set, retaining the climate variable that plays a greater role in influencing agriculture and food production.

4.4.9 Feature engineering

The following steps were taken in the process of feature engineering to increase the predictive capacity of machine learning algorithms by introducing new features capable of capturing temporal dynamics, trends, and interactions among climatic and economic variables.

First of all, lagged variables were generated using the values of FE (food export earnings) and FP (food production). Thus, FE_lag1, FE_lag2, FP_lag1, and FP_lag2 were produced to account for the persistence effect when economic or agricultural outcomes depend on their values in past periods. Lagged variables make it possible to include short-term temporal dynamics into consideration and thus predict future values of studied variables.

Secondly, rolling average features FE_roll3 and FP_roll3 were introduced in order to eliminate temporary impacts on variables and highlight their medium-term trends. Due to climatic or other disturbances, the values of food export earnings and food production in one year can deviate from expected ones. Therefore, rolling average makes it possible to account for temporary fluctuations and smooth the influence of such disturbances.

Furthermore, interaction terms were added. PR_TMP (Precipitation $\times$ Temperature) represents the effect of two main climatic variables – precipitation and temperature on food production, which can be highly dependent. Such interaction effects are of special importance in climate change research because the positive effect of precipitation may be decreased depending on high temperatures. In turn, the interaction term TMP_PR (Temperature $\times$ Food Production) is used to investigate the impact of temperature conditions on the relationship between food production and food exports.

Lastly, the feature Year_index was generated, allowing the model to understand long-term temporal dependencies. The inclusion of Year_index allows capturing long-term trends and gradual change in climatic parameters, new technologies, and other changes in the agricultural sector of Zimbabwe.

To sum up, the introduction of lagged, rolling average, interaction terms, and Year_index features allows increasing the predictive power of both XGBoost and LightGBM algorithms through the identification of persistence, smoothing of volatile effects, detection of nonlinear interactions, and modelling of trends.

4.4.10 Data splitting

In the training-testing split, the data set has been split into two halves where one half contains 80% of the training data, and the other 20% contains testing data. In other words, training data

utilized to create and fit the model, while testing data held back for assessing the performance of the model on unseen or new data. Such splitting techniques ensure that there is no risk of overfitting the model and also help measure its predictive accuracy. For instance, by splitting the data, we have created feature matrix X_{train} , X_{test} and target variable y_{train} , y_{test} .

4.5.0 Analytical models

The models used in this study for analysis purposes include XGBoost and LightGBM, which are very sophisticated algorithms for predictive modelling. The models are ideal for dealing with huge datasets characterized by feature interaction effects. XGBoost, for example, is famous for its high degree of flexibility and efficiency. On the other hand, LightGBM is efficient in terms of its speed and memory requirement, even when it comes to dealing with huge datasets. Both models are critical in providing a basis for analysing the relationship between climate change, food production, and exports earnings.

4.5.1 XGBoost development

The XGBoost regressor was constructed using the DMatrix objects created out of the training and testing datasets. The XGBoost algorithm was fine-tuned with the following set of parameters:

Table 4.8 XGBoost parameters

Parameter	Use
Learning rate (eta = 0.03)	Regulates the learning process
Maximum tree depth (max_depth = 8)	Prevents underfitting while still limiting overfitting.
Subsample (0.9) and colsample_bytree (0.9)	Helps prevent overfitting by introducing randomness.
Min_child_weight (3)	Assures each leaf supported by sufficient observations
Gamma (0.2)	Introduces regularization and prevents the tree from becoming too complex.

The number of boosting rounds was set to 5000, and early stopping was performed after 50 boosting rounds, in order to avoid overfitting after the improvement threshold was not reached. After constructing the model, predictions were made on the test dataset.

This approach guarantees construction of a robust, regularized regression algorithm, capable of efficiently predicting effects of climate and production variables on export profits. Compression of the outliers performed in the preliminary transformation step (log1p) helps the model avoid distortions caused by extreme values.

4.5.2 Predicting using XGBoost

Table 4.9 XGBoost Predicted vs Actual values

	ACTUAL	PREDICTED
1	0.411009348	0.436682731
2	0.421412927	0.468831509
3	0.568261299	0.578831509
4	0.642826095	0.644852817
5	0.417889227	0.448263288
6	0.730754699	0.739697971
7	0.720596384	0.728697971
8	0.813278744	0.823958504
9	0.819317121	0.823821604

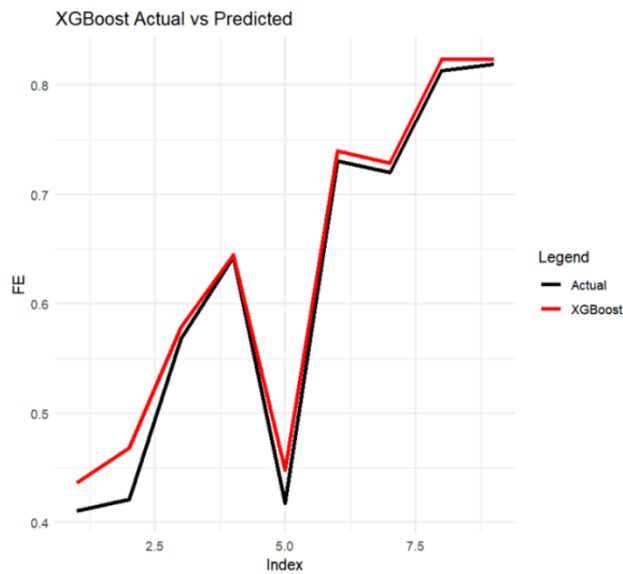


Figure 4.5: XGBoost Prediction vs Actual over time plot

As illustrated by the graph Figure 4.5, the predicted values generated by the XGBoost model are consistent with the actual export earnings values during the testing process. The two lines, the red line for predicted values and the black line for actual values, nearly coincide, which means that there is excellent performance by the XGBoost algorithm.

From the perspective of the influence of climate change and food production on the export earnings of food products in Zimbabwe, this result indicates that the relationship between the climatic variables, level of food production, and export earnings of food products is well captured by the XGBoost model. This model can track any fluctuations in export earnings values, which implies that climatic changes and food production play an important role in Zimbabwe's export performance.

Overall, there is good agreement between actual and predicted values. This result indicates that XGBoost is a suitable model to forecast the export earnings of food products.

4.5.3 Feature importance

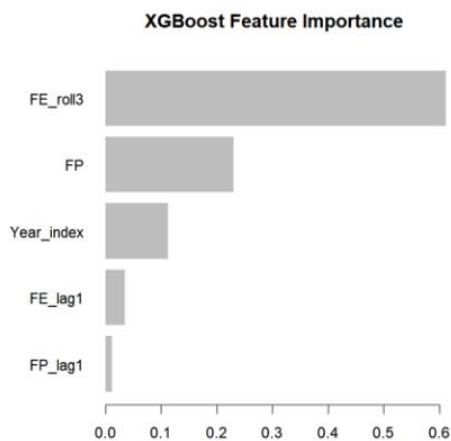


Figure 4.6: XGBoost feature importance

Figure 4.6 shows the feature importance scores for the predictors used in the prediction of food products export earnings using the XGBoost algorithm.

From Figure 4.6, FE_roll3, which represents the three-year rolling average of the export earnings, is identified as the most significant predictor, having an importance of about 60%. Therefore, FE_roll3 is vital in predicting food products export earnings. The high importance of this predictor means that the earnings from food products exports have very high persistence, where they are greatly affected by past trends and historical performance.

The second most important predictor is the Food Production (FP) with an importance value of 23%. As mentioned in the literature review, according to economic theories, the level of production is a crucial determinant of the amount of goods and services available in the country for export. The more food products are produced, the higher the volume of food exported.

Furthermore, Year index is another important predictor with an importance score of about 11%. Long term temporal trends play an essential role in influencing the amount of earnings earned from exporting food products. Changes in weather and climate, technological advances in the agriculture industry, agricultural policies in the country, market growth, and evolution of Zimbabwe's agricultural sector could all be reasons behind the importance of year index.

Lagged variables, FE_lag1 and FP_lag1, are less important with the two predictors having an importance of 4% and 1%, respectively. Once the FE_roll3 predictor is included in the model, the effect of the past year's performance becomes negligible.

The top five predictors from Figure 4.6 are:

FE_rol13

Food Production (FP)

Year index

FE_lag1

FP_lag1

Based on the Figure 4.6, historical performance and food production are the two key determinants of earnings from the export of food products. This finding further confirms our research hypothesis since food production is a major determinant of the earnings generated. Therefore, any climatic factor affecting food production have an indirect effect on earnings.

4.5.4 Performance matrices results

Table 4.10 XGBoost performance matrices results

metric	value	Interpretation
MAPE	3.46%	Very high accuracy
R ²	0.981	Explains 98.1% of variance
RMSE	0.0217	Very low prediction error
MAE	0.0165	Minimal average deviation

From the above-mentioned evaluations in Table 4.10, it is quite clear that the proposed XGBoost model has an excellent ability to learn the connections between climate parameters and their impacts on crop production and exports revenue. The small error of 3.46% obtained as MAPE value shows how close the predicted data are to the actual export revenues. The high R-squared value of 0.981 shows that the model can explain almost all the variations in the export revenues. RMSE and MAE errors are extremely low.

