

https://dspace.org/

Applying Machine Learning to Model South Africa's Equity Market Index Price Performance

Nokeri, Tshepo Chris

2023-07
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/41753
Nokeri, Tshepo Chris. (2023). Applying Machine Learning to Model South Africa's Equity Market Index Price Performance. [Master's dissertation, University of the Witwatersrand, Johannesburg]. https://hdl.handle.net/10539/41753

Downloaded from a DSpace repository. https://dspace.org/

UNIVERSITY OF THE WITWATERSRAND

M.SC. COMPUTER SCIENCE BY DISSERTATION

Applying Machine Learning to Model South Africa's Equity Market Index Price Performance

Author:
Tshepo Chris NOKERI

Supervisors:
Dr. Ritesh AJOODHA and Mr. Rudzani MULAUDZI



A thesis submitted in fulfillment of the requirements for the degree of M.Sc. Computer Science by Dissertation

in the

School of Computer Science and Applied Mathematics

July 14, 2023

Declaration of Authorship

I, Tshepo Chris NOKERI, declare that this thesis titled, "Applying Machine Learning to Model South Africa's Equity Market Index Price Performance" and the work presented in it, are my own. I confirm that:

- This work was done wholly or mainly done while in candidature for the M.Sc. Computer Science by Dissertation degree in the School of Computer Science and Applied Mathematics at the University of the Witwatersrand.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at the University of the Witwatersrand or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Except for such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signature:

T.C. Nokeri

Date: July 14, 2023

UNIVERSITY OF THE WITWATERSRAND

Abstract

Faculty of Science

School of Computer Science and Applied Mathematics

M.Sc. Computer Science by Dissertation

Applying Machine Learning to Model South Africa's Equity Market Index Price Performance

by Tshepo Chris NOKERI

Policymakers typically use statistical multivariate forecasting models to forecast the reaction of stock market returns to changing economic activities. However, these models frequently result in subpar performance due to inflexibility and incompetence in modeling non-linear relationships. Emerging research suggests that machine learning models can better handle data from non-linear dynamic systems and yield outstanding model performance. This research compared the performance of machine learning models to the performance of the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns to changing economic activities in South Africa. The vector autoregressive model was used to forecast the reaction of stock market returns. It achieved a *mean absolute percentage error (MAPE)* value of 0.0084. Machine learning models were used to forecast the reaction of stock market returns. The lowest *MAPE* value was 0.0051. The machine learning model trained on low economic data dimensions performed 65% better than the benchmark model. Machine learning models also identified key economic activities when forecasting the reaction of stock market returns. Most research focused on whole features, few models for comparison, and barely focused on how different feature subsets and reduced dimensionality change model performance, a limitation this research addresses when considering the number of experiments. This research considered various experiments, i.e., different feature subsets and data dimensions, to determine whether machine learning models perform better than the benchmark model when forecasting the reaction of stock market returns to changing economic activities in South Africa.

Contents

Declaration of Authorship ii

Abstract iii

1 Introduction to the Research 1
- 1.1 Introduction to the Research . . . 1
- 1.2 Problem Statement . . . 2
- 1.3 Purpose Statement . . . 2
- 1.4 Research Questions . . . 3
- 1.5 Research Contributions . . . 3
- 1.6 Research Motivation . . . 3
- 1.7 Research Structure . . . 4

2 Literature Review 6
- 2.1 The Stock Market in South Africa . . . 7
  - 2.1.1 The Financial Times Stock Exchange/Johannesburg Stock Exchange All-Share Index . . . 7
- 2.2 The use of Asset Pricing Models in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities . . . 8
  - 2.2.1 The Capital Asset Pricing Model . . . 8
  - 2.2.2 The Arbitrage Pricing Model . . . 8
  - 2.2.3 The Multi-Factor Model . . . 9
  - The Estimation of Alpha and Systematic Risk Factors . . . 9
  - The Estimation of Cumulative Systematic Risk Factors . . . 9
  - 2.2.4 The Modern Portfolio Model . . . 10
  - 2.2.5 The Drawbacks of Asset Pricing Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities . . . 11
  - 2.2.6 The Rationale for Comparing the Performance of Statistical and Machine Learning Models . . . 12
- 2.3 The use of Conventional Statistical Models in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities . . . 12
  - 2.3.1 Statistical Univariate Forecasting Models . . . 12
  - The Autoregressive Integrated Moving Average Model . . . 12
  - The Seasonal Autoregressive Integrated Moving Average Model . . . 13
  - 2.3.2 The Statistical Multivariate Forecasting Model . . . 13

The Vector Autoregressive Model . . . 13
2.3.3 The Drawbacks of Conventional Statistical Forecasting Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities . . . 14
2.4 State of Literature . . . 15
2.4.1 Related Research . . . 15
2.5 Research Gaps . . . 20
2.6 The use of Machine Learning Models in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities . . . 20
2.6.1 The Ordinary Least-Squares Regression Model . . . 20
2.6.2 The Decision Tree Model . . . 21
2.6.3 The Random Forest Tree Model . . . 21
2.6.4 The Extreme Gradient Boosting Tree Model . . . 22
2.7 The use of Neural Networks in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities . . . 22
2.7.1 The Recurrent Neural Network . . . 23
2.7.2 The Gated Recurrent Unit Network . . . 23
2.7.3 The Long-Short Term Memory . . . 24
2.7.4 The Restricted Boltzmann Machine . . . 24
2.7.5 The Multi-Layer Perceptron . . . 25
2.8 The Activation Function . . . 25
2.8.1 The Rectified Linear Unit Activation Function . . . 25

3 Research Methods **27**
3.1 The Acquisition of Economic and Stock Market Data of South Africa . . . 27
3.1.1 Sources of Economic and Stock Market Data of South Africa . . . 28
3.1.2 Properties of Economic and Stock Market Data of South Africa . . . 28
3.1.3 Strategies used to Cleanse and Preprocess Data Before Forecasting . . . 30
3.2 The Strategy used to Impute Missing Economic Data of South Africa . . . 31
3.2.1 The *k* Nearest Neighbor Imputation Strategy . . . 31
3.3 The Strategy used to Replace Outliers in Economic and Stock Market Returns Data of South Africa . . . 32
3.3.1 The Median Outlier Replacement Strategy . . . 32
3.4 The Strategy used to Reduce Economic Data Dimensions of South Africa . . . 32
3.4.1 The Principal Components Analysis Method . . . 32
3.5 The Strategy used to Partition Economic and Stock Market Returns Data of South Africa . . . 33
3.5.1 The Random Data Partitioning Strategy . . . 33
3.6 The Strategy used to Scale Economic Data of South Africa . . . 33
3.6.1 The Data Standardization Scaling Strategy . . . 34
3.7 Experiments for Model Comparison . . . 34
3.8 Strategies used to Regularize Regression Models . . . 36
3.8.1 The Ridge Model Regularization Strategy . . . 36
3.8.2 The Least Absolute Shrinkage and Selection Operator Model Regularization Strategy . . . 36
3.8.3 The Elastic Net Model Regularization Strategy . . . 37
3.9 The Metric used to Evaluate the Performance of Models . . . 38
3.9.1 The Mean Absolute Percentage Error Metric . . . 38
3.10 Ethical Consideration . . . 38
3.11 Research Methods Summary . . . 39

4 Experiment Results & Discussions **40**

4.1 The Exploration of the Distribution of Stock Market Returns Data in South Africa . . . 40
4.1.1 The Granger-Causal Relationship between Economic Activities and Stock Market Returns in South Africa . . . 41
4.2 The Benchmark for Model Comparison . . . 42
4.2.1 The Performance of the Vector Autoregressive Model when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 42
4.3 The Performance of Default Machine Learning Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 43
4.3.1 The Performance of Default Regression Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 43
4.3.2 The Performance of Default Tree-Based Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 43
4.3.3 The Performance of Default Neural Networks when Forecasting Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 44
4.4 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by Tree-based Models . . . 46
4.4.1 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by the Decision Tree Model . . . 46
4.4.2 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by the Random Forest Tree Model . . . 46
4.4.3 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by the Extreme Gradient Boosting Tree Model . . . 47
4.4.4 The Performance of Models Trained on a Feature Subset Containing Key Economic Features when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 48
4.5 The Reduction of Economic Data Dimensions . . . 49
4.5.1 Economic Features of South Africa in Different Dimensions . . . 49
4.5.2 An Index that Offers Insight into the Structure of the Economy in South Africa . . . 50
4.5.3 The Performance of Models Trained on Low Economic Data Dimensions when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa . . . 51
4.6 Overall Performance Summary . . . 52

5 Conclusions & Future Research **55**
5.1 Introduction . . . 55
5.2 Literature Review . . . 55
5.3 Research Methods . . . 55
5.4 Experimental Results . . . 56
5.5 Policy Recommendations . . . 57
5.6 Learned Model Application . . . 58
5.7 Future Research . . . 58

A A List of Economic Features used Forecast the Reaction of Stock Market Returns in South Africa **59**

B A Full Index that Offers Insight Into the Structure of the Economy in South Africa **69**

C Coefficients of the Optimal Model **73**

Bibliography **vii**

List of Figures

2.1 The reaction of stock market returns to changing economic activities, i.e., (A) economic activities, (B) international economic activities, (C) money and banking activities, (D) capital markets activities, and (E) national government finance activities 6

3.1 The modeling pipeline exhibiting the workflow that produces models that forecast the reaction of stock market returns to changing economic activities . 27

4.1 The distribution of stock market returns data in South Africa 40
4.2 The *loss* function value across epochs of default neural networks when forecasting the reaction of stock market returns to changing economic activities in South Africa 45
4.3 The top ten economic features in a feature subset selected by the decision tree model, as measured by the *gini impurity* value 46
4.4 The top ten economic features in a feature subset selected by the random forest tree model, as measured by the *gini impurity* value 47
4.5 The top ten economic features in a feature subset selected by the extreme gradient boosting tree model, as measured by the *gini impurity* value 47
4.6 Economic features of South Africa in two dimensions 49
4.7 Economic features of South Africa in three dimensions 50
4.8 The top performance of each machine learning model over four periods . . . 52
4.9 The top performance of each tree-based feature selection strategy 54
4.10 The learning curve of the optimal model that forecasts the reaction of stock market returns to changing economic activities in South Africa 54

List of Tables

1.1 Research structure 4

2.1 A summary of used economic features and their categories 7
2.2 The drawbacks of asset pricing models when forecasting the reaction of stock market returns to changing economic activities 11
2.3 The drawbacks of conventional statistical forecasting models when forecasting the reaction of stock market returns to changing economic activities . . . 14
2.4 Related research 15

3.1 Sources of economic and stock market data of South Africa, along with the description of data and the sample period 28
3.2 The properties of economic and stock market data of South Africa, and commentary on how those properties affect models that forecast the reaction of stock market returns to changing economic activities, along with strategies for addressing non-adherence to model requirements 29
3.3 Strategies used to cleanse and preprocess data strategies before forecasting the reaction of stock market returns to changing economic activities in South Africa 30
3.4 Experiments considered when determining whether machine learning models perform better than the benchmark model when forecasting the reaction of stock market returns to changing economic activities in South Africa 34
3.5 Research methods summary 39

4.1 The descriptive statistics of stock market returns data in South Africa 41
4.2 The Granger-Causality relationship between selected economic activities and the stock market returns in South Africa 41
4.3 Selected economic features in the Granger-Causality Matrix 42
4.4 The performance of the vector autoregressive model when forecasting the reaction of stock market returns to changing economic activities in South Africa 42
4.5 The performance of default regression models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 43

4.6 The performance of default tree-based models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 44
4.7 The performance of default neural networks when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 44
4.8 The performance of models trained on a feature subset containing key economic features when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 48
4.9 An index that provides insight into economic activities in South Africa 50
4.10 The performance of models trained on low economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 51
4.11 The top ten highest-performing machine learning models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 52
4.12 The top ten worst-performing machine learning models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$ 53
4.13 The ranking of economic features based on the *gini impurity* value calculated by each tree-based model 53

5.1 Research questions, various experiments, and research findings 56

A.1 A list of economic features used forecast the reaction of stock market returns in South Africa 59

B.1 A full index that offers insight into the structure of the economy in South Africa 69

C.1 Coefficients of the the optimal model (the ridge model) 73

1 Introduction to the Research

1.1 Introduction to the Research

The Johannesburg Stock Exchange (JSE) is legitimately regarded as the primary stock market area in Africa, because it was the 17^{th} largest stock exchange in the world in 2017 [SSEInitiative 2021; JSELtd 2021] and the largest stock market exchange in Africa, with a total market value of US$1.36 trillion and 339 listed stocks from various industries [SSEInitiative 2021]. Consequently, forecasting the returns of stocks listed on the JSE is important, but doing so for all stocks remains challenging given the listed stock count.

Information about historical stock market index returns of South Africa is critical for economic policymakers and investors, but it is insufficient for developing judgments, since it lacks present predictions about the future of the stock market index. Hence, policymakers rely on predictive analytical models to model the data creation process and create economic growth predictions, before developing and implementing policies that accomplish strategic economic objectives.

The principal objective of investors is to avert financial loss while maximizing predicted asset returns. However, this objective may be unaccomplished due to economic risk considerations. To attain this objective, investors focus heavily on mitigating unfavorable economic risk factors and capitalizing on favorable economic risk variables. Policymakers also require substantial information to develop policies that eliminate systematic market mistakes, conduct monetary and fiscal interventions for economic stability, and promote equity market-friendly policies.

Previous research on the reaction of stock market returns to changing economic activities in South Africa used conventional asset pricing models and statistical forecasting models that contained methodological difficulties and practical issues, necessitating the use of sophisticated analytical tools like machine learning models. Furthermore, there is little understanding of how machine learning methods compare to traditional forecasting models in anticipating the reaction of stock market returns in the modern economic setting.

This study looks at the dynamics of the economy and how stock market returns respond to changes in economic activity in South Africa. It considers several experimental situations, such as alternative feature sets, data dimensions, and model parameters, to test if machine learning algorithms outperform the benchmark model (the vector autoregressive model) in predicting stock market returns.