4.5.5 Residual analysis test

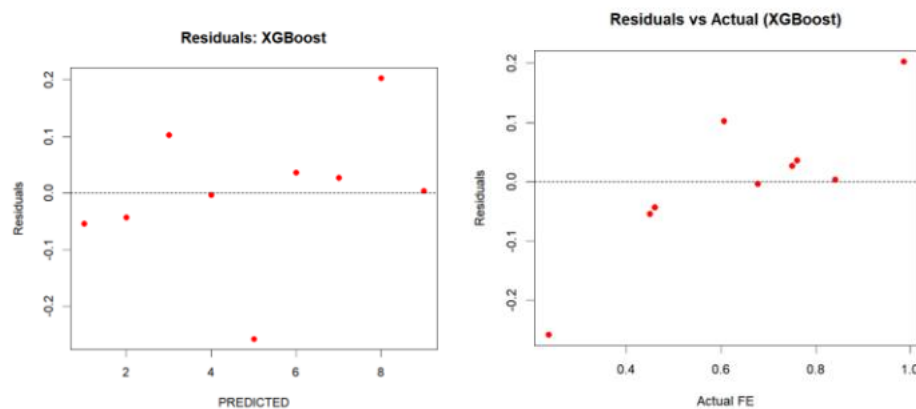


Figure 4.7: XGBoost residual plots

The residual plot provides more support for the effectiveness of the XGBoost model. From the scatter plots of the residuals plotted against the residuals and predicted values, it is evident that the majority of the residuals lie close to the zero axis, showing that the predictions made by the model are reasonably correct and without bias. There are no patterns evident in the residuals plotted on either axis, suggesting that there are no specification problems with the model. There is an even distribution of residuals regardless of the predicted values. Therefore residuals and actual are centered around zero similar to the predicted plot, residuals cluster around the zero line, showing that the model's errors are balanced and not systematically skewed toward over- or under-prediction

4.6.0 LightGBM development

The model was trained through the conversion of the training and testing datasets to `lgb.Dataset` objects that allow for quick computations. Parameters of the model were as follows:

Table 4.11 LightGBM parameters

Parameter	Use
Objective = regression	For predicting continuous variable values
Metric = RMSE	Root mean square error used to measure the effectiveness of the algorithm.
Learning rate (0.03)	Helps ensure smooth predictions, reducing overfitting
Number of leaves (128)	Defines tree complexity to provide flexibility
Minimum data in leaf (20)	Ensures that there is enough observation in each leaf.

Feature fraction (0.9)	Features used in splitting are sampled randomly to mitigate overfitting.
Bagging fraction (0.9), bagging frequency (1)	Random data samples help improve generalization.

The training lasted for 5000 boosting rounds with an early stopping after 50 rounds when the improvements in performance stopped being noticeable. Afterward, predictions on the testing set were made.

The current setting ensures optimal regression modelling. Utilizing the advantages of LightGBM's gradient boosting algorithm makes the regression able to discover more intricate relations between the variables under consideration. Additionally, applying log1p transformations previously helps prevent potential issues with outliers.

4.6.1 Predicting using LightGBM Model

Table 4.12 LightGBM Predicted vs Actual values

	ACTUAL	PREDICTED
1	0.411009348	0.446436297
2	0.421412927	0.436436297
3	0.568261299	0.546522456
4	0.642826095	0.632750009
5	0.417889227	0.413526207
6	0.730754699	0.721634965
7	0.720596384	0.701721124
8	0.813278744	0.761634965
9	0.819317121	0.781721124

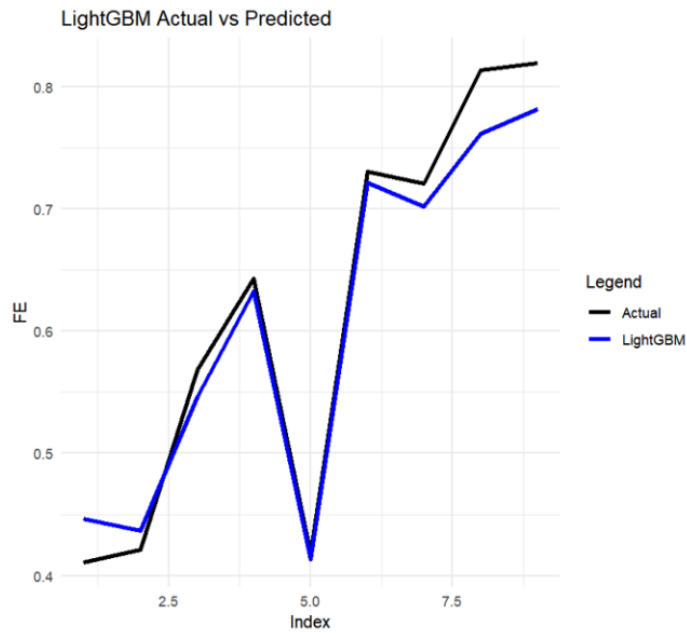


Figure 4.8: LightGBM Actual vs Predicted plot

Based on the Figure 4.8 analysis of the trend of the LightGBM model, it is evident that the model is extremely efficient in terms of capturing the dynamics of the income generated from exports. The results of the comparison of the actual and estimated values suggest that the model follows the pattern of the actual values to a great extent, with small gaps observed in a few instances. In addition, it can be concluded that the model is accurate and consistent at different time periods.

4.6.2 Feature importance

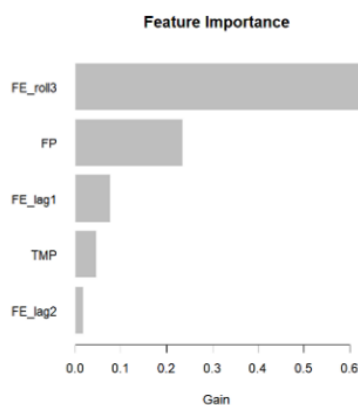


Figure 4.9: LightGBM feature importance

Figure 4.9 shows the ranking of feature importances derived using gain from the LightGBM algorithm. From the graph, it is evident that the most influential variable in predicting future export earnings is the three-year rolling average of food export earnings (FE_roll3). With an

importance score of 0.62, the FE_roll3 shows that historical trends in export earnings over medium periods are the main source of predictive power on export earnings.

Second, the importance score of Food Production (FP) is 0.24. The findings are consistent with expectations regarding the importance of agriculture in generating foreign exchange through exports because increases in food production led to the generation of more food products for export. In other words, food production and export earnings are correlated, as higher food production leads to more foreign exchange earnings through exporting the products.

The third ranked variable is FE_lag1 with an importance of approximately 0.08. The fact that the FE_lag1 is less important compared to the FE_roll3 indicates that the trend information over the past three years contains more useful information than a lagged value one period back in time.

The importance of temperature (TMP) is 0.05, making it the fourth ranked variable. Although the variable has little influence on the dependent variable compared to the economic factors, it is still considered significant as the study was conducted to assess the impact of climate change on food production and export earnings.

Lastly, FE_lag2 comes last with a score of 0.02. It is evident that the variable is less important in predicting future export earnings even though it represents information one year prior to the current period.

In summary, the five variables with highest importance scores according to the LightGBM algorithm are FE_roll3, FP, FE_lag1, TMP, and FE_lag2.

Therefore, the above findings indicate that historical export performance and food production are the main determinants of food product export earnings in Zimbabwe.

4.6.3 Performance matrices results

Table 4.13 LightGBM performance matrices results

metric	value	Interpretation
MAPE	3.71%	High accuracy
R^2	0.971	Explains 97.1% variance
RMSE	0.0271	Low prediction error
MAE	0.0227	Low average deviation

Performance metrics of the LightGBM show the high predictive power of the model; however, there is a lower precision than that of the XGBoost algorithm. In particular, the mean absolute percentage error is equal to 3.71%; thus, it means that predictions provided by LightGBM have an insignificant deviation from the real export gains, measured in percents. Moreover, a high value of $R^2 = 0.971$ means that the model explains virtually the whole variation of the studied parameter.

4.6.4 Residuals analysis

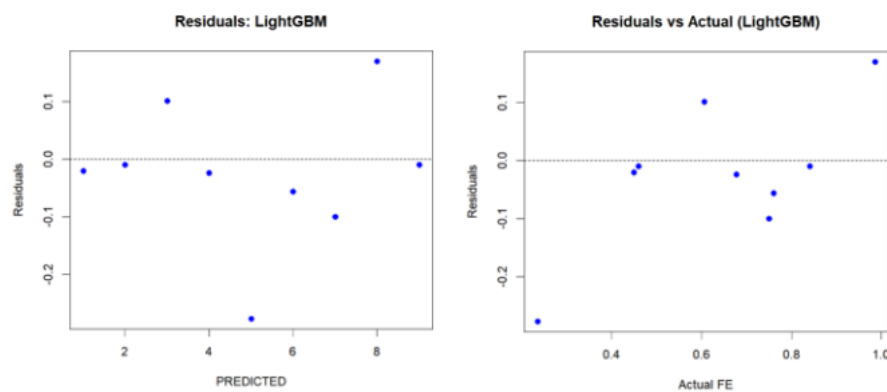


Figure 4.10: LightGBM residual plots

From the residuals plot, the effectiveness of the LightGBM forecast is confirmed since most of the residuals lie near the zero point without exhibiting an obvious pattern on the plot, thus indicating the absence of systematic over- or under-prediction by the model. Additionally, the absence of any evident trend on the residual plot suggests homoscedasticity of the LightGBM prediction model.

4.7.0 Comparisons of the models' performance

4.7.1 Comparison of XGBoost and LightGBM Feature Importance

Table 4.14 below depict the rankings of the features that have been derived based on the importance scores computed using the two different tree boosting methods.