This research is valuable for investors in the development phase of investment strategy management, as well as policymakers who want to identify appropriate analytical tools for answering critical practices and policy questions about economic activities and the reaction of stock market returns in South Africa.

1.2 Problem Statement

Policymakers typically use statistical multivariate forecasting models to forecast the reaction of stock market returns to changing economic activities, then use insights from models to structure economic policies that reduce systematic market errors and benefit the stock market [Olayode *et al.* 2021; Saxena and Bhadauriya 2021; Shamsudin *et al.* 2021].

Previous research on the reaction of stock market returns to changing economic activities focused on leading economies like the U.S. economy [Bordo and Jeanne 2002; Stadtmann and Dunsch 2018]. However, research focused on emerging economies has stalled and is currently focused on using conventional forecasting models [Macfarlane 2011; Tsepang Patrick 2013; Chifurira and Chinhamu 2019].

Prominent research used asset pricing models to understand the general equilibrium of exchange [Malamud 2015], some research used conventional statistical univariate forecasting models [Pillay 2020; Chitenderu *et al.* 2014; Makatjane and Moroke 2021], and most research repeatedly used a common statistical multivariate forecasting model, i.e., the vector autoregressive model [Adejayan and Oke 2022], which is the benchmark model in this research due to its widespread adoption.

Forecasting the reaction of stock market return to changing economic activities using statistical multivariate forecasting models remains a challenging undertaking due to inflexibility and incompetence in modeling non-linear relationships [Pratiwi *et al.* 2021]. On the other hand, Fu *et al.* [2018]; Wong *et al.* [2020]; Kamalov [2020], among others, suggested the adoption of machine learning models, because these models can better handle data from non-linear dynamic systems and yield outstanding model performance.

There is currently limited use of machine learning to forecast the reaction of stock market returns in the emerging economic context [Ramos-Pérez *et al.* 2021; Klibanov *et al.* 2021; Makatjane and Moroke 2021]. Furthermore, most research focused on whole features, few models for comparison, and barely focused on how different feature subsets and reduced dimensionality change model performance, a limitation this research addresses when considering the number of experiments.

This research established that machine learning models can be used to forecast the reaction of stock market returns to changing economic activities in the emerging economic context. It advances research by comparing various models trained on different feature subsets and economic data dimensions.

1.3 Purpose Statement

This research investigated the dynamics of the economy and how stock market returns react to changing economic activities in South Africa. It considered various experiments, i.e., different feature subsets and data dimensions, to determine whether machine learning models perform better than the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns. Not only that, but it also produced an index that offers

insight into the structure of the economy in South Africa.

1.4 Research Questions

This research responded to four crucial research questions:

1. How do stock market returns react to changing economic activities in South Africa?
2. Do machine learning models perform better than the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *mean absolute percentage error (MAPE)* metric?
3. Stock market returns react differently to changing economic activities based on the magnitude of change. Do models trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by tree-based models perform better than the models trained on whole features when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?
4. Is the performance of models trained on low economic data dimensions (found by the principal component analysis method) distinguishable from the performance of models trained on high economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?

1.5 Research Contributions

This research makes the following research contributions:

- Determine whether machine learning models perform better than the benchmark model when forecasting the reaction of stock market returns to changing economic activities in South Africa across various experiments, as measured by the *MAPE* metric.
- Present key economic features in South Africa when forecasting the reaction of stock market returns based on the *gini impurity* value calculated by tree-based models.
- Provide insight into the structure of the economy in South Africa by producing an index that ranks economic activities based on the *explained variance ratio* value calculated by the principal components analysis method.
- Show how different feature subsets and data dimensions change the *MAPE* value of models when forecasting the reaction of stock market returns to changing economic activities in South Africa.

1.6 Research Motivation

This study is significant because it analyzes how changes in economic activity affect stock market returns in South Africa. It increases public awareness of the technical measurements used in statistical and machine learning methodologies to forecast stock market index returns. Furthermore, innovative analytical solutions for increasing the performance of stock market index return prediction models.

This thesis contributes to the body of knowledge on stock market returns. Insights into the economic causes of stock market returns may be valuable to policymakers, because they may

better understand how their policies impact the market. Practitioners may also gain insight into the stock market's behavioral and important tendencies for making investment judgments.

1.7 Research Structure

Table 1.1 shows the research structure.

TABLE 1.1: Research structure

Number	Chapter Heading	Chapter Function
1	Introduction	This chapter states the research problem (section 1.2), the research purpose (section 1.3), research questions (section 1.4), and research contributions (section 1.5).
2	Literature Review	This chapter covers the stock market in South Africa (section 2.1), highlights primary asset pricing models for forecasting stock market returns (section 2.2) and statistical univariate and multivariate forecasting models commonly used to forecast stock market returns (subsection 2.3.1 and subsection 2.3.2), then presents the drawbacks of using asset pricing models and statistical forecasting models (subsection 2.2.5 and subsection 2.3.3). This chapter equally refers to the state of literature and related research (section 2.4), then reveals candidate machine learning models for addressing the drawbacks of using asset pricing models and statistical forecasting models (section 2.6, and section 2.7)
3	Research Methods	This chapter covers sources and properties of economic and stock market data of South Africa (subsection 3.1.1 and subsection 3.1.2), then proceeds to specify strategies for data imputation (section 3.2), outlier replacement (section 3.3), data dimension reduction (section 3.4), data partitioning (section 3.5), data scaling (section 3.6), and model regularization (section 3.8). The chapter concludes by revealing the metric used to evaluate the performance of candidate models (section 3.9).

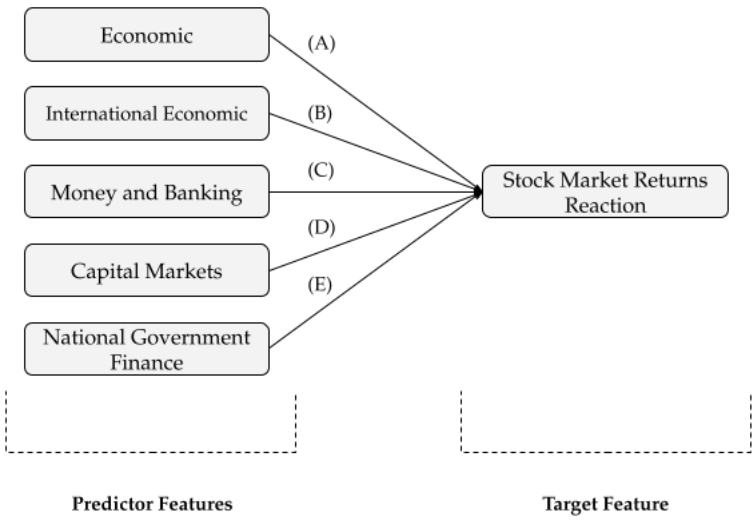
Continued on next page

Table 1.1 – Continued from previous page

Number	Chapter Heading	Chapter Function
4	Experiment Results & Discussions	This chapter shows exploratory descriptive statistical results of stock market returns in South Africa (section 4.1), produces an index that offers insight into the structure of the economy in South Africa (subsection 4.5.2), reports the performance of the benchmark model when forecasting the reaction of stock markets to changing economic activities in South Africa, as measured by the *MAPE* value (section 4.2), then report the performance of machine learning models when forecasting the reaction of stock markets to changing economic activities in South Africa, as measured by the *MAPE* value (section 4.3 and subsection 4.3.3). The last segment of the chapter determines how different feature subsets (subsection 4.4.4) and data dimensions (subsection 4.5.3) change the *MAPE* value of candidate models.
5	Conclusions & Future Work	This chapter recaps the reviewed literature (section 5.2), covers the used research method (section 5.3), summarizes experiment results (section 5.4), provides practice and policy recommendations (section 5.5), details the use of the learned model (section 5.6), and highlights the road-map for forthcoming research (section 5.7).

2 Literature Review

This research adopts the conceptual framework for forecasting the reaction of stock market returns to changing economic activities from previous research by Ahangar *et al.* [2010]; Jasra *et al.* [2012]; Ndikum [2020].



(A)

FIGURE 2.1: The reaction of stock market returns to changing economic activities, i.e., (A) economic activities, (B) international economic activities, (C) money and banking activities, (D) capital markets activities, and (E) national government finance activities

This research differs from previous research on the reaction of stock market returns to changing economic activities, in that it considers various economic features. The reaction of stock market returns in South Africa is the target feature, and economic features are predictor features (Table A.1 shows a thorough list of used economic features).

Table 2.1 shows a summary of used economic features and their categories.

TABLE 2.1: A summary of used economic features and their categories

Economic Feature	Category
Indices: consumer price, producer price, domestic mining and quarrying activities, domestic manufacturing, etc. Rates: interest, currency exchange, yields, bonds, securities, etc. Services: electricity, fuels, gas, supplies, water, etc. Goods: net trade and food, etc.	Economic category
Net average daily turnover, South African Reserve Bank (SARB) gross reserves in foreign currency, gold and other foreign reserves, etc.	International economic category
Money supply, credit, deposits and advances, liabilities, investment treasury bills, short-term credit, return on equity, etc.	Money and banking category
Traded shares value, fixed interest securities market, non-resident transactions, equity derivative markets, etc.	Capital markets category
National government revenue, expenditure, borrowing, financing of net borrowing requirements, etc.	National government finance category

2.1 The Stock Market in South Africa

The Johannesburg Stock Exchange (JSE) is the official stock market exchange of South Africa. It was the 17^{th} largest stock exchange in the world in 2017, as measured by the market capitalization [SSEInitiative 2021; JSELtd 2021]. It is also the largest exchange in Africa, with a total market value of US$1.36 trillion and 339 listed stocks across diverse sectors [SSEInitiative 2021].

It is reasonable to consider South Africa the primary stock market district in Africa. Forecasting returns of stocks listed in the JSE is valuable, but doing so for all stocks remains a challenging task, provided the listed stock count.

This research focuses solely on the reaction of stock market index returns (Financial Times Stock Exchange/Johannesburg Stock Exchange (FTSE/JSE) all-share index returns) to changing economic activities in South Africa.

2.1.1 The Financial Times Stock Exchange/Johannesburg Stock Exchange All-Share Index

The FTSE/JSE all-share index represents the benchmark performance of stocks listed in the JSE for 99.9% of the total market value. The benchmark performance is calculated using the market capitalization-weighted index method. Equation 2.1 defines the market capitalization-weighted index method:

$$MCPWI = w_1 \times p_1 + w_2 \times p_2, ..., w_n \times p_n, \tag{2.1}$$

where w denotes the weight share price and p denotes the share price. Equation 2.2 defines the weight share prices:

$$w_n = \frac{MC_i}{\sum_{i=1}^{n} MC_i}, \tag{2.2}$$

where MC_i denotes the total market capitalization.

2.2 The use of Asset Pricing Models in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities

The adoption of analytical tools for forecasting stock market returns and developing investment strategies is a long-standing phenomenon [Burdenko 2017; Ndikum 2020]. Research on stock market returns in South Africa repeatedly used asset pricing models [Carter *et al.* 2017].

The capital asset pricing model values assets (subsection 2.2.6), then anticipates the future cash flow of assets [Munk 2013]. This model focuses on the general equilibrium of exchange in the market [Malamud 2015].

Alternatives to the capital asset pricing model, i.e., the arbitrage pricing model (subsection 2.2.2) and multi-factor model (subsection 2.2.6), consider systematic economic risk factors, while the modern portfolio model considers portfolio diversification (subsection 2.2.6)

2.2.1 The Capital Asset Pricing Model

The capital asset pricing model identifies the weight of the expected asset return, the asset return as a function of the risk-free return and risk-premium, along with the discounted rate of the net present value [Lintner 1965; Mossin 1966].

Reddy and Thomson [2011] used the capital asset pricing model to explain expected excess stock market returns and determine the linkage between expected stock market returns and the *beta* in South Africa.

2.2.2 The Arbitrage Pricing Model

The arbitrage pricing model uses the linear-oriented framework to identify the extent to which systematic economic risk factors influence the expected asset return [Huberman 2005]. Equation 2.3 defines the arbitrage pricing model:

$$ER(x) = R_f + \beta_1 R_{p1} + \beta_1 R_{p2} \cdot, \beta_n R_{pn}, \tag{2.3}$$

where $ER(x)$ denotes the expected asset return, R_f denotes the risk-free asset return, β_n denotes the sensitivity of the asset price to fluctuations in systematic economic risk factors, and R_{pn} denotes a risk-premium emanating from systematic economic risk factors.

Muzindutsi and Niyimbanira [2012] used the arbitrage pricing model to determine the exposure of the returns of the top forty performing stocks listed on the JSE to a systematic

economic risk factor, the exchange rate.

2.2.3 The Multi-Factor Model

The multi-factor model uses the arbitrage pricing model as a baseline asset pricing model to identify the extent to which expected asset returns react to systematic economic risk factors, i.e., the inflation rate, interest rate, and economic cycle, among other systematic economic risk factors [Fama and French 1993].

The model equally considers market uncertainty, along with individual and joint variability among assets (or portfolios). Equation 2.4 defines the multi-factor model:

$$R_i = E\left(R_i\right) + \beta_{i1} + F_1 + \beta_{i2} + F_2 +, ..., +\beta_{ik} + F_k + \epsilon_i, \tag{2.4}$$

where R_i denotes the asset return, ER_i denotes the expected asset return, β_{i1} denotes the sensitivity of the asset or portfolio return to fluctuations in systematic economic risk factors, F_k denotes the systematic economic risk factor, and ϵ_i denotes market uncertainty.

Mukoyi and Ogujiuba [2022] compared the performance of various multi-factor models (i.e., the Fama and French three-factor model, Carhart four-factor model Fama and French five-factor model) when forecasting the reaction of stock market returns in the resource sector, industrial sector, and financial sector of South Africa to investment style risk.

The Estimation of Alpha and Systematic Risk Factors

Stock market research frequently considers the α *(alpha)* and β *(beta)* of the portfolio by arranging a *security characteristic line* in the linear model (Equation 2.5).

$$R_i - R_f = \alpha_i \beta_i (R_M - R_i) + \epsilon_i, \tag{2.5}$$

where R_i denotes the realized portfolio return, R_M denotes the market return, R_f denotes the risk-free return, α_i denotes the *alpha* of the portfolio, and β_i denotes the *beta* of the portfolio.