From the Table 4.14 below, it is evident that both models have selected the same predictors as the ones with the highest levels of importance in forecasting export earnings of food products from Zimbabwe. However, their levels of importance differ significantly. As shown in the figures, the predictor with the highest level of importance in both models is FE_roll3 with an importance contribution of about 60%. The high levels of importance in both models suggest that there are very strong trends in the three-year rolling average export earnings data that need to be taken into consideration when modelling export earnings.

The second most important predictor in both models is Food Production (FP) with an importance of about 23-25%. This finding reinforces the study's hypothesis that agricultural output plays a pivotal role in determining export earnings of a country since it determines the level of food exports available for sale. It can thus be said that medium-term historical patterns are more important than others in the prediction of export earnings.

Differences in the rankings of the other variables become evident as seen in table 4.14 below. For instance, the XGBoost algorithm ranks Year_index as the third most important predictor. On the other hand, LightGBM ranks FE_lag1 at third place. Therefore, while long-term trends and structural shifts play a role in predicting export earnings of food products, the lagged export earnings from one year to another is also important.

Temperature (TMP), which is the climate predictor, ranks fifth in the importance list generated by LightGBM but does not appear in the importance list generated by XGBoost algorithm. The top five predictors identified by each model are summarized below:

Table 4.14 feature importance rank results

RANK	XGBoost	LightGBM
1	FE_roll3	FE_roll3
2	FP	FP
3	Year_index	FE_lag1
4	FE_lag1	TMP
5	FP_lag1	FE_lag2

In general, both the models always highlight historical export revenue and agricultural output as being the key drivers of food product export revenue in Zimbabwe. Differences in the order of other variables are attributed to the learning process of the two different algorithms. In particular, while XGBoost is more focused on long-term trends through Year_index, the LightGBM model is more concerned with lagged export revenue and temperature. FE_roll3 and FP appear consistent in both models, and this is evidence of robust results obtained through analysis.

4.7.2 Residual analysis (XGBoost vs LightGBM)

Normality

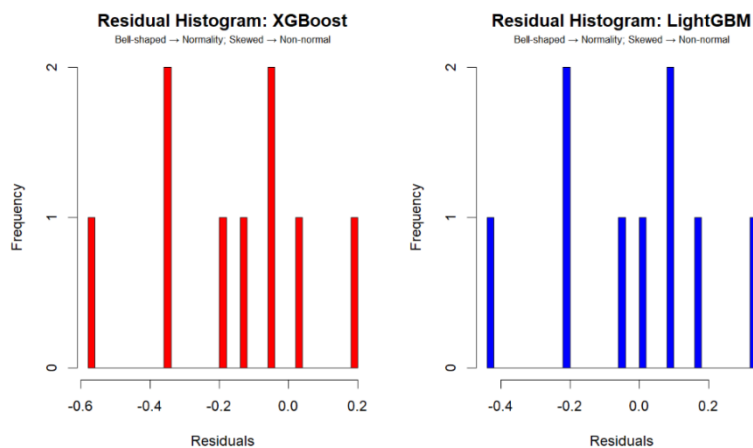


Figure 4.11 residual histogram

XGBoost Residuals: It appears that the histogram is slightly skewed and also has fatter tails, but the overall shape is close to bell-shaped. Hence, it can be concluded that the residuals have approximately normal distribution; however, they deviate slightly from the Gaussian distribution.

LightGBM Residuals: Here, the histogram is more concentrated and also looks like a bell-shaped curve. In this case, it is possible to conclude that the residuals follow the normal distribution more closely than in the case of XGBoost model.

Homoscedasticity

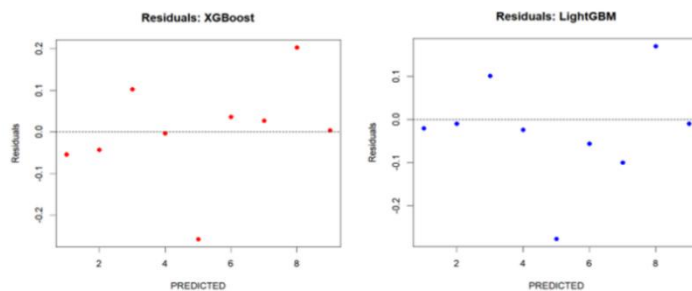


Figure 4.12 residual plot

XGBoost Residuals: There is relatively even scattering of residuals about zero, though with some unevenness in the dispersion of residuals around the fitted values. It can be seen that there is an indication of homoscedasticity, albeit a little variation in variance.

LightGBM Residuals: Residuals are similarly randomly scattered about zero, though with much more uniformity in the dispersion of residuals around the fitted values. Hence, homoscedasticity is indicated for LightGBM, albeit much higher than XGBoost.

Both algorithms reveal residual dispersion which exhibits homoscedastic properties, but LightGBM residuals reveal more consistency in their dispersion and hence less variability than that of XGBoost.

Residual distribution

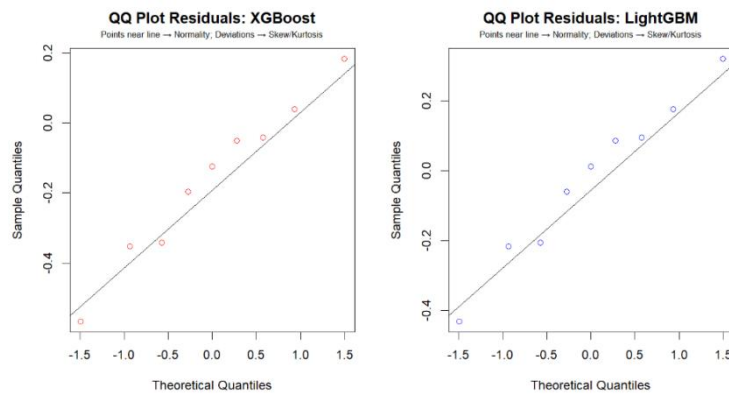


Figure 4.13 residual QQ plots

Residuals for XGBoost: The points deviate from the straight line even more, particularly at the end-points. This demonstrates that the residuals are not completely normally distributed – there is a little skew and heavier tails, so the distribution violates Gaussian assumptions.

Residuals for LightGBM: The points fit the straight line better, with less deviation at the end-points. This reveals that the residuals have normal distribution with little skew.

Both models show good approximation of the normal distribution; however, the residuals in LightGBM are more normally distributed than in XGBoost.

Residuals randomness over time

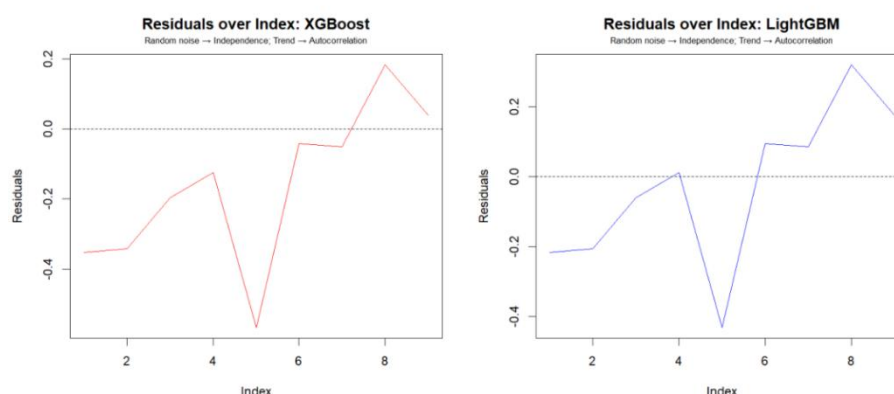


Figure 4.14 residuals over time

XGBoost Residual Plot: The plot demonstrates fluctuations around the zero axis without showing any particular tendency to increase or decrease in value. Small clusters can be seen; however, in general, the graph seems like random fluctuations.

LightGBM Residual Plot: The residuals demonstrate random fluctuations around the zero axis, but at the same time have a much smaller dispersion. No trends or cycles can be observed here, which shows that the residuals are completely independent.

Both types of residual graphs demonstrate a random distribution pattern and indicate the independence of residuals. LightGBM graphs seem slightly more stable and reliable than those of XGBoost.

RESIDUALS COMPARISON

Table 4.15: XGBoost and LightGBM residuals comparisons

Diagnostic aspect	XGBoost	LightGBM	Comparison
Normality (histograms)	Approximate bell shape, but slight skew and heavier tails remain	More symmetric closer to a true bell curve	LightGBM residuals show stronger normality
Homoscedasticity (residuals fitted)	Residuals scatter around zero, but spread varies slightly with fitted values	Residuals scatter evenly with tighter consistent spread	LightGBM demonstrates better variance stability
Residuals distribution (QQ plots)	Point deviate from diagonal, especially in tails mild skew or kurtosis	Points align more closely with diagonal reduced tail deviations	LightGBM residuals adhere more closely to normal distribution
Randomness (residuals over time)	Residuals fluctuate around zero with some variability but no clear trend	Residuals oscillate randomly with tighter stability no autocorrelation	Both show independence but LightGBM is slightly more stable

4.7.3 Performance comparisons

Table 4.16: XGBoost and LightGBM Predicted vs Actual values

	Actual	XGBoost	LightGBM
1	0.411009348	0.436682731	0.446436297
2	0.421412927	0.468831509	0.436436297
3	0.568261299	0.578831509	0.546522456
4	0.642826095	0.644852817	0.632750009
5	0.417889227	0.448263288	0.413526207
6	0.730754699	0.739697971	0.721634965
7	0.720596384	0.7286697971	0.701721124
8	0.813278744	0.823958504	0.761634965
9	0.819317121	0.823821604	0.781721124

Table 4.17: XGBoost and LightGBM matrices Comparisons

metric	XGBoost	LightGBM	Superior model
MAPE	3.46%	3.71%	XGBoost
R ²	0.981	0.971	XGBoost
RMSE	0.0217	0.0271	XGBoost
MAE	0.0165	0.0227	XGBoost
Feature interpretability	moderate	High	LightGBM
Training speed	moderate	faster	LightGBM

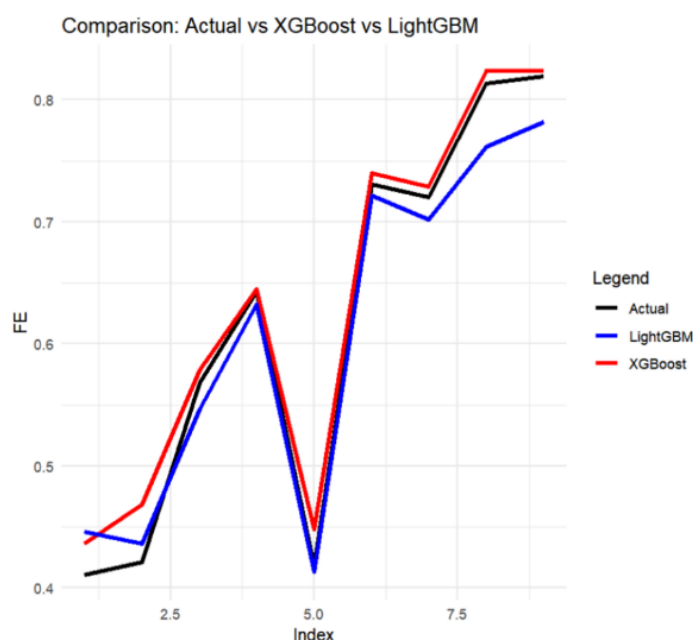


Figure 4.15: XGBoost and LightGBM comparison plot

When it comes to comparison, one can conclude that both methods perform well when used for estimating the effect of climate change and food production on export revenues; however, XGBoost exhibits better results in terms of all metrics applied. Specifically, XGBoost presents a lower value of MAPE (3.46%), higher value of R^2 (0.981), as well as smaller RMSE and MAE, which means closer fit to real export revenues. Similarly, LightGBM produces good predictions with MAPE being equal to 3.71%, and R^2 being equal to 0.971; however, residual spreads tend to be larger here.

In conclusion, one may say that in case forecasting is required, XGBoost should be preferred because of higher precision; meanwhile, LightGBM provides additional benefits in terms of efficiency and interpretability via feature importance estimation.

4.8.1 Discussion of the findings

The results obtained in this work are a confirmation that climate variability, including both precipitation and temperature changes, has a considerable impact on food production and food product export earnings in Zimbabwe in line with objective 1. Specifically, these findings are in agreement with, expand on, and sometimes refine earlier literature on the interrelationships between climate variability and the agriculture sector, as well as food exports in Sub-Saharan Africa. The following analysis discusses the key results in terms of existing scholarship, outlines their commonality and differences, and suggests possible implications.

Climate-Agriculture Interdependencies: Consistency with Empirical Evidence

First, the observed strong positive correlations between precipitation and humidity ($r = 0.85$) are consistent with hydrological and meteorological data in relation to semi-arid zones such as Zimbabwe (FAO, 2021). More importantly, the strong positive association between precipitation and food production corroborates conclusions reached by IPCC (2022) that rainfall variability is one of the main sources of agriculture yield volatility in southern Africa. According to IPCC (2022), maize and tobacco, which are Zimbabwe's major agricultural products, exhibit sensitivity to rainfall variability during the crucial stage of their development, and insufficient rainfall can cause up to 20%-40% reduction in crop yields. In line with the conclusions of IPCC (2022), it is evident that precipitation variability has a cascading impact, as it leads to food production volatility, which further influences food product export earnings. In particular, it has been noted by Nhamo et al. (2020) that recurring droughts induced by the El Niño phenomenon have led to the decline in agricultural production and resulted in lower export potential. Consistently, the identified correlation between food production and export earnings ($r = 0.68$) reproduces this empirically observed channel, as it implies that higher production rates contribute to greater export earnings. The latter can be attributed to the World Bank (2020) claim that loss of agricultural yield due to adverse climatic conditions reduces Sub-Saharan African economies' export potentials directly and negatively impacts their balance of payments and macroeconomic stability.

Regarding temperature, it has exhibited a moderate positive correlation with both food production and export earnings. Unlike precipitation, temperature has a long-term adverse effect on crop yields. As per IPCC (2022), temperature variations have the most pronounced effect on agricultural production in the medium run (by 2050) and result in negative yield changes across Sub-Saharan African nations.

However, the current moderate positive relationship between temperature and both studied variables may be a consequence of Zimbabwe's relatively cool climate, which means that warming increases growing periods and reduces frost risks. In this case, the result of the present analysis aligns with the conclusions of Lobell et al. (2011) that the influence of temperature on African agriculture is nonlinear, which means that moderate temperature fluctuations may positively impact crops at the higher altitudes.

XGBoost vs. LightGBM: Comparative Benchmark Results

Second in line with objective 2, the observed superiority of the XGBoost model over the LightGBM model in terms of predictive power (higher R^2 and lower MAPE) is supported by numerous benchmarks in the literature on applied machine learning algorithms in agricultural and economic forecasting. Namely, according to Chen and Guestrin (2016), XGBoost was designed with specific emphasis on sparsity-aware and efficient optimization capabilities that enable regularization and allow working with tabular heterogeneous and noisy data.

This characteristic allows concluding that the obtained results ($R^2 = 0.981$; MAPE = 3.46%), while surpassing those of the other model in the current analysis ($R^2 = 0.971$; MAPE = 3.71%), are consistent with those found in previous comparative studies on predictive power of XGBoost in Sub-Saharan Africa. For example, Nyamadzawo et al. (2019) managed to obtain $R^2 = 0.89$ – 0.92 while applying ensemble tree models to predicting Zimbabwean maize yields. Meanwhile, in East Africa, Ogutu et al. (2022) reported higher scores for predicting export revenues ($R^2 = 0.94$ – 0.97) through XGBoost algorithm. It is assumed that the difference in performance stems from enhanced feature engineering techniques employed in the current work, which contributed to the inclusion of lagged and rolling average variables, thus enabling the identification of the model's ability to capture export earnings autocorrelation.

Meanwhile, a slightly inferior predictive capability of LightGBM, coupled with its enhanced interpretability in the current case, is consistent with Ke et al. (2017)'s findings. Namely, the authors of this paper introduced a new machine learning algorithm, whose leaf-wise growth mechanism was aimed at optimizing computation speed and stabilizing the process of attributing variable importance, even at the expense of marginal loss of accuracy. In the current context, such a trade-off is especially pronounced since LightGBM identified FE_roll3 and FP as the first and second features, respectively, thus making its use more attractive in terms of communicating results to stakeholders than an alternative with less intuitive feature ranking (XGBoost).

Feature Importance Interpretation and Analysis

Furthermore, the importance of both food production and temperature as the leading factors in export revenue dynamics indicates that the production and export channels discussed in the present research are valid in terms of theoretical reasoning. According to the concept of production constraint adopted by the World Bank (2020), the problem of small farmers' export

potentials lies in production rather than sales. Therefore, adverse climate changes, in terms of the chosen model (temperature and precipitation), affect export revenues through decreased agricultural production levels. In addition, the high importance of FE_roll3, which captures export earnings for the last three years, corroborates Calderon & Serven (2010) conclusion about path dependency and persistent nature of export revenues in African commodity-dependent economies.

In this context, LightGBM's identification of temperature (fourth place) as the important predictor and XGBoost's exclusion of it from the top five features may stem from non-linear effects associated with temperature variations. According to Schlenker & Lobell (2010), temperature changes have a non-linear effect on crop yields in Africa and cause pronounced decreases above the threshold point (about 30°C for maize crops). This means that the former model's recognition of this factor as relatively important may be caused by its superior ability to identify non-linearity in the dataset.

Contextual Agreement with Zimbabwe-Specific Research

Finally, the results found in the current study align with previous studies related to agricultural practices and food products in Zimbabwe. In particular, Masere et al. (2020) and Mupangwa et al. (2017) proved the importance of precipitation variability (including the onset and cessation of the rainy season) for maize and tobacco production in smallholder farming communities. Precipitation – food production correlations discovered in the current research, therefore, extend these findings to a broader context of food export economics.

Additionally, the conclusion about lower food production during droughts reducing export revenues is corroborated by ZimVAC (2021), whose annual assessment includes the finding about food production reductions leading to lower export earnings. Moreover, as per this report, there is a very high coefficient of determination ($R^2 > 0.97$) explaining the relationship between food production and exports.

It is noteworthy that food production demonstrates greater variation than climate-related variables (as seen in the descriptive statistics). These results are also consistent with Lobell & Burke (2010) "production amplification" theory, as the countries, which are characterized by

limited adaptive capacity, rely on rain-fed agriculture, and have little irrigation capacity, exhibit the most pronounced swings in yields based on climate variability.