The Estimation of Cumulative Systematic Risk Factors

The linear function considers systematic economic risk factors, idiosyncratic risk, and the expected portfolio return, along with transaction cost. Equation 2.6 defines the linear function:

$$f(h) = \frac{1}{2}\kappa h_t^T Q^T Q h_t + \frac{1}{2}\kappa h_t^T S h_t - \alpha^T h_t + (h_{t-1})^T \Lambda (h_{t-1}), \tag{2.6}$$

where κ denotes a risk aversion factor.

The multi-factor model equally considers market neutrality, along with the position size and diversification of the portfolio. Equation 2.7 defines the gradient *(slope coefficient)* in the linear model:

$$f^{'}(h) = \frac{1}{2}\kappa(2Q^T Q h)\frac{1}{2}\kappa(2Sh) - \alpha + 2(h_{t-1})\Lambda. \tag{2.7}$$

Each common risk considers a systematic economic risk factor in the linear model. Equation 2.8 defines common risk:

$$C_{risk} = \frac{1}{2} h_t^T \beta F \beta \beta^T h_t. \tag{2.8}$$

The multi-factor model considers the portfolio return variance, diminishes the variance in proximity to zero, and vector estimates of α.

Not only that, but the model equally considers the response of the asset price to the market impact for each currency unit exchanged. Equation 2.9 transforms β to Q to diminish matrix expansion:

$$C_{risk} = \frac{1}{2} h_t^T Q^T Q h_t. \tag{2.9}$$

Equation 2.10 defines the linear impact model:

$$\sum_{i=1}^{N} = \lambda_{(i,t)} \left(h_{(i,t)} - h_{(i,t-1)} \right)^2, \tag{2.10}$$

where

$$\lambda_{(i,t)} = \frac{1}{10 \times ADV_{(i,t)}}, \tag{2.11}$$

where $ADV_{(i,t)}$ denotes the average daily volume traded for each 10 basis point.

2.2.4 The Modern Portfolio Model

The modern portfolio model considers portfolio return volatility a risk proxy, along with its expected return volatility and return weights. Not only that, but the model also confirms the statistical dependence among assets and identifies the efficient frontier of the portfolio [Elton and Gruber 2018]. Equation 2.12 estimates the expected portfolio return:

$$E(R_p) = \sum_{i=1}^{n} w_i E\left(R_i\right), \tag{2.12}$$

where R_p denotes the portfolio return, R_i denotes the expected asset return, and w_i denotes the weights of the asset return. Equation 2.13 defines the portfolio return variance:

$$\sigma_p^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} w_i w_j \sigma_i \sigma_j p_{ij}, \tag{2.13}$$

where σ_i denotes the extent to which the asset return deviates from the *mean* asset return, w_i denotes the weights of the asset return, and p_{ij} denotes the statistical dependence among assets. Equation 2.14 defines the portfolio return volatility:

$$\sigma_p = \sqrt{\sigma_p^2}. \tag{2.14}$$

Equation 2.15 defines the expected portfolio return containing two assets (i.e., asset A and B):

$$E(R_p) = w_A E(R_A) w_B E(R_B) = w_A E(R_A) + (1 - w_A) E(R_B). \tag{2.15}$$

Equation 2.16 defines the portfolio return variance:

$$\sigma_p^2 = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2 w_A w_B \sigma_A \sigma_B p_{AB}. \tag{2.16}$$

Taljaard and Maré [2021] used the modern portfolio model to identify changes in the concentration of market capitalization-weights in the top forty performing stocks listed on the JSE.

2.2.5 The Drawbacks of Asset Pricing Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities

This research acknowledges the importance of asset pricing models and their contribution to our understanding of forecasting the reaction of stock market returns to changing economic activities. However, these models have certain drawbacks (Table 2.2).

TABLE 2.2: The drawbacks of asset pricing models when forecasting the reaction of stock market returns to changing economic activities

Model	Drawback
The capital asset pricing model	The capital asset pricing model is criticized for its plainness and impracticality, along with its failure to consider the statistical significance between systematic economic risk factors and the expected asset return [Muthama *et al.* 2014; Andrei *et al.* 2018].
The arbitrage pricing model	Prevailing research on the reaction of stock market returns to changing economic activities hold that exploiting the expected asset return should not remain the sole focus of investors. Alternatively, they should equally consider the asset return volatility or the portfolio return volatility [Sinha 2016; Khudoykulov 2017].
The modern portfolio model	The concern of investors about the downside risk, which represents the financial risk of losses from investing in assets or a portfolio, is unrealized by the modern portfolio model [Otuteye and Siddiquee 2017; Crack and Grieves 2017]. The model equally does not consider a scenario, whereby the expected portfolio return exceeds the actual portfolio return [Hou *et al.* 2017].

2.2.6 The Rationale for Comparing the Performance of Statistical and Machine Learning Models

The previous section discussed the most conventional models for forecasting stock market returns, namely the capital asset pricing model (), the arbitrage pricing model (lsec:TheArbitragePricingModel), the multi-factor model (), and the modern portfolio model (). Furthermore, the the inadequacies of these models in stock market forecasting. Furthermore, the current portfolio model was undiscussed in depth since this study does not focus on several South African stocks, but rather on the stock market index, which includes all South African stocks.

In conclusion, the typical asset pricing methods outlined above are unable to manage the complexity of the data set employed in this study. This dissertation examined the effectiveness of statistical models and machine learning models in forecasting the impact of economic activity on stock market returns.

2.3 The use of Conventional Statistical Models in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities

Previous research on the reaction of stock markets in South Africa used conventional statistical forecasting models. For instance, Pillay [2020]; Chitenderu *et al.* [2014]; Makatjane and Moroke [2021] used statistical univariate forecasting model, and Aye *et al.* [2020]; Ilesanmi and Tewari [2020]; Adejayan and Oke [2022] used statistical multivariate models.

2.3.1 Statistical Univariate Forecasting Models

Statistical univariate forecasting models identify predictive patterns of a temporal feature and forecast successive patterns [Babu and Reddy 2014]. These forecasting models spearhead the frontier of research on stock market returns in South Africa [Pillay 2020; Chitenderu *et al.* 2014; Makatjane and Moroke 2021].

This research covers two conventional statistical forecasting models, i.e., the autoregressive integrated moving average model (section 2.3.1) and seasonal autoregressive integrated moving average model (section 2.3.1), to provide the background of statistical univariate forecasting.

The Autoregressive Integrated Moving Average Model

The autoregressive integrated moving average $(ARIMA)\ (p,d,q)$ *(or Box-Jenkins)* model bundles p—the lag k, d of the temporal feature—property changes of the temporal feature, and q—the moving average order, to identify predictive patterns and forecast a temporal feature [Young and Shellswell 1972]. Equation 2.17 defines the $ARIMA\ (p,d,q)$ model:

$$\hat{y}_t - y_{t-1} = \hat{\mu} + \hat{\phi}(y_{t-1} - y_{t-2}), ..., +\hat{\epsilon}_t, \tag{2.17}$$

where $\hat{y}_t$ denotes estimates of the temporal feature at period t, μ denotes unbiased estimates of the temporal feature, and $\hat{\epsilon}_t$ denotes residuals of the $(ARIMA)\ (p,d,q)$ model.

Mallikarjuna and Rao [2019] used the $ARIMA\ (p,d,q)$ model, along with other models like the self-exciting threshold autoregressive model, recurrent neural network, and a hybrid

model of the $ARIMA\ (p,d,q)$ model and recurrent neural network, to forecast stock market returns in South Africa.

The Seasonal Autoregressive Integrated Moving Average Model

The seasonal autoregressive integrated moving average $(SARIMA)\ (p,d,q) \times (P,D,Q,s)$ model amplifies the $ARIMA\ (p,d,q)$ model with additional parameters (i.e., p and seasonal P, d and seasonal D, q and seasonal Q) [Hyndman and Athanasopoulos 2013]. Equation 2.18 defines the $(SARIMA)\ (p,d,q) \times (P,D,Q,s)$ model:

$$\hat{y}_t^s = y_t - y_{t-4} + \hat{\epsilon}_i. \tag{2.18}$$

Equation 2.19 translates Equation 2.18:

$$y_t^S = (1 - B^4) \times y_t. \tag{2.19}$$

Makatjane and Moroke [2021] used the $(SARIMA)\ (p,d,q) \times (P,D,Q,s)$ model to forecast stock market returns in South Africa.

2.3.2 The Statistical Multivariate Forecasting Model

Statistical multivariate forecasting models identify predictive patterns of temporal features and forecast successive patterns while considering model residuals. Research on the reaction of stock market returns to changing economic activities in South Africa repeatedly used a popular conventional statistical multivariate forecasting model, the vector autoregressive model [Aye *et al.* 2020; Ilesanmi and Tewari 2020; Adejayan and Oke 2022].

This research considers the vector autoregressive model, the benchmark model, because of its widespread adoption in research.

The Vector Autoregressive Model

The vector autoregressive model maintains a stochastic system that interprets multiple features as linear p lag combinations, along with p lags [Gouriéroux *et al.* 2017]. For clarification, assume a temporal problem with two features, Equation 2.20 and Equation 2.21 defines the vector autoregressive model:

$$\hat{y}_t = \alpha + \hat{\beta}_{11} y_{t-1} + \hat{\beta}_{12} y_{t-2} + \hat{\gamma}_{11} x_{12} + \hat{\gamma}_{12} x_{t-2} + \hat{\epsilon}_{1t}, \tag{2.20}$$

$$x_t = \alpha + \hat{\beta}_{21} y_{t-1} + \hat{\beta}_{22} y_{t-2} + \hat{\gamma}_{21} x_{22} + \hat{\gamma}_{22} x_{t-2} + \hat{\epsilon}_{2t}, \tag{2.21}$$

where $\hat{y}_t$ denotes $k \times 1$ temporal feature estimates, $\hat{\alpha}$ denotes $n \times 1$ intercept vector estimates, $\hat{\gamma}_i$ denotes $k \times k$ coefficient matrix estimates, and $\hat{\epsilon}_t$ denotes a serial uncorrelated random vectors order with $\bar{x}_t = 0$ and sum joint variability among temporal features.

Pillay [2020] and Macfarlane [2011] used the vector autoregressive model to forecast the reaction of stock market returns to changing economic activities in South Africa.

The structural vector autoregressive model bundles Equation 2.20 and Equation 2.21, then identifies causality among temporal features and concludes on the statistical significance among them [Gouriéroux *et al.* 2017; Tank *et al.* 2021]. Equation 2.22 defines the structural vector autoregressive model:

$$\hat{y}_t = \hat{\alpha}_1 + \hat{\beta}_1 x_t + \hat{\phi}_{11} y_{t-1} + \hat{\phi}_{12} y_{t-2} + \hat{\phi}_{11} x_{12} + \hat{\phi}_{12} x_{t-2} + \hat{v}_{1t}, \tag{2.22}$$

$$x_t = \alpha_2 + \hat{\beta}_1 x_t + \hat{\phi}_{21} y_{t-1} + \hat{\phi}_{22} y_{t-2} + \hat{\phi}_{21} x_{22} + \hat{\phi}_{22} y_{t-2} + \hat{v}_{2t}. \tag{2.23}$$

2.3.3 The Drawbacks of Conventional Statistical Forecasting Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities

Conventional statistical forecasting models are universally regarded in previous research on the reaction of stock market returns to changing economic activities in South Africa, but they have drawbacks (Table 2.3).

TABLE 2.3: The drawbacks of conventional statistical forecasting models when forecasting the reaction of stock market returns to changing economic activities

Drawback	Description
Non-stationarity	Temporal features rarely follow a stationary process, and the model residual expansion problem is common in analysis, resulting in non-adherence to some requirements of conventional statistical forecasting models.
Non-linearity	Because temporal features are frequently non-linear, conventional statistical forecasting models are unsuitable for modeling complex data structures, i.e., when the data set contains multiple temporal features with high dimensions [Liu *et al.* 2021; Pahlawan *et al.* 2021a].
The curse of data dimensionality problem	As specified, conventional statistical forecasting models are incapable of handling a data set containing multiple temporal features with high dimensions [Ahangar *et al.* 2010].
Subpar model performance	Due to non-adherence of temporal features to conventional statistical forecasting model requirements, the model performance tends to be subpar [Zimmerman 1994; Anscombe and Guttman 1960; Pincus 1995].

The machine learning approach addresses the drawbacks of conventional statistical forecasting models [BenSaïda and Litimi 2013; Liu *et al.* 2020; Kennedy *et al.* 2020].

2.4 State of Literature

Policymakers became concerned about the instability in the stock market during and after the global market crisis started in 2008, particularly those in emerging economies [Hedging and Umoetok 2013]. Such occurrences amplified the need for sophisticated analytical tools like machine learning models (section 2.6), because these models can capture complex market data structures [Ramos-Pérez *et al.* 2021; Liu *et al.* 2021; Pahlawan *et al.* 2021a].

2.4.1 Related Research

This research reviews previous research, then identifies related research debates, along with research inconsistencies and gaps it intends to fill, before selecting candidate machine learning models and the research method to use.

Table 2.4 shows the model criteria guiding this research, along with feature sets, model specifications, and the performance of models that forecast the reaction of stock market returns to changing economic activities.

TABLE 2.4: Related research

Author	Features	Model Specifications	Model Performance
Ndikum [2020]	200 economic features across multiple categories (i.e., the economic category, and money and banking category, among other economic categories).	Compared to the performance of the restricted Boltzmann machine and the deep belief network when forecasting the reaction of S&P 500 returns to 200 economic features, as measured by the *mean squared error (MSE)*.	The restricted Boltzmann machine achieved a *mean squared error (MS)E* value of 0.36. Whereas, the deep belief network achieved a *MSE* value of 0.35.