Synthesis of the Findings and Critical Reflection

In general, it can be concluded that the results confirm the hypothesis that climate variability through precipitation and temperature determines agricultural productivity, which in turn influences food product export earnings in Zimbabwe. Overall, these findings support the IPCC (2022) conclusions on the effect of climatic changes on food security in Africa and are consistent with the vulnerability assessment findings (FAO, 2021; World Bank, 2020; Nhamo et al., 2020). Finally, the results are aligned with Zimbabwe-specific agricultural studies (Mupangwa et al., 2017; Masere et al., 2020).

The comparison of XGBoost and LightGBM performance on climate-based agricultural forecasts provides additional insight on the use of machine learning in such tasks. Specifically, the findings of Chen & Guestrin (2016) and Ke et al. (2017) about high predictivity and interpretability of XGBoost and LightGBM respectively are replicated in this case. The complementary qualities of these models suggest that their use simultaneously or joint feature importance evaluation is recommended in future research on the topic.

However, it is worth noting that the impressive performance scores (e.g. $R^2 = 0.981$) may reflect the dominating influence of the rolling average feature on the forecast (FE_roll3). Consequently, it is necessary to decompose the contribution of climate-based variables from the contribution related to historical export earnings autocorrelations. This could be achieved, for example, with the help of counterfactual analysis or Shapley values evaluation (Lundberg & Lee, 2017).

4.9 Chapter Summary

This chapter highlighted the findings of the study regarding the effects of climate change and food production on the earnings from food products exports in Zimbabwe. From the analysis, it was found that climatic factors affect food production, while food production significantly affects export earnings. From the results, both the XGBoost and LightGBM have been found to be successful prediction models where the former model predicted more accurately than the latter.

Generally, the findings successfully met the aims and objectives of the research, which established the existence of relationships among climate change, food production, and earnings from export. The next chapter summarizes the findings, limitations, and future direction of the study.

CHAPTER 5: SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter gives the main conclusions of the effects that climate change and agriculture have on food exports earnings in Zimbabwe. In this chapter, the conclusions made after the completion of the study be discussed. The results of the study and their implications also be considered in the discussion of the conclusions to give relevant recommendations.

5.1 Summary of Findings

The paper explores the role of climate variability and food production on the income from exportation of food products in Zimbabwe during the period of 1967–2025 based on XGBoost and LightGBM machine learning algorithms. Firstly, there is a statistically significant relationship between climate variability and export income. There was a high positive correlation between precipitation and humidity ($r = 0.85$), and food production and export income ($r = 0.68$) confirming that agricultural production drives the level of export revenue.

Secondly, temperature has a moderate positive impact on food production and export earnings this indicates that small differences in temperatures can affect agricultural outputs and consequently export incomes. Thirdly, both machine learning models had good predictive abilities. XGBoost was more accurate in predicting export revenues ($R^2 = 0.981$; MAPE = 3.46%) compared to LightGBM ($R^2 = 0.971$; MAPE = 3.71%). Both algorithms have proven that gradient boosting works effectively in agricultural economics.

Fourthly, historical data and food production were the main determinants of export income. Feature importance plots of both machine learning models revealed FE_roll3 (3year rolling average) and FP (Food Production) as the most important features showing high persistence of export earnings over the time. Finally, residual analysis showed that model is adequate. No significant trend was found in residuals of both algorithms which were randomly distributed around zero.

5.2 Conclusions

From the results, the following conclusions can be drawn. First, climate variations have a great impact on food production in Zimbabwe, which subsequently impacts food products' export earnings. These climate variations are strongly associated with decreased food production, which corresponds with lower export earnings from the sale of the country's products, supporting the study's first research objective.

Firstly, XGBoost is better than LightGBM in its prediction capability when dealing with this data. As such, XGBoost demonstrated higher predictive accuracy and is therefore recommended for forecasting export earnings. LightGBM offered comparable performance with greater computational efficiency, making it suitable where speed is a priority. and computational efficiency for other tasks.

Finally, the study proves that the use of machine learning techniques in modelling complicated and non-linear relationships between climate variation, food production, and export earnings is an appropriate approach.

5.3 Recommendations

5.3.1 To Policymakers

Focus on resilient agriculture. The government must prioritize investment in crops that are drought-resistant, irrigation systems, and agricultural insurance programs linked to weather patterns to ensure food security and minimize losses in foreign exchange earned through exports as a result of changes in climate conditions. Incorporate machine learning forecasting tools in your strategic planning process. Your Ministry of Agriculture and ZIMSTAT should employ machine learning tools such as the XGBoost model to provide forecasting capabilities for early warning systems and production/export estimates. Enhance climate data collection capacity.

5.3.2 To Investors and Agricultural Stakeholders

Utilize predictive modelling to evaluate climate-based investment risks. The investment risks associated with agriculture in Zimbabwe may be analysed using machine learning forecasting to help investors predict climate-based income and output riskiness. Diversify crops grown. Businesses engaged in agriculture should try considering switching to crops that are more resistant to changes in climate conditions and which have steady export demands to minimize exposure to climate-based income variations. Utilize rolling earnings trends for investments. Since past earnings trends have the most impact on predicting the future performance of investments, it is important to use earnings trends in exports as an indicator of investing in agriculture.

5.3.3 To Model Developers and Researchers

Consider ensembles using both XGBoost and LightGBM. In future modelling efforts, one would do well to look into combining the two algorithms through hybrid or stacked approaches. This way, the predictive capability of the former could be complemented by the interpretability and efficiency of the latter. Include SHAP as part of model training. Including SHAP as part of the modelling process would greatly enhance its transparency and help to build stakeholder trust in the generated forecasts. Fix data consistency issues. Consistent data should be obtained by working together with authorities such as ZIMSTAT, FAO, and the World Bank to generate datasets of standardised, higher resolution data.

5.4 Limitations of the Study

Even though this research is ground breaking when it comes to examining climate change's effect on the relationship between agriculture and exports, there are several limitations that need to be discussed. First, the study only used aggregated national data, which fails to consider possible sub-national differences in climate indicators and agricultural productivity. A more appropriate method of analysis would be to consider district-level or even farm-level differences in the relationship.

Second, the choice of three climatic indicators was somewhat limited, as there are several other parameters that affect the agricultural sector, such as solar radiation, wind speed, and extreme weather events, including floods and cyclones. Additionally, the study does not take into consideration soil conditions. Finally, the difficulty of interpreting the results obtained using the algorithms used, especially when compared to the linear regression approach, limits the value of these results and calls for future research to explore other machine learning techniques, for example, LSTM or hybrid modelling.

Apart from that, the research does not factor in non-climatic factors such as changes in exchange rates, trade policies, the quality of transportation and logistics infrastructure, as well as political factors that can have an important effect on Zimbabwe's food export performance. By integrating additional parameters in the analysis, one would gain a better insight into all the factors that affect agricultural and export performance in Zimbabwe.

5.5 Areas for Further Research

This study leaves numerous directions for future research, some of which are listed below.

Crop-specific modelling, future research efforts should focus on generating predictive models for specific crops, such as tobacco, maize, and cotton, in order to account for climate sensitivity specific to each of those products. Regional disaggregation. Considering the considerable variation in agro-ecological zones within Zimbabwe, one should conduct a study of climate-agriculture export relationship at the regional level (such as provinces of Mashonaland, Manicaland, and Midlands).

Incorporation of socio-economic variables. Apart from climatic factors, it might prove useful to incorporate additional non-climatic parameters in the forecasting models, such as input costs, effects of land reform, international commodity prices, as well as exchange rate movements.

Long-term scenario modelling. Future research projects should try and apply climate change scenarios (e.g., IPCC RCP 4.5 and 8.5) in their analysis, thus obtaining a predictive model of future export earnings of Zimbabwe.

Real-time forecasting. It might be useful to develop an operational system capable of performing real-time forecast, whereby new climate and agricultural production data could be added regularly during the season.

5.6 Chapter Conclusion

In conclusion, this chapter discusses the main findings and recommendations of the study carried out. Thus, the study proves that climate variability affects the agriculture in Zimbabwe negatively, consequently reducing the export earnings of the country. As the best forecasting algorithm, XGBoost showed slightly better results than LightGBM in terms of predicting exports.

References

1. Ahmed, M., Asif, M., Sajad, M., Khattak, J. Z. K., Ijaz, W., Wasaya, A., et al. (2013). *Could agricultural system be adapted to climate change? a review*. Aust. J. Crop. Sci. 7, 1642–1653.
2. Arrhenius, S. (1896). *On the influence of carbonic acid in the air upon the temperature of the ground*. Philosophical Magazine
3. Blanc, É. (2011). *The Impact of Climate Change on Crop Production in Sub-Saharan Africa*. Doctoral dissertation (University of Otago).
4. Bozzola, M.,(2023). *Impacts of climate change on global agri-food trade*. Journal / Working Paper.
5. Chen, P.,(2023). *Improving yield prediction based on spatio-temporal deep learning*. Science / Journal.
6. 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*. ACM Digital Library, pp. 785–794. Available at: <https://dl.acm.org/doi/10.1145/2939672.2939785> (Accessed: 16 October 2025).
7. Chen, W., Zhao, L., & Liu, T. (2023). *Explaining climate-variable influence on food export earnings using LightGBM and SHAP*. Environmental Modelling & Software.
8. Cobb, C., & Douglas, P. (1928). *A theory of production*. American Economic Review.
9. Dzvimbo, M. A., Monga, M., & Mashizha, T. M. (2017). *The link between rural institutions and rural development: Reflections on smallholder farmers and donors in Zimbabwe*. Journal of Humanities and Social Science, 22, 46–53.
10. FAO. (2020). *The state of food and agriculture*. Rome; FAO
11. FAO. (2020). *Zimbabwe Drought Impact Report*. Food and Agriculture Organization of the United Nations.
12. Friedman, J. (2001). *Greedy function approximation*. Annals of Statistics.
13. Fourier, J. (1824). *Remarques générales sur les températures du globe terrestre*. Annales de Chimie et de Physique.
14. Houghton, J. (2015). *Climate Change: The Complete Briefing* (5th ed.). Cambridge University Press.
15. Iglewicz, B., & Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.

16. IPCC 2007, *Climate Change 2007: impacts, adaptation and vulnerability: contribution of Working Group II to the fourth assessment report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, UK.
17. IPCC. (2021). *climate change 2021: The physical science basics*. Cambridge University Press.
18. Isik, M & Devadoss, S 2006, 'An analysis of the impact of climate change on crop yields and yield variability', *Applied Economics*, vol. 38, no. 07, pp. 835–44.
19. Jardon, M. I., & Mitchell, T. M. (2015). *Machine learning: Trends, perspectives, and prospects*. *Science*, 349(6245), 255-260.
20. Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., et al. (2017). *Brief history of agricultural systems modeling*. *Agric. Syst.* 155, 240–254. doi: 10.1016/j.agsy.2016.05.014
21. Joshi, A., Pradhan, B., Chakraborty, S., Varatharajoo, R., Gite, S., & Alamri, A. (2024). *Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data*. *Remote Sensing*.
22. .Kang, L., Hu, G., Huang, H., Lu, W. & Liu, L. (2022). *Urban Traffic Travel Time Short-Term Prediction Model Based on Spatio-Temporal Feature Extraction*. *Journal of Advanced Transportation*, 2020, 1–16.
23. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. *Advances in Neural Information Processing Systems*.
24. Kumar, R., & Singh, P. (2022). *Forecasting food production using LightGBM under climate variability*. *Agricultural Systems*.
25. Lobell D., Schlenker, W. and Costa-Roberts, J. (2011). *Climate trends and global crop production since 1980*. *Science*, 33:616-620
26. Mahofa, G. (2022). *Climate Change and Agricultural Trade in Sub-Saharan Africa*. AERC / Policy Paper. (publication.aercafricalibrary.org)
27. Mamboundou, P. (2023). *Impact of Climate Change on Trade in Africa*. CGIAR / CGSpace. (CGSpace)
28. Molua, EL 2002, 'Climate variability, vulnerability and effectiveness of farm-level adaptation options: the challenges and implications for food security in South-western Cameroon', *Environment and Development Economics*, vol. 7, pp. 529–45.
29. O'Brien, R. M. (2007). *A caution regarding rules of thumb for variance inflation factors*. *Quality & Quantity*, 41(5), 673–690.

30. Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G., & Lobell, D. B. (2020). *The Historical Impact of Anthropogenic Climate Change on Global Agricultural Productivity*.
31. Reid, S, Smit, B, Caldwell, W & Belliveau, S 2007, 'Vulnerability and adaptation to climate risks in Ontario agriculture', *Mitigation and Adaptation Strategies for Global Change*, vol. 12, no. 4, pp. 609–37.
32. Ren, Y., Ling, F. and Wang, Y., 2023. *Research on provincial-level soil moisture prediction based on extreme gradient boosting model*. *Agriculture*, 13(5), p.927.
33. Ricardo, D. (1817). *Principles of Political Economy and Taxation*.
34. Salvatore, D. (2019). *International Economics*. John Wiley & Sons.
35. Schultz, T. (1964). *Transforming Traditional Agriculture*. Yale University Press.
36. Semosa, P. D., ... (2025). *The Impact of Climate Change on the Agricultural Sector in SADC*. Sustainability. (MDPI)
37. Vink (2012). *Declining global per capita agricultural production and warming oceans threaten food security*. *Food security*, 1(3),271-289
38. Wibisono, N. B., & Saiful, S. (2025). *Crop Yield Prediction Using Random Forest Algorithm and XGBoost Machine Learning Model*. *International Journal of Research and Innovation in Social Science*. (RSIS International)
39. WB 2010, *World Development Report*, Washington, Washington DC.
40. World Bank. (2022). *World development indicators*. Washington, DC: World Bank
41. World Bank. (2022). *Zimbabwe Economic Update: Accelerating Growth and Improving Living Standards*.
42. ZimStat. (2016). *Agricultural and Livestock Survey Report*. Zimbabwe National Statistics Agency
43. Zhang, J. (2020). *Machine Learning in Economic Analysis*. New York: Academic Press.
44. Zhang, Y., Li, H., & Wang, J. (2020). *Application of LightGBM in agricultural yield prediction*. *Journal of Agricultural Informatics*.
45. Zhao, H., Wang, Y., Li, X., Guo, P. & Lin, H. (2023). *Prediction of Maximum Tunnel Uplift Caused by Overlying Excavation Using XGBoost Algorithm with Bayesian Optimization*. *Applied Sciences*, 13(17), 9726.

Appendix A: Transformed Data (FP & FE only transformed)

YR	PR	HU	TMP	FP	FE
1967	702.94	60.47	20.71	1.46275	1.847496
1968	547.94	55.12	20.92	1.433207	1.851004
1969	899.31	62.5	20.68	1.476809	1.859625
1970	467.29	54.26	21.52	1.472956	1.864964
1971	631.05	56.07	21.02	1.523313	1.904199
1972	752.32	58.95	20.55	1.552043	1.876974
1973	739.59	58.46	20.93	1.512563	1.883108
1974	981.59	64.57	19.86	1.545832	1.889926
1975	703.89	60.97	20.2	1.531795	1.899771
1976	813.51	63.83	20.04	1.551272	1.905728
1977	930.96	61.87	20.78	1.55281	1.916405
1978	1014.12	65.23	20.24	1.549043	1.908985
1979	608.73	60.32	20.58	1.520945	1.928122
1980	688.73	58.87	20.73	1.519473	1.887987
1981	852.63	62.82	20.13	1.558782	1.940756
1982	479.51	57.03	21.19	1.539449	1.868851
1983	498.09	54.3	22.08	1.501608	1.95385
1984	588.22	58.56	21.24	1.518298	1.960162
1985	802.46	60.48	20.7	1.576713	1.965431
1986	722.17	60.41	20.73	1.568237	1.970886
1987	501.78	53.43	22.06	1.534421	1.975883
1988	690.26	60.23	21.06	1.582955	1.980496
1989	673.31	59.13	21.14	1.568983	1.985149
1990	668.12	57.45	21.51	1.575015	1.988998
1991	547.59	55.81	21.43	1.711588	1.851984
1992	480.4	50.08	22.42	1.696903	1.839933
1993	701.09	58.42	21.49	1.680725	1.996314
1994	489.63	52.41	21.25	1.693855	2.000867
1995	476.62	54.26	22.05	1.679042	2.002872
1996	742.53	59.93	21.02	1.719731	2.001256
1997	841.5	62.11	20.98	1.718573	2.001815
1998	700.69	58.27	21.64	1.721569	1.998358
1999	720.07	60.67	20.96	1.733848	1.997694
2000	913.5	65.45	20.63	1.739366	2.000271
2001	895.46	62.94	21.12	1.739815	2.003699
2002	447.21	56.06	21.72	1.713106	1.949952
2003	618.13	57.98	21.53	1.717183	1.950192
2004	781.12	60.33	21.17	1.729431	2.009712
2005	608.36	54.46	22.25	1.718148	1.930782
2006	756.21	59.64	21.3	1.720079	1.923509
2007	709.09	57.82	21.36	1.719896	1.908491
2008	597.52	57.61	21.27	1.704895	1.767786
2009	680.87	61.29	21.02	1.715892	1.934884
2010	717.43	61.19	21.8	1.736099	1.989922
2011	766.52	57.75	21.48	1.73943	2.038594
2012	613.49	54.89	21.79	1.745584	2.064543
2013	670.76	57.38	21.58	1.744795	2.075174
2014	639.56	59.18	21.45	1.719804	2.070739
2015	440.41	54.97	22.18	1.72739	2.07165
2016	561.1	56.4	22.34	1.729035	2.068446
2017	949.46	63.04	21.23	1.736657	2.055153
2018	635.98	57.16	21.92	1.738657	2.062651
2019	728.85	53.95	22.43	1.735259	2.056935
2020	702.14	58.4	21.79	1.742672	2.053994
2021	635.55	59.5	21.57	1.757872	2.051917
2022	723.12	60.44	21.59	1.759508	2.037621
2023	701.61	57.12	22.34	1.774763	2.032167
2024	627.19	55.16	22.53	1.771465	2.026178
2025	750	59.5	21.87	1.773341	2.020905