Continued on next page

Table 2.4 – Continued from previous page

Author	Features	Model Specifications	Model Performance
Ahangar *et al.* [2010]	Data comprised 40 microeconomic features and economic features across multiple categories (i.e., the rates category, and the money and banking category, among other economic categories).	7 economic features were selected based on the *explained variance ratio* value found by the principal component analysis method. The deep belief network (with three *hidden* layers and 14 neurons) was used to forecast the reaction of stock market returns to changing economic activities in Iran.	The *MSE* value for the deep belief network was 31.6.
Pahlawan *et al.* [2021a]	Data included 20 economic features spread over multiple economic categories (i.e., the rates category, indices category, and money and banking category, among other economic categories).	Compared the performance of the recurrent neural network, gated recurrent unit, and long-short term memory when forecasting the reaction of stock market returns (particularly the S&P 500 returns) to changing economic activities in the U.S., as measured by the *MSE*.	The *MSE* value for the long-short term memory was 1.35, the *MSE* value for the gated recurrent unit was 1.55, and the *MSE* value for the recurrent neural network was 1.55.
Wong *et al.* [2020]	Data comprises 74 industry-specific features and 102 economic features.	The deep belief network (with an online early stopping strategy) was used to the reaction of stock market returns (particularly the S&P 500 returns) to changing industry-specific activities and economic activities in the U.S.	The *MSE* value for the deep belief network was 50.22.

Continued on next page

Table 2.4 – Continued from previous page

Author	Features	Model Specifications	Model Performance
Kamalov [2020]	Data comprises the S&P 500 price volatility simulation data.	Compared the performance of the long-short term memory, multi-layer perceptron, and convolutional neural network when classifying the reaction of stock market return volatility (particularly the S&P 500 return volatility) to changing economic activities in the U.S.	The long-short term memory outperformed the multi-layer perceptron and convolutional neural network, with a 0.85 *area under the curve* value.
Xiong *et al.* [2015]	Data contained multiple broad economic features.	The *Shapley* additive method was used to select features. The long-short term memory was used to forecast the reaction of stock market returns to a feature subset and whole economic features.	The first long-short term memory achieved a *MSE* value of 2 890, and the second long-short term memory achieved a *MSE* value of 2 880.
Klibanov *et al.* [2021]	Data comprised Russell index returns simulation data.	The deep belief network was used to classify Russell index returns.	The *accuracy* value for the deep belief network was 55.42.
Ramos-Pérez *et al.* [2021]	Data comprised a few economic features.	Compared the performance of the long-short term memory and multi-layer perceptron when forecasting the reaction of stock market return volatility (particularly S&P 500 return volatility) to changing economic activities in the U.S., as measured by the *root mean squared error (RMSE)*.	The *RMSE* value for the multi-layer perceptron, which was astonishingly near zero, was the highest.

Continued on next page

Table 2.4 – Continued from previous page

Author	Features	Model Specifications	Model Performance
Fu *et al.* [2018]	Data comprised 244 technical and fundamental features.	Compared the performance of deep belief networks (with different structures) when forecasting the reaction of stock market returns (particularly S&P 500 returns) to changing technical features and economic activities in the U.S., as measured by the *mean absolute error (MAE)*.	The *MAE* values for deep belief networks were 2.98 and 0.97, respectively.
Alhomadi [2021]	Data comprised 30 macroeconomic features and U.S. stock market indices features.	Compared the performance of the ordinary least-squares regression model, elastic net model, support vector regression model, random forest model, and extreme gradient boosting model, when forecasting the reaction of stock market returns (particularly the S&P 500 returns) to changing economic activities in the U.S., as measured by the *R-Squared*.	The *R-Squared* value of the ordinary least-squares regression model was 0.1967, the *R-Squared* value of the elastic net model was 0.4559, the *R-Squared* value of the support vector regression model was 0.4970, the *R-Squared* value of the random forest model was 0.3363, and the *R-Squared* value of the extreme gradient boosting model was 0.4215.
Nengovhela [2022]	Data comprised economic features and stock market indices features.	Compared to the performance of the random forest model, k nearest neighbor model, support vector regression model, decision tree model, and neural network, when forecasting the reaction of stock market returns to changing economic activities in South Africa.	The *MAE* value of the random forest model was 0.9609, the *MAE* value of the k nearest neighbor model was 0.9817, the *MAE* value of the support vector regression model was 1.1231, the *MAE* value of the decision tree model was 1.3247, and the *MAE* value of the neural network was 0.9819.

Continued on next page

Table 2.4 – Continued from previous page

Author	Features	Model Specifications	Model Performance
Mallikarjuna and Rao [2019]	Data comprised stock market indices features of developed, emerging, and frontier economies.	Compared the performance of different models (i.e., the $ARIMA\ (p,d,q)$ model, self-exciting threshold autoregressive model, recurrent neural network, singular spectrum analysis model, and a hybrid model of the $ARIMA\ (p,d,q)$ model and recurrent neural network) when forecasting stock market returns in South Africa, as measured by the *RMSE*.	The *RMSE* of the$ARIMA\ (3,0,1)$ model was 1.062336, the *RMSE* of the self-exciting threshold autoregressive model was 1.064449, the *RMSE* of the recurrent neural network was 1.063028, the *RMSE* of the singular spectrum analysis model was 1.066819, and the *RMSE* of the hybrid model was 1.061791.

Table 2.4 shows most research focused on whole features, few models for comparison, and barely focused on how different feature subsets and reduced dimensionality change model performance, a limitation this research addresses when considering the number of experiments.

This research established that machine learning models can be used to forecast the reaction of stock market returns to changing economic activities in the emerging economic context. It advances research by comparing various models with different feature subsets and data dimensions.

2.5 Research Gaps

While more preliminary research has examined how economic variables impact stock market returns in South Africa, only a few studies have looked into the performance differences between statistical models and machine learning models. Those that did so concentrated on sophisticated economies, employed restricted experimental scenarios, and made out-of-sample forecasts over a single time horizon.

This study, on the other hand, focuses on comparing the effectiveness of statistical models and machine learning models in projecting the reaction of stock market returns to changing economic activity in the developing economic scenario. This study is unusual in that it tested several models trained on distinct feature subsets and economic data components over varied time periods.

2.6 The use of Machine Learning Models in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities

The machine learning approach denotes a practical systematic approach that maps out logical steps for sophisticated computer systems to learn tasks T based on experience [Liu 1996; Mitchell 1997; Kennedy *et al.* 2020]. This approach advances on task T using knowledge of prior performance P estimates, identifies complex data structures, and enhances the generalization capacity [Dietterich 1996], which is useful for forecasting the reaction of stock market returns to changing economic activities.

This research used the ordinary least-squares regression model, decision tree model, random forest tree model, extreme gradient boosting tree model, recurrent neural network, gated recurrent unit network, long-short term memory, restricted Boltzmann machine, and multilayer perceptron.

2.6.1 The Ordinary Least-Squares Regression Model

The ordinary least-squares regression model learns x_i *(predictor features)* and predicts y_i *(a target feature)* while deflating $\hat{\epsilon}_i$ *(ordinary least-squares regression model residuals)*. Equation 2.24 defines the ordinary least-squares regression model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\epsilon}_i, \tag{2.24}$$

where $\hat{y}_i$ denotes predicted y_i estimates, $\hat{\beta}_0$ *(the y-intercept)* denotes $\bar{y}_i$, where $x_i = 0$, and $\hat{\beta}_i$ *(a slope coefficient)* denotes the path of corresponding changes between x_i and $\hat{y}_i$. Equation 2.25 defines $\hat{\beta}_0$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}_i)}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}, \tag{2.25}$$

where $\bar{x}_i$ denotes the *mean* value of x_i and $\bar{y}_i$ denotes the *mean* value of y_i.

Equation 2.26 defines $\hat{\epsilon}_i$:

$$\hat{\epsilon}_i = y_i - \hat{y}_i. \tag{2.26}$$

Equation 2.24 with more x_i results in Equation 2.27:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3, ..., + \hat{\epsilon}_i. \tag{2.27}$$

[Marozva 2020] used the ordinary least-squares regression model to forecast the reaction of stock market return volatility to changing economic activities and political activities in South Africa. Equally, Mpofu [2011] used the model to forecast the reaction of stock market return to the manufacturing index and prime overdraft rate, among other features.

2.6.2 The Decision Tree Model

The decision tree model adopts a recursive partitioning strategy to isolate feature values. Successively, the model reduces the *impurity* using the *gini impurity* estimator and splits nodes while inflating the *entropy* (Equation 2.29) [Moore II 1987]. Equation 2.28 defines the *gini impurity* estimator:

$$\hat{f}(x_i) = 1 - \sum_{i=1}^{c} \hat{p}_j^2, \tag{2.28}$$

where p_j denotes class c of the node in the sample proportion. By dividing feature values into manageable chunks, the *entropy* estimator isolates homogeneous feature values of nodes and then identifies irregularities. Equation 2.29 defines the *entropy* estimator:

$$\hat{f}(x_i) = -\sum_{i=1}^{c} \left(log_2 \hat{p}_j\right). \tag{2.29}$$

For Equation 2.29, $p_j \neq 0$, provided class c is unfilled. $Entropy = 0$ for feature values is similar to the class c of the node in the sample proportion.

Nengovhela [2022] used decision tree models to forecast the reaction of stock market returns to changing economic activities in South Africa.

2.6.3 The Random Forest Tree Model

The random forest tree model unifies decision tree models produced through a random process to enhance model performance by using the *loss* minimization approach at multiple iterations [Vijayakumar and Cheung 2018]. Equation 2.30 defines the random forest tree model:

$$\hat{y}_i = \frac{1}{N} \sum_{i=1}^{n} \hat{f}\left(x_i^{'}\right), \tag{2.30}$$

where $\hat{f}\left(x_i^{'}\right)$ denotes the function (a linear function for this research).

Nengovhela [2022] used random forest tree models to forecast the reaction of stock market returns to changing economic activities in South Africa.

2.6.4 The Extreme Gradient Boosting Tree Model

The extreme gradient boosting tree model bundles hollow decision tree models to enhance model performance by using the *loss* minimization approach at multiple iterations [Hastie *et al.* 2009; Nokeri 2021]. This involves evaluating the performance of decision tree models and referring to the previous subpar model performance, while considering decision tree model residuals from previous iterations [Mason *et al.* 1999]. Equation 2.31 learns x_i:

$$\hat{y}_i = \hat{m}_i\left(x_i\right) + \hat{\epsilon}_1. \tag{2.31}$$

Equation 2.32 investigates the $\hat{\epsilon}_i$ dependency:

$$\hat{\epsilon}_3 = \hat{h}_i\left(x_i\right) + \hat{\epsilon}_3. \tag{2.32}$$

Equation 2.33 bundles the regressed $\hat{\epsilon}_i$ in Equation 2.32:

$$\hat{y}_i = \hat{m}_i\left(x_i\right) + \hat{g}_i\left(x_i\right) + \hat{h}_i\left(x_i\right) + \hat{\epsilon}_3. \tag{2.33}$$

Equation 2.34 completes $\hat{y}_i$:

$$\hat{y}_i = \alpha \times \hat{m}_i\left(x_i\right) + \hat{\beta}_i \times \hat{g}_i\left(x_i\right) + \hat{\gamma} \times \hat{h}_i\left(x_i\right) + \hat{\epsilon}_4. \tag{2.34}$$

Alhomadi [2021] used the extreme gradient boosting tree model to forecast the reaction of stock market returns to changing economic activities in the U.S. No research used the model to forecast the reaction of stock market returns in South Africa.

2.7 The use of Neural Networks in Forecasting the Reaction of Stock Market Returns to Changing Economic Activities

Neural networks are distinct machine learning model classes that are a replica of animals' biological neural networks. A neural network accumulates x_i *(predictor features)* in the *input* layer (the first layer in a neural network), then use $\hat{f}\left(x_i\right)$ *(an activation function)* to identify complex data structures and route feature values to successive *hidden* layers (layers between the *input* layer and the *output* layer in a neural network), designate dissimilar w_i *(weights)* and β_i *(biases)*, then use $\hat{f}\left(x_i\right)$ in the *output* layer (the last layer in a neural network) to forecast subsequent y_i *(predicted target feature estimates)*.

This research considers a subset of machine learning models, acknowledged as neural networks, i.e., the recurrent neural network (subsection 2.7.1), gated recurrent unit (subsection 2.7.2), long-short term memory (subsection 2.7.3), restricted Boltzmann machine (subsection 2.7.4), and multi-layer perceptron (subsection 2.7.5).

2.7.1 The Recurrent Neural Network

The recurrent neural network accumulates x_t, then uses $\hat{f}(x_t)$ to identify x_t and $\hat{h}_{t-1}$ (previous *hidden* states), designate distinctive w_t, and forecast y_t *(target feature values)* and successive $\hat{h}_t$ using the *tangenthyperbolic* activation function. Equation 2.35 defines the recurrent neural network:

$$\hat{h}_t = tanh(w_h \times h_{t-1} \times w_x \times x_t), \tag{2.35}$$

where $\hat{h}_t$ denotes *hidden* states, w_h denotes weights of h_{t-1}, w_x denotes weights of x_t, and *tanh* denotes a *tangent hyperbolic* activation function that predicts y_t restricted to [-1, 1]. Equation 2.36 predicts y_t:

$$\hat{y}_t = w_{hy} - \hat{h}_t. \tag{2.36}$$

Sako *et al.* [2022] used the recurrent neural network to forecast the reaction of stock market returns to the exchange rate (ZAR/USD) in South Africa.

Mallikarjuna and Rao [2019] compared the performance of different models (i.e., the $ARIMA$ (p,d,q) model, self-exciting threshold autoregressive model, recurrent neural network, singular spectrum analysis model, and a hybrid model of the $ARIMA$ (p,d,q) model and recurrent neural network) when forecasting stock market returns in South Africa, among other countries, as measured by the *root mean squared error*. The research found that all candidate models outperform the recurrent neural network.

2.7.2 The Gated Recurrent Unit Network

The gated recurrent unit uses $\hat{f}(x_t)$ to learn x_t and h_{t-1}, and contains a *forget* gate that $forgets$ invaluable x_t and h_{t-1}. The neural network equally contains a *reset* gate (or an *update* gate) [Chung *et al.* 2014], whereby t and $h = 0$:

$$z_t = \sigma_i \left(w_z[h_{t-1}, x_t]\right), \tag{2.37}$$

$$r_t = \sigma_i \left(w_r[h_{t-1}, x_t]\right), \tag{2.38}$$

$$\hat{h}_t = tanh \left(w_r[r_t h_{t-1}, x_t]\right), \tag{2.39}$$

$$\hat{h}_t = (1 - z_t) \times \left(w_o[h_{t-1} + z_t + \hat{h}_t]\right), \tag{2.40}$$

where z_t denotes an *update* gate, r_t denotes a *reset* gate, $\hat{h}_t$ denotes *hidden* states, x_t denotes predictor features, *tanh* denotes a *tangent hyperbolic* activation function, and w_t denotes the *weight* matrix.

Sako *et al.* [2022] compared the performance of various neural networks (i.e., the gated recurrent unit, among other neural networks like the recurrent neural network and long-short term memory) when forecasting the reaction of stock market returns to the exchange rate, the

South African rand to the U.S. dollar.

2.7.3 The Long-Short Term Memory

The long-short term memory accumulates x_t and uses $\hat{f}(x_t)$ in an *update* gate to conclude by evoking some x_t and h_{t-1}, before routing the rest to cell state vectors of h_{t-1} that use $\hat{f}(x_t)$ to forecast y_t [Hochreiter and Schmidhuber 1997].

$$\hat{f}(x)_t = \sigma(W_f[h_{t-1}, x_t] + \hat{\beta}_f), \tag{2.41}$$

$$i_t = \sigma(w_i[h_{t-1}, x_t] + \hat{\beta}_i), \tag{2.42}$$

$$\tilde{x}_t = tanh(w_i[h_{t-1}, x_t] + \hat{\beta}_C), \tag{2.43}$$

$$\tilde{c}_t = f_t \times C_{t-1} + i_t + c_t, \tag{2.44}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + \hat{\beta}_0), \tag{2.45}$$

$$\hat{h}_t = o_t \times tanh(c_t), \tag{2.46}$$

where x_t denotes predictor features, f_t denotes the activation vector of the *forget* gate, $\hat{h}_t$ denotes predicted *hidden* states, i_t denotes the activation vector of *input* gate or *update* gate, $\tilde{x}_t$ denote cell state vectors of $\tilde{c}_t$, and w_t denotes a *weight* matrix, and $\hat{\beta}_C$ denotes the *bias* of x_t in cell state vectors.

Balusik *et al.* [2021] compared the performance of the long-short term memory to the performance of the $(SARIMA)\ (p,d,q) \times (P,D,Q,s)$ model when forecasting stock market returns in South Africa, as measured by the *RMSE* metric. The research found the long-short term memory outperforms the $(SARIMA)\ (p,d,q) \times (P,D,Q,s)$ model.

2.7.4 The Restricted Boltzmann Machine

The restricted Boltzmann machine denotes an abstract neural network with $\hat{v}_i$—a *visible* layer and $\hat{h}_j$—a *hidden* layer attached to $w_{i,j}$. Equation 2.47 defines the *energy* function of $\hat{v}_i$ and $\hat{h}_j$:

$$E(v_i, h_i) = \sum_{i=1}^{n} \alpha_i v_i - \sum_{i=1}^{n} \hat{\beta}_j h_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{v}_i w_{i,j} \hat{h}_j, \tag{2.47}$$

where α_i denotes weights and biases of $\hat{v}_i$, and b_j denotes biases of $\hat{v}_j$. Equation 2.48 translates Equation 2.47.

$$E\left(\hat{v}_i, \hat{h}_i\right) = -\alpha^T v_i - \hat{\beta}_j{}^T \hat{h}_j - \hat{v}_i{}^T w_{i,j} \hat{h}_j. \tag{2.48}$$

da Costa and Gebbie [2020] used the restricted Boltzmann machine stacked with auto-encoders to forecast the stock market price in South Africa.

2.7.5 The Multi-Layer Perceptron

The multi-layer perceptron maintains an *input* layer and *output* layer, along with two layers between them (*hidden* layers) at most. This feed-forward network operates the back-propagation learning approach to estimate the gradient of the *loss* function, along with its weights. Equation 2.49 accumulates and converts x_i and designates dissimilar w_{ji} and $\hat{\beta}_j$, then routes them to an initial *hidden* layer:

$$\hat{f}(x_i) = \hat{\beta}_j + \sum_{i=1}^{n} w_{ij} x_i. \tag{2.49}$$

Equation 2.50 uses φ_m in an initial *hidden* layer, which incrementally routes feature values to successive *hidden* layers:

$$\varphi_m = [1 + exp(f(x_i))]^{-1}. \tag{2.50}$$

Equation 2.51 uses an activation function to forecast y_i, which are treated as x_i by successive *hidden* layers:

$$\hat{y}_i = \hat{\beta}_j + \sum_{j=1}^{n} w_j \varphi_j. \tag{2.51}$$

Ataman and Kahraman [2021] compared the performance of the ordinary least-squares model, multi-layer perceptron, and a hybrid model (integrating the ordinary least-squares model and multi-layer perceptron) when forecasting the reaction of stock market returns to changing economic activities in Brazil, Russia, India, China, and South Africa, as measured by the *R-Squared* metric. The research found the hybrid model outperforms the ordinary least-squares model and multi-layer perceptron.

2.8 The Activation Function

To forecast y_i (*target feature values*), $\hat{f}(x_i)$ (an activation function) first accumulates x_i and identifies complex data structures, then predicts subsequent y_i. There are multiple activation functions (i.e., the *sigmoid* activation function, *tangent hyperbolic* activation function, and *rectified linear unit* activation function, among other activation functions).

2.8.1 The Rectified Linear Unit Activation Function

This research uses a rectified linear unit *(relu)* activation function at layers of neural networks, because it does not bound $\hat{y}_i$, since it produces $\hat{y}_i$ that range from 0 to ∞ *(infinity)* after accumulating x_i and identifying their predictive patterns. Equation 2.52 defines the *relu* activation function:

$$Relu = max(0, x_i). \tag{2.52}$$

The *relu* activation function acquires feature values (and/or *hidden* states) from the *input* layer and absorbs them, then routes them to concurrent *hidden* layers, which learns predictive patterns and attaches varying *biases* and *weights*.

In the last *hidden* layer, the *relu* activation function directs predictor feature values (and/or *hidden* state) to the *output* layer, which predicts concealed target feature values (and/or concurrent*hidden* states).

3 Research Methods

This research adopts a research method from previous research. Figure 3.1 shows the modeling pipeline exhibiting the workflow that produces models that forecast the reaction of stock market returns to changing economic activities.

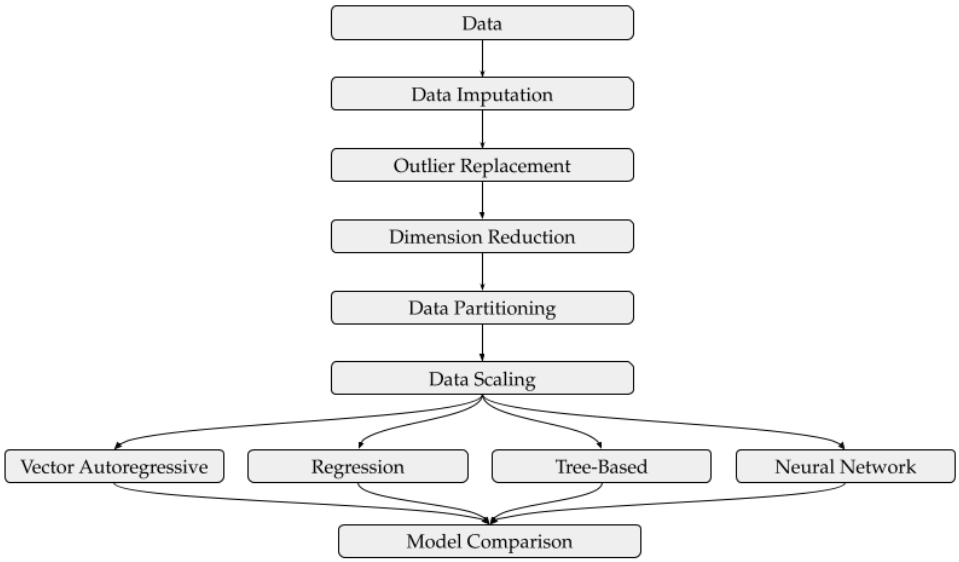


FIGURE 3.1: The modeling pipeline exhibiting the workflow that produces models that forecast the reaction of stock market returns to changing economic activities

3.1 The Acquisition of Economic and Stock Market Data of South Africa

This research extracts 54 monthly economic features from the South African Reserve Bank (SARB) database, and 99 monthly economic features from the Federal Reserve Economic Data (FRED) database, because those were available monthly features in the specified databases.

It extracts the monthly FTSE/JSE all-share index price feature from the JSE database. The sample period was selected because the Financial Times Stock Exchange/Johannesburg Stock Exchange (FTSE/JSE) all-share index was established in 2002 by the FTSE and JSE. For some unexplained reason, the majority of the factors addressed in this research began to be covered around 2002. The sample period will now begin in 2002.

3.1.1 Sources of Economic and Stock Market Data of South Africa

Table 3.1 shows sources of economic and stock market data of South Africa, along with the description of data and the sample period.

TABLE 3.1: Sources of economic and stock market data of South Africa, along with the description of data and the sample period

Source	**Description**	**Features**	**Sample Period**
FRED	A regional reserve bank that is part of the United States Central Bank with headquarters in Washington, D.C. is known as the Federal Reserve Bank of St. Louis [1].	99 monthly changing economic features *(predictor features)* sourced from the FRED database.	2002 - 2022
SARB	SARB is the Central Bank of South Africa [2].	54 monthly changing economic features *(predictor features)* sourced from the SARB database	2002 - 2022
JSE	JSE is the official stock exchange in South Africa.	A single monthly temporal feature, the FTSE/JSE all-share index price *(the target feature)* sourced from the SARB [3].	2002 - 2022

The Fred database combines data from many sources, including Statistics South Africa (Stats SA), the Organization for Economic Co-operation and Development (OCED), and the South African Reserve Bank (SARB). Some variables are unavailable in the SARB database. As a result, the FRED database was utilized as a supplementary data source to collect South African economic information.

3.1.2 Properties of Economic and Stock Market Data of South Africa

Table 3.2 shows properties of economic and stock market data of South Africa, and commentary on how those properties affect models that forecast the reaction of stock market returns

[1]The Federal Reserve Bank of St. Louis periodically maintains economic data of many nations, along with changing economic activities, in a database, acknowledged as FRED.

[2]The principal SARB mandate is to structure and influence South African economic expansionary policies, and produce banknotes and coins, among other mandates.

[3]The FTSE/JSE all-share index represents the benchmark performance of equities in South Africa for 99.9% of the total market capitalization.

to changing economic activities, along with strategies for addressing non-adherence to model requirements. The purpose of identifying the properties was to identify which research techniques to use.

TABLE 3.2: The properties of economic and stock market data of South Africa, and commentary on how those properties affect models that forecast the reaction of stock market returns to changing economic activities, along with strategies for addressing non-adherence to model requirements

Data Properties	**Vector Autoregressive**	**Linear Regression**	**Tree-based**	**Neural Network**	**Comments**
Non-stationarity		***	***		Certain economic features are not stationary. To adhere to the vector autoregressive model stationary requirement. This research used an order differentiation strategy to produce economic features that proceed linearly and prevent the model residual expansion problem.
Few economic feature values	***	***	*	***	Economic features contain a few feature values, which makes it challenging to identify complex predictive patterns using neural networks, as neural networks are intended for large data sets.
Missing economic feature values			*		While neural networks and the decision tree model do not hold rigid model requirements regarding missing values, the vector autoregressive model and ordinary least-squares regression model hold rigid requirements. The k nearest neighbor imputation strategy imputes missing economic feature values.
High dimensions of economic activities	*	***	**	***	The vector autoregressive and ordinary-least squares regression model cannot identify complex structures in economic and stock market data and generalize unknown observations when there are high dimensions of economic activities. The principal components analysis method reduces the dimensions of economic activities of South Africa into meaningful eigenvectors.
Different scales and quantities of economic features	***	***	***	***	Economic features come in different sizes and quantities. The data standardization scaling strategy scales changing economic activities before modeling.

Legend

*	Poor data modeling capacity.
**	Regular data modeling capacity.
***	Optimal data modeling capacity.

3.1.3 Strategies used to Cleanse and Preprocess Data Before Forecasting

Table 3.3 shows strategies used to cleanse and preprocess data strategies before forecasting the reaction of stock market returns to changing economic activities in South Africa.

TABLE 3.3: Strategies used to cleanse and preprocess data strategies before forecasting the reaction of stock market returns to changing economic activities in South Africa

Strategy	Description
Redundant economic features removal	There were duplicate economic features present in original data sets extracted from the Federal Reserve Economic Data (FRED) database and South African Reserve Bank (SARB) database. Duplicate features were removed before combining data sets for redundancy avoidance in analysis.
Data labeling	Economic features extracted from the FRED database had economic features codes as column names. Column names were renamed to their official economic feature names.
Data concatenation	Economic features came from different data sources (i.e., the FRED database, SARB database, and JSE database). Economic features were concatenated to complete the data set.
Indexing	The index *(date)* of Economic features extracted from the FRED database. Stock market returns had a format that was incompatible with platform and model requirements. The format was set to a *Year-Month-Day* format.
Sample period specification	Economic features had varying sample periods. Sample periods were set over 22 years (from 2002 to 2022), because the FTSE/JSE all-share index price was established in 2002 by the FTSE and JSE.
Separator conversion	Economic features extracted from the SARB database contained comma separators. The conversion from a comma separator to a decimal point separator, so features adhere to platform and model requirements.
Infinity conversion	Certain economic features contained negative and positive infinity values of a float type. Infinity values of a float type were converted to *NaN* (missing feature values), then replaced using the k nearest neighbor data imputation strategy.
Feature stationarity by order differencing	Features were made stationary using the order differentiation strategy.