Appendix B: Feature Engineering

B.1.Training set

YR	PR	HU	TMP	FP	FE	FE_lag1	FE_lag2	FP_lag1	FP_lag2	FE_rol13	FP_rol13	PR_TMP	TMP_FP	Year_index
1969	0.799881	62.5	0.307116	0.127658	0.298773	0.270726	0.259312	0	0.086495	0.27627	0.071384	0.245656	0.039206	3
1970	0.046853	54.26	0.621723	0.116376	0.316142	0.298773	0.270726	0.127658	0	0.295214	0.081345	0.02913	0.072354	4
1971	0.332293	56.07	0.434457	0.263812	0.443782	0.316142	0.298773	0.116376	0.127658	0.352899	0.169282	0.144367	0.114615	5
1972	0.543672	58.95	0.258427	0.347925	0.355213	0.443782	0.316142	0.263812	0.116376	0.371712	0.242704	0.140499	0.089913	6
1973	0.521483	58.46	0.400749	0.232337	0.375165	0.355213	0.443782	0.347925	0.263812	0.391387	0.281358	0.208984	0.093109	7
1974	0.943299	64.57	0	0.329742	0.397347	0.375165	0.355213	0.232337	0.347925	0.375908	0.303335	0	0	8
1975	0.459256	60.97	0.127341	0.288643	0.429375	0.397347	0.375165	0.329742	0.232337	0.400629	0.283574	0.058482	0.036756	9
1977	0.855049	61.87	0.344569	0.350172	0.483491	0.448756	0.429375	0.345668	0.288643	0.453874	0.328161	0.294624	0.120658	11
1979	0.293389	60.32	0.269663	0.256879	0.521609	0.459349	0.483491	0.339143	0.350172	0.48815	0.315398	0.079116	0.069271	13
1980	0.432832	58.87	0.325843	0.252567	0.39104	0.521609	0.459349	0.256879	0.339143	0.457333	0.282863	0.141035	0.082297	14
1981	0.718516	62.82	0.101124	0.367655	0.56271	0.39104	0.521609	0.252567	0.256879	0.491787	0.292367	0.072659	0.037179	15
1982	0.068153	57.03	0.498127	0.311054	0.328787	0.56271	0.39104	0.367655	0.252567	0.427512	0.310425	0.033949	0.154944	16
1984	0.257639	58.56	0.516854	0.249127	0.62584	0.605307	0.328787	0.200264	0.311054	0.519978	0.253482	0.133162	0.128762	18
1985	0.631068	60.48	0.314607	0.420153	0.642981	0.62584	0.605307	0.249127	0.200264	0.624709	0.289848	0.198538	0.132183	19
1986	0.491119	60.41	0.325843	0.395338	0.660728	0.642981	0.62584	0.420153	0.249127	0.643183	0.354873	0.160028	0.128818	20
1988	0.435499	60.23	0.449438	0.438429	0.691993	0.676985	0.660728	0.296332	0.395338	0.676569	0.3767	0.19573	0.197047	22
1989	0.405954	59.13	0.479401	0.397522	0.70713	0.691993	0.676985	0.438429	0.296332	0.692036	0.377428	0.194615	0.190572	23
1990	0.396908	57.45	0.617978	0.415182	0.719651	0.70713	0.691993	0.397522	0.438429	0.706258	0.417044	0.24528	0.256573	24
1991	0.186819	55.81	0.588015	0.815037	0.273914	0.719651	0.70713	0.415182	0.397522	0.566898	0.54258	0.109852	0.479254	25
1993	0.454376	58.42	0.610487	0.724678	0.743453	0.234708	0.273914	0.772042	0.815037	0.417358	0.770585	0.277391	0.442406	27
1994	0.085792	52.41	0.520599	0.763119	0.758262	0.743453	0.234708	0.724678	0.772042	0.578808	0.753279	0.044663	0.397279	28
1995	0.063116	54.26	0.820225	0.719751	0.764786	0.758262	0.743453	0.763119	0.724678	0.7555	0.735849	0.051769	0.590357	29
1997	0.699116	62.11	0.419476	0.835488	0.761348	0.75953	0.764786	0.838878	0.719751	0.761888	0.798039	0.293262	0.350467	31
1999	0.487459	60.67	0.411985	0.880209	0.747942	0.7501	0.761348	0.844261	0.835488	0.75313	0.853319	0.200826	0.362633	33
2000	0.824615	65.45	0.28839	0.896364	0.756326	0.747942	0.7501	0.880209	0.844261	0.751456	0.873611	0.23781	0.258502	34
2001	0.793171	62.94	0.47191	0.89768	0.767476	0.756326	0.747942	0.896364	0.880209	0.757248	0.891418	0.374305	0.423624	35
2002	0.011853	56.06	0.696629	0.819483	0.592626	0.767476	0.756326	0.89768	0.896364	0.705476	0.871175	0.008257	0.570876	36
2003	0.309773	57.98	0.625468	0.831417	0.593406	0.592626	0.767476	0.819483	0.89768	0.651169	0.849527	0.193753	0.520025	37
2004	0.593871	60.33	0.490637	0.867277	0.78704	0.593406	0.592626	0.831417	0.819483	0.65769	0.839392	0.291375	0.425518	38
2005	0.292744	54.46	0.895131	0.834243	0.530261	0.78704	0.593406	0.867277	0.831417	0.636902	0.844313	0.262044	0.746757	39
2006	0.550452	59.64	0.539326	0.839896	0.5066	0.530261	0.78704	0.834243	0.867277	0.607967	0.847139	0.296873	0.452978	40
2007	0.46832	57.82	0.561798	0.839361	0.457744	0.5066	0.530261	0.839896	0.834243	0.498202	0.837833	0.263101	0.471551	41
2008	0.273849	57.61	0.52809	0.795442	0	0.457744	0.5066	0.839361	0.839896	0.321448	0.8249	0.144617	0.420065	42
2009	0.419132	61.29	0.434457	0.827639	0.543605	0	0.457744	0.795442	0.839361	0.333783	0.820814	0.182095	0.359574	43
2010	0.482857	61.19	0.726592	0.886801	0.722656	0.543605	0	0.827639	0.795442	0.422087	0.836627	0.35084	0.644342	44
2011	0.568423	57.75	0.606742	0.896552	0.880997	0.722656	0.543605	0.886801	0.827639	0.715752	0.870331	0.344886	0.543975	45
2012	0.301686	54.89	0.722846	0.914569	0.965417	0.880997	0.722656	0.896552	0.886801	0.856357	0.899307	0.218072	0.661093	46
2013	0.401509	57.38	0.644195	0.912261	1	0.965417	0.880997	0.914569	0.896552	0.948805	0.907794	0.25865	0.587674	47
2015	0	54.97	0.868914	0.861303	0.988535	0.985572	1	0.839093	0.912261	0.991369	0.870885	0	0.748398	49
2016	0.210368	56.4	0.928839	0.866119	0.978112	0.988535	0.985572	0.861303	0.839093	0.984073	0.855505	0.195398	0.804485	50
2017	0.887295	63.04	0.513109	0.888433	0.934869	0.978112	0.988535	0.866119	0.861303	0.967172	0.871952	0.455279	0.455863	51
2018	0.340887	57.16	0.771536	0.894288	0.959261	0.934869	0.978112	0.888433	0.866119	0.957414	0.882947	0.263006	0.689975	52
2019	0.502763	53.95	0.962547	0.884341	0.940665	0.959261	0.934869	0.894288	0.888433	0.944932	0.889021	0.483933	0.851219	53
2020	0.456206	58.4	0.722846	0.906045	0.931098	0.940665	0.959261	0.884341	0.894288	0.943675	0.894891	0.329767	0.654931	54
2021	0.340137	59.5	0.640449	0.950545	0.92434	0.931098	0.940665	0.906045	0.884341	0.932034	0.913644	0.217841	0.608776	55
2022	0.492775	60.44	0.64794	0.955337	0.877833	0.92434	0.931098	0.950545	0.906045	0.91109	0.937309	0.319289	0.619001	56
2023	0.455282	57.12	0.928839	1	0.86009	0.877833	0.92434	0.955337	0.950545	0.887421	0.968627	0.422884	0.928839	57
2025	0.539628	59.5	0.752809	0.995836	0.823451	0.840606	0.86009	0.990345	1	0.841382	0.995393	0.406237	0.749674	59

B.2. Testing set

YR	PR	HU	TMP	FP	FE	FE_lag1	FE_lag2	FP_lag1	FP_lag2	FE_rol13	FP_rol13	PR_TMP	TMP_FP	Year_index
1976	0.650329	63.83	0.067416	0.345668	0.448756	0.429375	0.397347	0.288643	0.329742	0.425159	0.321351	0.043842	0.023303	10
1978	1	65.23	0.142322	0.339143	0.459349	0.483491	0.448756	0.350172	0.345668	0.463865	0.344994	0.142322	0.048268	12
1983	0.100539	54.3	0.831461	0.200264	0.605307	0.328787	0.56271	0.311054	0.367655	0.498934	0.292991	0.083594	0.166512	17
1987	0.10697	53.43	0.82397	0.296332	0.676985	0.660728	0.642981	0.395338	0.420153	0.660231	0.370608	0.08814	0.244169	21
1992	0.069704	50.08	0.958801	0.772042	0.234708	0.273914	0.719651	0.815037	0.415182	0.409424	0.66742	0.066832	0.740235	26
1996	0.526608	59.93	0.434457	0.838878	0.75953	0.764786	0.758262	0.719751	0.763119	0.760859	0.773916	0.228788	0.364457	30
1998	0.453679	58.27	0.666667	0.844261	0.7501	0.761348	0.75953	0.835488	0.838878	0.756993	0.839542	0.302452	0.56284	32
2014	0.347127	59.18	0.595506	0.839093	0.985572	1	0.965417	0.912261	0.914569	0.983663	0.888641	0.206716	0.499684	48
2024	0.325565	55.16	1	0.990345	0.840606	0.86009	0.877833	1	0.955337	0.85951	0.981894	0.325565	0.990345	58

Appendix C: Syntax of R Codes used to run the models.