Continued on next page

Table 3.3 – Continued from previous page

Strategy	Description
Data imputation	Economic features contained missing feature values. The k nearest neighbor data imputation strategy replaced missing feature values with values in proximity to them.
Outlier replacement	Economic features had varying sample periods. In this manner, features contained outliers. The *median* outlier replacement strategy replaced outliers with the *median* value.
Data partitioning	The random data partitioning strategy ensured models learn economic activities and forecast the reaction of stock market returns in South Africa. This strategy was adopted with varying split ratios to compare the performance of models over four periods.
Data dimension reduction	The principal components analysis method reduced dimensions of economic activities in South Africa into meaningful eigenvectors. The principal component analysis method produced an index that offers insight into the structure of the economy in South Africa based on the explained variance ratio.
Data standardization	Economic features came in different scales and quantities. The data standardization scaling strategy contains economic activities on a standard scale.

3.2 The Strategy used to Impute Missing Economic Data of South Africa

Most machine learning models cannot handle missing values [Rubin 1975], hence Leke and Marwala [2019]; Thulare *et al.* [2021] suggested the use of data imputation strategies. This research used the k nearest neighbor imputation strategy to replace missing values (subsection 3.2.1).

3.2.1 The *k* Nearest Neighbor Imputation Strategy

Previous research suggested the use of the k nearest neighbor imputation strategy to impute missing values of economic features with values in proximity using the Euclidean distance (d) method [Zhang 2012; Mulaudzi and Ajoodha 2020]. Equation 3.1 defines the Euclidean d method:

$$d\left(x_i, y_i\right) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{3.1}$$

where d denotes the distance between feature values, n denotes the number of x_i, whereas x_i and y_i denote feature values in a *Euclidean* space.

3.3 The Strategy used to Replace Outliers in Economic and Stock Market Returns Data of South Africa

Outliers pose a problem in analysis, because some models are sensitive to outliers. The research uses the *median* outlier replacement strategy to replace outliers with the *median* value (subsection 3.3.1).

3.3.1 The Median Outlier Replacement Strategy

The *median* outlier replacement strategy uses the *median* value to replace outliers (i.e., $x_i <$ 5% percentile or $x_i >$ 95% percentile) in the sample probability distribution. Equation 3.2 defines the *median (M)*, where n is odd:

$$M = \frac{n+1}{2}, \tag{3.2}$$

where n denotes the sample size. Equation 3.3 defines M, where n is even:

$$M = \frac{\left(\frac{n}{2}\right) + \left(\frac{n}{2}+1\right)}{2}. \tag{3.3}$$

3.4 The Strategy used to Reduce Economic Data Dimensions of South Africa

The *curse of data dimensionality* problem frequently leads to an upsurge in the amount of power required for computation [Young 2020]. This research used the principal components analysis method to reduce dimensions of economic activities in South Africa (subsection 3.4.1).

3.4.1 The Principal Components Analysis Method

The principal component analysis method, an unsupervised machine learning method, reduces feature sets into meaningful eigenvectors. This method is suitable for addressing the *curse of data dimensionality* problem [Shen 2009]. Equation 3.4 defines the principal component analysis method:

$$\sum_{i=1}^{n} \frac{1}{m} \left(x_t - \bar{s}\right) \left(x_i - \bar{x}\right)^T, \tag{3.4}$$

where $\bar{x}$ emanates from eigenvectors.

Interestingly, the research ranks economic activities based on the explained variance ratio (σ_i^2) value to produce an index that offers insight into the structure of the economy in South Africa. Equation 3.5 defines σ_i^2:

$$\sigma_i^2 = \sum_{i=1}^{n} \frac{\hat{v}\left(\tilde{x}_i\right)}{\hat{v}\left(x_i\right)}, \tag{3.5}$$

where $\hat{v}$ denotes the explained variance ratio. A scree plot identifies the number of components needed to complete the principal component analysis method. The starting point of eigenvalue misrepresentation defines the n component selection criterion.

3.5 The Strategy used to Partition Economic and Stock Market Returns Data of South Africa

This research used the random data partitioning strategy to partition data, so each feature value has an odd chance of being included in the selection (subsection 3.5.1). It considers four split ratios to identify the performance of machine learning models over four periods.

3.5.1 The Random Data Partitioning Strategy

The random data partitioning strategy designates feature values to either the training or test data set through a probabilistic process. Estimating the population proportion establishes the confidence interval. Equation 3.6 defines the population proportion:

$$\hat{p}_i \pm z \sum_{i=1}^{n} \frac{\hat{p}_i \left(1 - \hat{p}_i\right)}{n}, \tag{3.6}$$

where $\hat{p}_i$ denotes sample proportion estimates, n denotes the sample size, and z denotes the critical value. Equation 3.7 defines the mean value of the sample proportion:

$$\hat{\mu}_i = \hat{p}_i. \tag{3.7}$$

Equation 3.8 defines $\sigma\left(\hat{p}_i\right)$:

$$\sigma\left(\hat{p}_i\right) = \sum_{i=1}^{n} \frac{\hat{p}_i \left(1 - \hat{p}_i\right)}{n}. \tag{3.8}$$

Equation 3.9 defines the confidence interval (CI) of the sample proportion:

$$CI = \hat{p}_i \left(-z < \sum_{i=1}^{n} \frac{\hat{p}_i \left(1 - \hat{p}_i\right)}{n} < z \right), \tag{3.9}$$

where $0 < CI < 1$ and $\pm\, z$ denotes critical values.

3.6 The Strategy used to Scale Economic Data of South Africa

Machine learning models require features to fit in a scale before modeling, so objective functions perform calculations efficiently. This research used the data standardization scaling strategy to designate feature values in the training data set into a standard scale to ensure values are contained in a standard scale for apt model comparison (subsection 3.6.1).

3.6.1 The Data Standardization Scaling Strategy

The data standardization scaling strategy alters the probability distribution, whereby $\mu_i = 0$, $\sigma_i = 1$, and x_i is in proximity to 0, so features proceed in a *Gaussian* process (or follow a normal distribution) [Tabak 2004]. Equation 3.10 defines the data standardization scaling strategy:

$$Scale_{standard} = \sum_{i=1}^{n} \frac{x_i - \hat{x}_i}{\sigma_i}, \tag{3.10}$$

where x_i denotes feature values, and $\sigma(x_i)$ denotes the divergence of x_i from $\bar{x}_i$. Equation 3.11 defines σ_i:

$$\sigma(x_i) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}. \tag{3.11}$$

3.7 Experiments for Model Comparison

Table 3.4 shows various experiments considered for model comparison.

TABLE 3.4: Experiments considered when determining whether machine learning models perform better than the benchmark model when forecasting the reaction of stock market returns to changing economic activities in South Africa

Scenario	Description	Principal Component Analysis	Regression	Tree-based	Neural Network	Vector Autoregressive

Continued on next page

Table 3.4 – Continued from previous page

Scenario	Description	Principal Component Analysis	Regression	Tree-based	Neural Network	Vector Autoregressive
Scenario 1	Investigate the dynamics of the economy and how stock market returns react to changing economic activities in South Africa using the default vector autoregressive model, along with default supervised machine learning models (i.e., the ordinary least-squares regression model, ridge model, least absolute shrinkage and selection operator model, elastic net model, decision tree model, random forest tree model, extreme gradient boosting tree model, recurrent neural network, gated recurrent unit, long-short term memory, restricted Boltzmann machine, and multi-layer perceptron). Following that, compare the performance of default supervised machine learning models against the performance of the benchmark model (the default vector autoregressive model) when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.		*	*	*	*
Scenario 2	Present feature subsets containing key economic features selected based on the *gini impurity* value calculated by tree-based models. Concurrently, compare the performance of models trained on a feature subset to the performance of models trained on whole features when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.		*	*	*	
Scenario 3	Produce low economic data dimensions found by the principal component analysis method. Subsequently, compare the performance of models trained on low economic data dimensions to the performance of models trained on high economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.	*	*	*	*	

3.8 Strategies used to Regularize Regression Models

Machine learning models operate a ι_1 norm function or ι_2 norm function (or both) to confine model parameters by introducing a penalty term. The ridge model regularization strategy used the ι_{2-norm} function, least absolute shrinkage and selection operator model regularization strategy used the ι_{1-norm} function, and elastic net model regularization strategy used the ι_{2-norm} function and ι_{2-norm} function.

3.8.1 The Ridge Model Regularization Strategy

The ridge model regularization strategy, equally known as the ι_2 norm model regularization strategy, stabilizes the *bias* and *variance (var)* by applying a strengthening parameter similar to weights of β_t while learning predictor feature values c. Equation 3.12 defines the ridge model regularization strategy:

$$\hat{\beta}_i = \sum_{i=1}^{n} \left(\hat{y}_i - \sum_{j=1}^{n} \hat{\beta}_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} |\hat{\beta}_j|, \tag{3.12}$$

where λ denotes the penalty term, leading to Equation 3.13:

$$\hat{\beta}_i = \left(x_i' x_i + \lambda I \right)^{-1} \left(x_i' \hat{y}_i \right). \tag{3.13}$$

The penalty term penalizes a machine learning model, as it commits errors when learning predictor feature values and forecasting concealed target feature values. Equation 3.14 defines the *bias*:

$$bias = \lambda \left(x_i' x_i + \lambda I \right)^{-1} \hat{\beta}_i. \tag{3.14}$$

Equation 3.15 defines the *var*:

$$var = \lambda \left(x_i' x_i + \lambda I \right)^{-1}, x_i' x_i \left(x_i' x_i \lambda I \right)^{-1}. \tag{3.15}$$

3.8.2 The Least Absolute Shrinkage and Selection Operator Model Regularization Strategy

The least absolute shrinkage and selection operator model regularization strategy, equally known as the ι_1 norm model regularization strategy, enhances the performance of machine learning models by normalizing parameters of models and penalizing model residuals [Tibshirani 1996]. Equation 3.16 defines the least absolute shrinkage and selection operator model regularization strategy:

$$\hat{\beta}_i = \sum_{i=1}^{n} \left(\hat{y}_i - \sum_{j=1}^{n} \hat{\beta}_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} \left(\hat{\beta}_j \right)^2. \tag{3.16}$$

3.8.3 The Elastic Net Model Regularization Strategy

The elastic net regularization strategy bundles the λ of the ridge model regularization strategy and least absolute shrinkage and selection operator model regularization strategy to manage *bias* and *var*, provided a limited n with higher dimensions [Zou and Hastie 2005a]. It achieves this by eliminating invaluable features and prioritizing noteworthy features, and contains a quadratic function (λ) [Zou and Hastie 2005b; Meier *et al.* 2008].

Equation 3.17 defines the elastic net model regularization strategy:

$$\hat{\beta}_i = \frac{\sum_{i=1}^{n} \left(\hat{y}_t - x_i' \hat{\beta}_j\right)^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{n} \hat{\beta}_i\right)^2 \alpha \sum_{j=1}^{n} |\hat{\beta}_j|, \tag{3.17}$$

where $\alpha = 0$ for the least absolute shrinkage and selection operator model regularization strategy, and $\alpha = 1$ for the ridge model regularization strategy.

3.9 The Metric used to Evaluate the Performance of Models

This research identifies errors models commit when forecasting reactions of stock market returns to evaluate model performance. The *mean absolute percentage error* metric (subsection 3.9.1) evaluates the benchmark model (the vector autoregressive model), along with candidate machine learning models, when forecasting the reaction of stock market returns to changing economic activities in South Africa.

3.9.1 The Mean Absolute Percentage Error Metric

Compared to other regression model performance metrics (i.e., the *mean squared error* metric, *root mean squared error* metric, and *mean absolute error* metric, among other regression model performance metrics), the *mean absolute percentage error (MAPE)* metric or *mean absolute percentage deviation (MAPD)* metric is useful in evaluating the performance of regression-based models when the data set contains temporal features, since the metric is intuitive in model interpretation [Swamidass 2000].

The *MAPE* metric denotes the divergence ratio of $\hat{y}_i$ from y_i. It does not consider positive divergence ratio values or negative divergence ratio values, and it helps prevent a scenario in which positive errors and negatives do not revoke each other, which is common with alternative regression-based model performance metrics [Myttenaere *et al.* 2016]. Equation 3.18 defines the $MAPE$ metric:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right|. \tag{3.18}$$

The $MAPE$ metric serves as the primary model performance metric in this research. A model with a $MAPE$ value in proximity to 100% is ideal. Investors may use the *MAPE* metric to inform investing decisions.

If the *MAPE* value is incredibly high, they may reconsider investing in the stock market in South Africa or exclude the index from their portfolio. Else, if the *MAPE* value is exceptionally low, they can consider investing in the stock market in South Africa or include the index in their portfolio.

3.10 Ethical Consideration

This research extracts monthly economic features from FRED and SARB databases, and the monthly stock market price in South Africa from the JSE database. All databases are for general access and academic use. This research does not require an ethics clearance certificate, because all features come from secondary data sources.

3.11 Research Methods Summary

Table 3.5 shows links between research questions and research methods.

TABLE 3.5: Research methods summary

Research Question	Research Method
How do stock market returns react to changing economic activities in South Africa? Do machine learning models perform better than the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?	Investigate the dynamics of the economy and how stock market returns react to changing economic activities in South Africa using the default vector autoregressive model, along with default supervised machine learning models (i.e., the ordinary least-squares regression model, ridge model, least absolute shrinkage and selection operator model, elastic net model, decision tree model, random forest tree model, extreme gradient boosting tree model, recurrent neural network, gated recurrent unit, long-short term memory, restricted Boltzmann machine, and multi-layer perceptron). Following that, compare the performance of default supervised machine learning models against the performance of the benchmark model (the default vector autoregressive model) when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.
Do models trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by tree-based models perform better than the models trained on whole features when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?	Present feature subsets containing key economic features selected based on the *gini impurity* value calculated by tree-based models. Concurrently, compare the performance of models trained on a feature subset to the performance of models trained on whole features when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.
Is the performance of models trained on low economic data dimensions (found by the principal component analysis method) distinguishable from the performance of models trained on high economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?	Produce low economic data dimensions found by the principal component analysis method. Subsequently, compare the performance of models trained on low economic data dimensions to the performance of models trained on high economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.

4 Experiment Results & Discussions

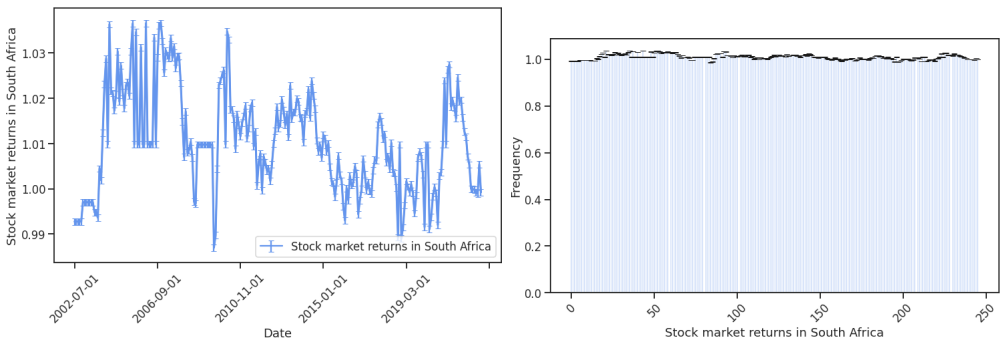
This research investigated the dynamics of the economy and how stock market returns react to changing economic activities in South Africa. It considered various experiments, i.e., different feature subsets and data dimensions, to determine whether machine learning models perform better than the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns.

Download the source code and data of the research project here:

http://www.github.com/tshepochris

4.1 The Exploration of the Distribution of Stock Market Returns Data in South Africa

Figure 4.1 shows the distribution of stock market returns in South Africa of the same sample period.



(A) Series (B) Series Distribution

FIGURE 4.1: The distribution of stock market returns data in South Africa

Figure 4.1 (A) shows reversion or steady stock market returns that calm the market after extreme lows or highs from 2002 to 2022. (B) shows stock market returns proceed in a Gaussian process.

Table 4.1 shows the *p value* of the augmented Dickey-Fuller *(ADF)* test, which infers stock market returns stationarity (at $\alpha = 0.05$), where a *p value* < 0.05 denotes the stock market returns in South Africa are stationary, and *p value* > 0.05 denotes the feature is non-stationary.

It also shows the central tendency and divergence of stock market returns in South Africa (i.e., the *mean* value, *standard deviation* value, *skew* value, and *kurtosis* value).

TABLE 4.1: The descriptive statistics of stock market returns data in South Africa

	Mean	**Std**	**Skew**	**Kurtosis**	**ADF p-value**
Stock market returns in South Africa	1.0105	0.0113	0.3357	-0.4565	0.0055

Table 4.1 shows the *mean* stock market returns in South Africa value is 1.0105, the *standard deviation* is 0.0113, the *skew* is 0.3357, and the *kurtosis* is -0.4565. The *p value* of the ADF test is 0.0055, suggesting stationary stock market returns in South Africa, which is a requirement for the vector autoregressive model.

4.1.1 The Granger-Causal Relationship between Economic Activities and Stock Market Returns in South Africa

Table 4.2 shows *p values* of the Granger-Causality test about the causal relationship between selected economic activities and the stock market returns in South Africa, whereby a *p value* < 0.05 entails Granger-Causality among features, and a *p value* > 0.05 entails no Granger-Causality among features.

TABLE 4.2: The Granger-Causality relationship between selected economic activities and the stock market returns in South Africa

	x1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**
y1	1.0000	0.0500	0.0163	0.0027	0.0472	0.3143	0.1244	0.0660	0.8230	0.4134
y2	0.0001	1.0000	0.1957	0.2711	0.1442	0.1034	0.5286	0.8887	0.3102	0.0276
y3	0.0043	0.0000	1.0000	0.0003	0.3358	0.0772	0.0580	0.0311	0.4313	0.1157
y4	0.3339	0.0408	0.0396	1.0000	0.1201	0.0756	0.1065	0.0314	0.6256	0.6680
y5	0.0139	0.0049	0.0888	0.1613	1.0000	0.1297	0.0007	0.0894	0.5912	0.1686
y6	0.0332	0.0007	0.0142	0.3642	0.4706	1.0000	0.1902	0.2288	0.1049	0.2237
y7	0.0033	0.0262	0.0287	0.0008	0.0000	0.3392	1.0000	0.0001	0.0887	0.2822
y8	0.5282	0.1685	0.1077	0.0431	0.0096	0.1919	0.0000	1.0000	0.3623	0.2366
y9	0.0700	0.0045	0.0695	0.2368	0.3477	0.0882	0.7771	0.0789	1.0000	0.6431
y10	0.3490	0.1124	0.0017	0.0057	0.2584	0.0012	0.0199	0.4887	0.0168	1.0000

At most, *p value* < 0.05, implying the Granger-Causality between most economic features and the stock market returns in South Africa.

This research considered a manageable set of economic features for the Granger-Causality test due to multiple economic features. Table 4.3 shows selected economic features in the Granger-Causality matrix.

TABLE 4.3: Selected economic features in the Granger-Causality Matrix

Legend
Stock market returns in South Africa
Coincident indicator (2015=100)
Sales: wholesale trade: value
LL OECD: reference series: gdp trend
Long-term government stock
Broad money and components: M3
Total liabilities to the public
Required liquid assets
Real effective exchange rates based on manufacturing consumer price
Number of shares traded (millions)

4.2 The Benchmark for Model Comparison

The vector autoregressive was repeatedly used in previous research on the reaction of stock market returns to changing economic activities in South Africa. As such, it was the benchmark model. The structural vector autoregressive identified the Granger-Causal relationship between selected economic activities and the stock market returns in South Africa.

4.2.1 The Performance of the Vector Autoregressive Model when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa

The vector autoregressive model firmly maintained a stochastic-oriented system that handled some changing economic activities as linear combinations of p lags, along with p lags, then cleared up the temporal problem.

Table 4.4 shows the performance of the vector autoregressive model when forecasting the reaction of stock market returns to changing economic activities in South Africa.

TABLE 4.4: The performance of the vector autoregressive model when forecasting the reaction of stock market returns to changing economic activities in South Africa

	Mean absolute percentage error					
Model	**Selection**	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
VAR	Default	0.022604	0.004828	0.004828	0.029785	0.0155
VAR	DT	0.008941	0.007010	0.007010	0.010541	0.0084
VAR	RFT	0.008941	0.007010	0.007010	0.010541	0.0084
VAR	XGBT	0.008941	0.007010	0.007010	0.010541	0.0084

Table 4.4 shows the default vector autoregressive model trained on whole features achieved the highest average *MAPE* value (at 0.0155). Default vector autoregressive models trained on feature subsets containing key economic features selected based on the *gini impurity* value calculated by the decision tree model, random forest tree model, and extreme gradient boosting tree model achieved the lowest average *MAPE* value (at 0.0084).

4.3 The Performance of Default Machine Learning Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa

Used machine learning models include regression models (i.e., the ordinary least-squares model, ridge model, and least absolute shrinkage and selection operator model), along with tree-based models (i.e., the decision tree model, random forest tree model, and extreme gradient boosting tree model) and neural networks (i.e., the recurrent neural network, gated recurrent unit, long-short term memory, restricted Boltzmann machine, and multi-layer perceptron).

4.3.1 The Performance of Default Regression Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa

The ordinary least-squares regression model fitted a straight line to forecast the reaction of stock market returns to changing economic activities in South Africa, while considering model residuals. This segment also considered other default regression models (i.e., the ridge model, least absolute shrinkage and selection operator model, and elastic net model).

Table 4.5 shows the performance of default regression models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.5: The performance of default regression models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error				
Model	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
OLS	0.0101	0.0074	0.0109	0.0136	0.0105
Ridge	0.0090	0.0073	0.0084	0.0102	0.0087
Lasso	0.0094	0.0103	0.0075	0.0080	0.0088
ElasticNet	0.0094	0.0103	0.0075	0.0080	0.0088

Table 4.5 shows the default ordinary least-squares model has the highest average *MAPE* value (at 0.0105), and the default ridge model has the lowest average *MAPE* value (at 0.0087). The highest-performing default regression model performs 3% worse than the highest-performing default vector autoregressive model.

4.3.2 The Performance of Default Tree-Based Models when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa

Through the recursive data partitioning strategy, tree-based models developed tree-like structures to forecast the reaction of stock market returns to changing economic activities in South Africa. Table 4.6 shows the performance of default tree-based models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.6: The performance of default tree-based models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error				
Model	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
DT	0.0052	0.0050	0.0078	0.0067	0.0061
RFT	0.0027	0.0063	0.0072	0.0062	0.0056
XGBT	0.0039	0.0039	0.0080	0.0099	0.0064

Table 4.6 shows the default extreme gradient boosting tree model has the highest average *MAPE* value (at 0.0064), and the default random forest tree model has the lowest average *MAPE* value (at 0.0056).

The highest-performing default tree-based model performs 50% better than the highest-performing default vector autoregressive model, and 55% better than the highest-performing default regression model.

Nengovhela [2022] used the decision tree model and random forest tree model to forecast the reaction of stock market returns to changing economic activities in South Africa. The random forest tree model outperformed the decision tree model as measured by the *MAE* metric.

The findings of this research are inconsistent with previous research, the random forest tree model outperforms other default tree-based models.

4.3.3 The Performance of Default Neural Networks when Forecasting Reaction of Stock Market Returns to Changing Economic Activities in South Africa

Neural networks (i.e., the recurrent neural network, gated recurrent unit, long-short term memory, restricted Boltzmann machine, and multi-layer perceptron) forecast the reaction of stock market returns to changing economic activities in South Africa by designating dissimilar *weights* and *biases* at each neuron of each layer in the neural network infrastructure through an iterative process.

Table 4.7 shows the performance of default neural networks when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.7: The performance of default neural networks when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error				
Model	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
RBM	1.0000	1.0000	1.0000	1.0932	1.0233
MLP	0.9145	1.4740	2.1006	3.8992	2.0970
RNN	0.7380	0.9501	0.7931	0.8421	0.8308
GRU	0.3798	0.6116	1.0859	1.2058	0.8208
LSTM	0.3786	0.4657	0.7951	0.9007	0.6350

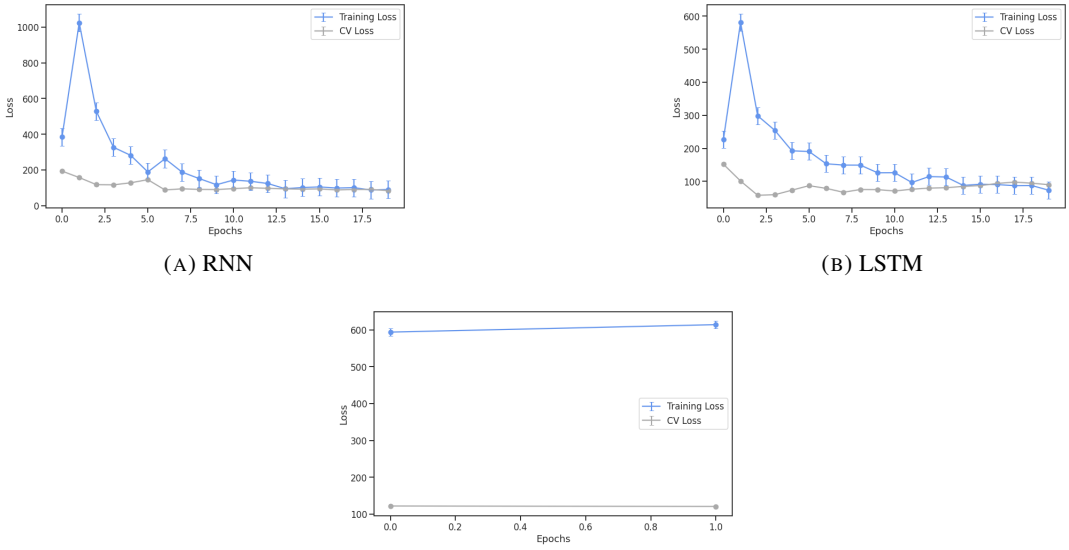
Table 4.7 shows the default multi-layer perceptron has the highest average *MAPE* value (at 2.0970), and the default long-short term memory has the lowest average *MAPE* value (at 0.6350).

The highest-performing default neural model performs 99% worse than the highest-performing default vector autoregressive model, 98% worse than the highest-performing default regression model, and 99% worse than the highest-performing default tree-based model.

Previous research, such as Pahlawan *et al.* [2021b] used a recurrent neural network, gated recurrent unit, and long-short term memory to forecast the reaction of stock market returns. The long-short term memory had the lowest *mean squared error* value (at 1.35).

Equally, Fu *et al.* [2018] used a deep belief network to forecast stock market returns to changing economic activities, along with technical indicators. The highest-performing deep belief network had a *MAPE* of 0.97.

The *loss* function (or *cost* function) evaluates the performance of a model. Figure 4.2 shows the *loss* function value of default neural networks across epochs.



(A) RNN

(B) LSTM

(C) GRU

FIGURE 4.2: The *loss* function value across epochs of default neural networks when forecasting the reaction of stock market returns to changing economic activities in South Africa

Figure 4.2 (A) shows the training *loss* function value is above the cross-validation *loss* function value of the recurrent neural network across all *epochs*. (B) shows the training *loss* function value is above the cross-validation *loss* function value of the long-short term memory across all *epochs*. (C) shows the training *loss* function value and cross-validation *loss* function value of the gated recurrent unit barely change across all *epochs*.

4.4 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by Tree-based Models

The *gini impurity* estimator identified key economic features to tree-based models, i.e., the decision tree model (subsection 4.4.1), random forest tree model (subsection 4.4.2), and extreme gradient boosting tree model (subsection 4.4.3), because of their convenience. Feature importance is useful in model interpretation, as it allows modelers to identify the importance of features, then devise approaches for modifying features to achieve the desired effect.

4.4.1 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by the Decision Tree Model

The decision tree model selected a feature subset containing key economic features selected based on the *gini impurity* value. Subsequently, economic features were ranked based on the *gini impurity* value (Figure 4.3).