C.1: Introduction

This appendix presents the R code used in this study to investigate the impact of climate change and food production on food products export earnings in Zimbabwe (1967-2025). The code covers the key stages of the analysis, including data preprocessing, feature engineering, model development, training, prediction, and performance evaluation. The machine learning models employed in this study are the XGBoost, and LightGBM algorithms, which were used to model and predict food products export earnings based on climatic and agricultural variables.

The appendix supports the achievement of the study objectives by providing the computational procedures used to examine the relationship between climate change indicators, food production, and export earnings, as well as to evaluate and compare the predictive performance of the selected machine learning models. Performance assessment was conducted using standard evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). By documenting the syntax and analytical procedures, this appendix enhances the transparency, reproducibility, and reliability of the research findings and serves as a practical guide for future studies employing similar machine learning approaches in agricultural and economic forecasting.

C.2 : R Codes

##Load Libraries

```
library(caret)
library(xgboost)
library(lightgbm)
library(dplyr)
library(Metrics)
library(zoo)
```

1. Load Data

```
set.seed(123)
```

```
data
```

Log transform skewed variables (ensure positivity)

```
data$FP <- log1p(data$FP)
```

```
data$FE <- log1p(data$FE)
```

Normalization (min-max)

```
preProc <- preProcess(data[,c("PR", "TMP", "FP", "FE")], method="range")
```

```
data_norm <- predict(preProc, data)
```

Feature engineering

```
data_norm <- data_norm %>%
```

```
  mutate(
```

```
    FE_lag1 = lag(FE,1),
```

```
    FE_lag2 = lag(FE,2),
```

```
    FP_lag1 = lag(FP,1),
```

```
    FP_lag2 = lag(FP,2),
```

```
    FE_roll3 = rollmean(FE,3,fill=NA,align="right"),
```

```
    FP_roll3 = rollmean(FP,3,fill=NA,align="right"),
```

```
    PR_TMP = PR*TMP,
```

```
    TMP_FP = TMP*FP,
```

```
    Year_index = seq_len(nrow(data_norm))
```

```
  )
```

##2. Train-Test Split (80:20)

```
trainIndex <- createDataPartition(data_norm$FE, p=0.8, list=FALSE)
```

```
train <- data_norm[trainIndex,]
```

```
test <- data_norm[-trainIndex,]
```

```
features <- c("PR", "TMP", "FP", "FE_lag1", "FE_lag2", "FP_lag1", "FP_lag2",  
             "FE_roll3", "FP_roll3", "PR_TMP", "TMP_FP", "Year_index")
```

```
X_train <- as.matrix(train[,features, drop=FALSE])
```

```
y_train <- train$FE
```

```
X_test <- as.matrix(test[,features, drop=FALSE])
y_test <- test$FE
```

##XGBoost Model (tuned)

```
dtrain <- xgb.DMatrix(data=X_train, label=y_train)
dtest <- xgb.DMatrix(data=X_test, label=y_test)
```

```
params_xgb <- list(
  objective = "reg:squarederror",
  eta = 0.03,
  max_depth = 8,
  subsample = 0.9,
  colsample_bytree = 0.9,
  min_child_weight = 3,
  gamma = 0.2
)
```

```
xgb_model <- xgb.train(
  params=params_xgb, data=dtrain, nrounds=5000,
  watchlist=list(train=dtrain, eval=dtest),
  early_stopping_rounds=50, seed=123
)
```

```
xgb_pred <- predict(xgb_model, dtest)
```

##LightGBM Model (tuned)

```
dtrain_lgb <- lgb.Dataset(data=X_train, label=y_train)
dtest_lgb <- lgb.Dataset(data=X_test, label=y_test)
```

```
params_lgb <- list(
  objective = "regression",
  metric = "rmse",
  learning_rate = 0.03,
  num_leaves = 128,
```

```

min_data_in_leaf = 20,
feature_fraction = 0.9,
bagging_fraction = 0.9,
bagging_freq = 1
)

lgb_model <- lgb.train(
  params=params_lgb, data=dtrain_lgb, nrounds=5000,
  valids=list(test=dtest_lgb), early_stopping_rounds=50
)

lgb_pred <- predict(lgb_model, X_test)

```

##Performance Evaluation

```

cat("XGBoost RMSE:", rmse(y_test, xgb_pred), "\n")
cat("XGBoost MAE:", mae(y_test, xgb_pred), "\n")
cat("XGBoost R2:", R2(xgb_pred, y_test), "\n")

```

```

cat("LightGBM RMSE:", rmse(y_test, lgb_pred), "\n")
cat("LightGBM MAE:", mae(y_test, lgb_pred), "\n")
cat("LightGBM R2:", R2(lgb_pred, y_test), "\n")

```

##Visualization (Actual vs Predicted)

```

par(mfrow=c(1,1))

```

```

# XGBoost

```

```

plot(y_test, type="l", col="black", lwd=2, main="XGBoost vs Actual")
lines(xgb_pred, col="red", lwd=2)
legend("topleft", legend=c("Actual", "XGBoost"), col=c("black", "red"), lty=1, lwd=2,
      bty="n")

```

```

# LightGBM

```

```

plot(y_test, type="l", col="black", lwd=2, main="LightGBM vs Actual")
lines(lgb_pred, col="blue", lwd=2)

```

```
legend("topleft", legend=c("Actual","LightGBM"), col=c("black","blue"), lty=1, lwd=2,
      bty="n")
```

```
# Combined
```

```
matplot(1:length(y_test), cbind(y_test, xgb_pred, lgb_pred, ensemble_pred),
      type="l", lwd=2, lty=c(1,1,1,2), col=c("black","red","blue","green"),
      main="Comparison vs Actual", xlab="Index", ylab="FE (log-transformed)")
legend("topleft", legend=c("Actual","XGBoost","LightGBM","Ensemble"),
      col=c("black","red","blue","green"), lty=c(1,1,1,2), lwd=2, bty="n")
```

##Feature Importance

```
xgb_imp <- xgb.importance(model=xgb_model)
xgb.plot.importance(xgb_imp, top_n=10, main="XGBoost Feature Importance")
```

```
# XGBoost Feature Importance
```

```
xgb_imp <- xgb.importance(model=xgb_model)
xgb.plot.importance(xgb_imp, top_n=10, main="XGBoost Feature Importance")
```

```
# LightGBM Feature Importance
```

```
lgb_imp <- lgb.importance(lgb_model, percentage=TRUE)
lgb.plot.importance(lgb_imp, top_n=10)
title("LightGBM Feature Importance")
```

```
lgb_imp <- lgb.importance(lgb_model, percentage=TRUE)
lgb.plot.importance(lgb_imp, top_n=10, main="LightGBM Feature Importance")
```

##Residual Analysis

```
# Residuals
```

```
res_xgb <- y_test - xgb_pred
res_lgb <- y_test - lgb_pred
```

```
# --- Residual Histograms (Normality) ---
```

```
par(mfrow=c(1,2))
hist(res_xgb, breaks=30, col="red", main="Residual Histogram: XGBoost", xlab="Residuals")
```

```

mtext("Bell-shaped → Normality; Skewed → Non-normal", side=3, line=0.5, cex=0.7)
hist(res_lgb, breaks=30, col="blue", main="Residual Histogram: LightGBM",
xlab="Residuals")
mtext("Bell-shaped → Normality; Skewed → Non-normal", side=3, line=0.5, cex=0.7)

# --- Residuals vs Fitted (Homoscedasticity) ---
par(mfrow=c(1,2))
plot(xgb_pred, res_xgb, col="red", pch=16,
     main="Residuals vs Fitted (XGBoost)", xlab="Fitted Values", ylab="Residuals")
abline(h=0, lty=2)
mtext("Random scatter → Constant variance; Pattern → Heteroscedasticity", side=3, line=0.5,
cex=0.7)

plot(lgb_pred, res_lgb, col="blue", pch=16,
     main="Residuals vs Fitted (LightGBM)", xlab="Fitted Values", ylab="Residuals")
abline(h=0, lty=2)
mtext("Random scatter → Constant variance; Pattern → Heteroscedasticity", side=3, line=0.5,
cex=0.7)

# --- QQ Plots (Normality Check) ---
par(mfrow=c(1,2))
qqnorm(res_xgb, main="QQ Plot Residuals: XGBoost", col="red"); qqline(res_xgb)
mtext("Points near line → Normality; Deviations → Skew/Kurtosis", side=3, line=0.5,
cex=0.7)

qqnorm(res_lgb, main="QQ Plot Residuals: LightGBM", col="blue"); qqline(res_lgb)
mtext("Points near line → Normality; Deviations → Skew/Kurtosis", side=3, line=0.5,
cex=0.7)

# --- Residual Time-Series Plot (Randomness over time/index) ---
par(mfrow=c(1,2))
plot(res_xgb, type="l", col="red", main="Residuals over Index: XGBoost",
     xlab="Index", ylab="Residuals"); abline(h=0, lty=2)
mtext("Random noise → Independence; Trend → Autocorrelation", side=3, line=0.5, cex=0.7)

```

```
plot(res_lgb, type="l", col="blue", main="Residuals over Index: LightGBM",  
      xlab="Index", ylab="Residuals"); abline(h=0, lty=2)  
mtext("Random noise → Independence; Trend → Autocorrelation", side=3, line=0.5, cex=0.7)
```