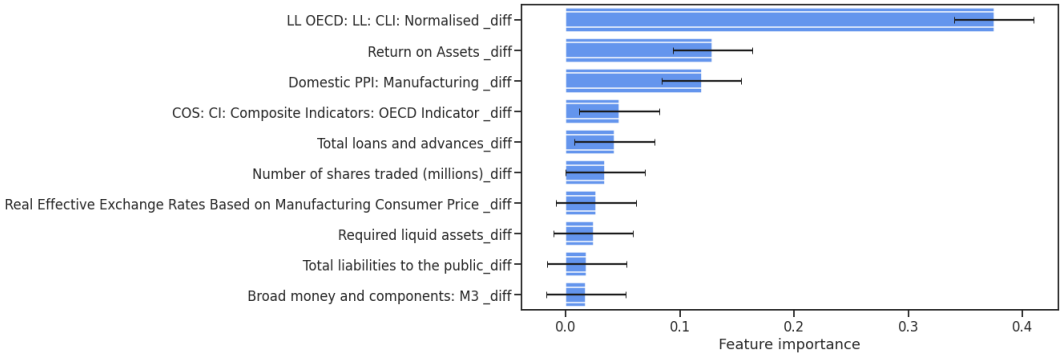


FIGURE 4.3: The top ten economic features in a feature subset selected by the decision tree model, as measured by the *gini impurity* value

Figure 4.3 shows the most economic features in South Africa to the decision tree model is the *LL OECD: LL: CLI: normalised* with *gini impurity = 0.3752*, followed by *Return on Assets* with *gini impurity = 0.1284*, then *domestic PPI: manufacturing* with *gini impurity = 0.1189*.

Broad money and components: M3 represents the least important economic activity in South Africa, with *gini impurity = 0.0176*.

4.4.2 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by the Random Forest Tree Model

The random forest tree model selected a feature subset containing key economic features selected based on the *gini impurity* value. Concurrently, economic features were ranked based on the *gini impurity* value (Figure 4.4).

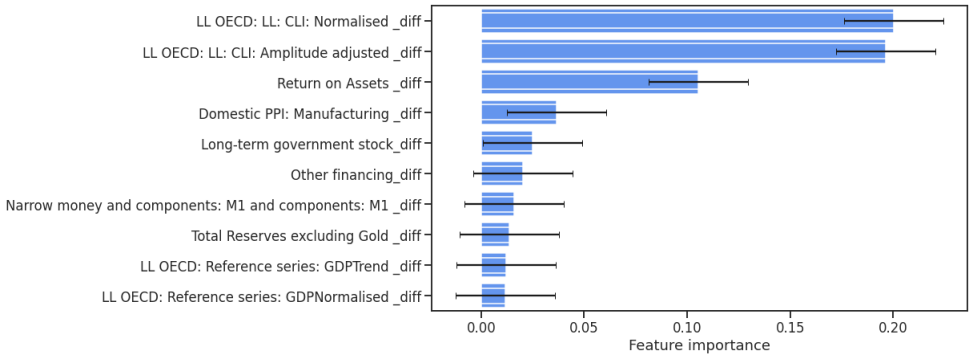


FIGURE 4.4: The top ten economic features in a feature subset selected by the random forest tree model, as measured by the *gini impurity* value

Figure 4.4 shows the most economic features in South Africa to the random forest tree model is the *LL OECD: LL: CLI: normalised* with *gini impurity = 0.200176*, followed by *LL OECD: LL: CLI: amplitude adjusted* with *gini impurity = 0.196495*, then *Return on Assets* with *gini impurity = 0.105387*.

LL OECD: reference series: GDP normalized represents the least important economic activity in South Africa, with *gini impurity = 0.0118*.

4.4.3 The Selection of a Feature Subset Containing Key Economic Features based on the *Gini Impurity* Value Calculated by the Extreme Gradient Boosting Tree Model

The extreme gradient boosting tree model selected a feature subset containing key economic features selected based on the *gini impurity* value. Following that, economic features were ranked based on the *gini impurity* value (Figure 4.5).

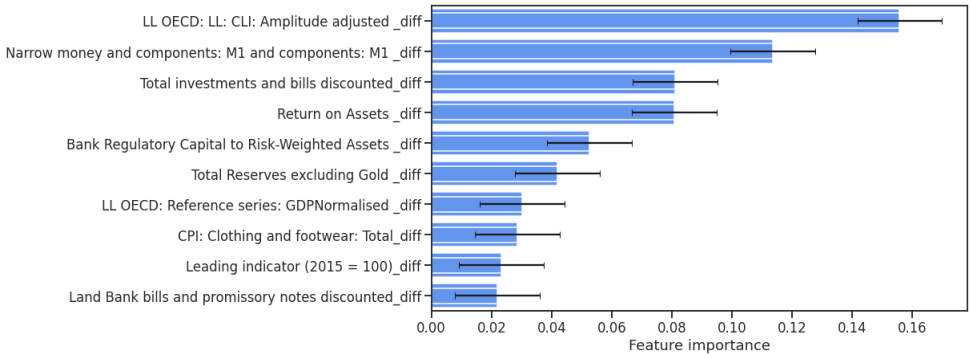


FIGURE 4.5: The top ten economic features in a feature subset selected by the extreme gradient boosting tree model, as measured by the *gini impurity* value

Figure 4.5 shows the most economic features in South Africa to the extreme gradient tree model is the *LL OECD: LL: CLI: amplitude adjusted* with *gini impurity = 0.1557*, followed by *narrow money and components: M1 and components: M1* with *gini impurity = 0.1136*, then *total investments and bills discounted* with *gini impurity = 0.0810*.

Land bank bills and promissory notes represents the least important economic activity in South Africa, with *gini impurity = 0.0220*.

4.4.4 The Performance of Models Trained on a Feature Subset Containing Key Economic Features when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa

Following the identification of key economic features. The concurrent step was to compare the performance of models trained on a feature subset containing key economic features against the performance of models trained on whole features.

Table 4.8 shows the performance of models trained on a feature subset containing key economic features when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.8: The performance of models trained on a feature subset containing key economic features when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error					
Model	**Selection**	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
OLS	DT	0.0081	0.0062	0.0079	0.0094	0.0079
Ridge	DT	0.0081	0.0062	0.0079	0.0093	0.0079
Lasso	DT	0.0094	0.0103	0.0075	0.0080	0.0088
ElasticNet	DT	0.0094	0.0103	0.0075	0.0080	0.0088
DT	DT	0.0038	0.0072	0.0089	0.0092	0.0073
RFT	DT	0.0049	0.0070	0.0075	0.0059	0.0063
XGBT	DT	0.0034	0.0059	0.0044	0.0068	0.0051
RBM	DT	0.6559	0.6410	0.5293	0.5824	0.6022
MLP	DT	0.1045	0.1618	0.2727	0.2010	0.1850
RNN	DT	7.7934	1.0464	1.0468	1.2549	2.7854
LSTM	DT	2.3890	1.0430	1.1607	0.9286	1.3803
GRU	DT	1.5455	0.6883	0.8243	0.9703	1.0071
OLS	RFT	0.0043	0.0052	0.0074	0.0077	0.0061
Ridge	RFT	0.0045	0.0053	0.0073	0.0075	0.0062
Lasso	RFT	0.0094	0.0103	0.0075	0.0080	0.0088
ElasticNet	RFT	0.0094	0.0103	0.0075	0.0080	0.0088
DT	RFT	0.0054	0.0072	0.0076	0.0138	0.0085
RFT	RFT	0.0025	0.0079	0.0065	0.0065	0.0058
XGBT	RFT	0.0045	0.0070	0.0064	0.0100	0.0070
RBM	RFT	0.5971	0.3697	0.6759	0.5518	0.5486
MLP	RFT	0.1019	0.2209	0.2843	0.1269	0.1835
RNN	RFT	2.8547	1.3562	7.3089	1.9658	3.3714
LSTM	RFT	2.4146	0.7562	0.8271	0.9345	1.2331
GRU	RFT	0.6070	1.3584	0.8998	0.9864	0.9629
OLS	XGBT	0.0041	0.0072	0.0081	0.0089	0.0071
Ridge	XGBT	0.0042	0.0072	0.0081	0.0089	0.0071
Lasso	XGBT	0.0094	0.0103	0.0075	0.0080	0.0088
ElasticNet	XGBT	0.0094	0.0103	0.0075	0.0080	0.0088
DT	XGBT	0.0040	0.0031	0.0103	0.0052	0.0056
RFT	XGBT	0.0037	0.0056	0.0066	0.0053	0.0053
XGBT	XGBT	0.0033	0.0040	0.0070	0.0077	0.0055
RBM	XGBT	1.0000	1.0000	1.0000	0.9602	0.9900
MLP	XGBT	0.1726	0.2236	0.2845	0.2990	0.2449

Continued on next page

Table 4.8 – continued from previous page

Model	Selection	H1	H2	H3	H4	AVG MAPE
RNN	XGBT	25.7889	2.0354	2.3727	1.1041	7.8253
LSTM	XGBT	2.3909	1.9734	0.9071	0.7706	1.5105
GRU	XGBT	0.5378	1.4749	0.7689	1.0917	0.9683

Table 4.8 shows the recurrent neural network trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the extreme gradient boosting tree model has the highest average *MAPE* value (at 7.8253), and the extreme gradient boosting model trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the decision tree model has the lowest average *MAPE* value (at 0.0051).

The extreme gradient boosting tree model trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the decision tree model performed 65% better than the highest-performing default vector autoregressive, 71% better than the highest-performing default regression model trained on key economic features, 10% better than the highest-performing default tree-based model trained on key economic features, and 12 351% better than the highest-performing default neural network.

4.5 The Reduction of Economic Data Dimensions

This research used the principal components analysis method to reduce economic data dimensions. The following segment shows economic features of South Africa in various dimensions (subsection 4.5.2).

4.5.1 Economic Features of South Africa in Different Dimensions

It is obscure to visualize all the economic features of South Africa. The principal component analysis method reduced dimensions of economic activities into meaningful eigenvectors to ensure interpretation by components.

Figure 4.6 shows the economic features of South Africa in two dimensions.

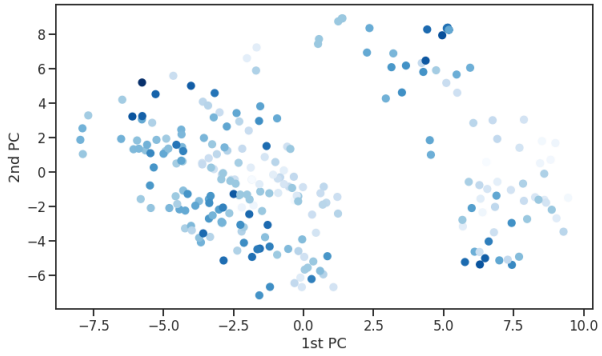


FIGURE 4.6: Economic features of South Africa in two dimensions

Figure 4.6 shows multiple condensed 1^{st} principal components were vastly concentrated before approaching 0. However, after approaching 0, 2^{nd} principal components diminished and scattered vastly, with an upsurge in 1^{st} principal components.

Figure 4.7 shows economic features of South Africa in three dimensions.

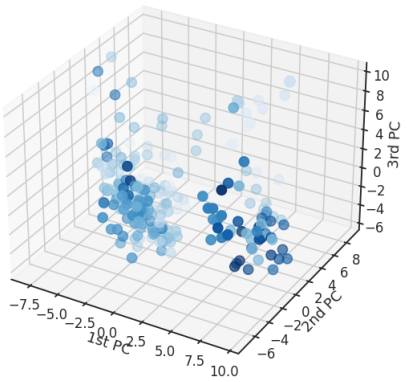


FIGURE 4.7: Economic features of South Africa in three dimensions

Figure 4.7 confirms multiple principal components before 0, in the 1^{st} principal component axis, but few principal components in the $2nd$ principal component axis. There were multiple principal components in the 3^{rd} principal component axis.

4.5.2 An Index that Offers Insight into the Structure of the Economy in South Africa

The principal component analysis method produced an index that offers insight into the structure of the economy in South Africa based on the explained variance *(EV)* ratio.

Table 4.9 shows an index that provides insight into economic activities in South Africa (Table B.1 for a thorough list of economic activities in the index) based on the explained variance *(EV)* ratio.

TABLE 4.9: An index that provides insight into economic activities in South Africa

Rank	Economic activities	EV Ratios
1	Business Tendency Surveys for Manufacturing: CI: Composite Indicators: OECD Indicator	20.2738
2	Real Effective Exchange Rates Based on Manufacturing Consumer Price	14.0164
3	National Currency to US Dollar Exchange Rate: Average of Daily Rates	12.9776
4	National Currency to US Dollar Spot Exchange Rate	8.0563
5	CPI: Total Food Excluding Restaurants	6.3136
6	CPI: OECD Groups: Fuel, Electricity, and Gasoline	5.8097
7	CPI: OECD Groups: All Items Non-Food and Non-Energy	4.6073
8	Commercial Real Estate Loans to Total Loans	4.3957
9	COS: CI: Composite Indicators: OECD Indicator	3.8083
10	U.S. Exports of Goods by F.A.S. Basis to South Africa	3.1414

Table 4.9 shows the *business tendency surveys for manufacturing: CI: composite indicators: OECD indicator* explains most of the variance in changing economic activities with

the *explained variance ratio = 20.2738*, followed by *real effective exchange rates based on manufacturing consumer price* with the *explained variance ratio = 14.0164*, then *national currency to U.S. dollar exchange rate: average of daily rates* with the *explained variance ratio = 8.0563*.

The *U.S. exports of goods by F.A.S. basis to South Africa* explains the least of the variance in the top ten economic activities in the index, with an *explained variance ratio = 3.1414*.

4.5.3 The Performance of Models Trained on Low Economic Data Dimensions when Forecasting the Reaction of Stock Market Returns to Changing Economic Activities in South Africa

The principal component analysis method reduced economic data dimensions into meaningful eigenvectors. This segment compares the performance of models trained on low economic data dimensions to the performance of models trained on high economic data dimensions.

Table 4.10 shows the performance of models trained on low economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.10: The performance of models trained on low economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error				
Model	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
OLS	1.5017	0.9991	3.7365	3.2257	2.3657
Ridge	1.0154	0.8492	4.2344	4.0395	2.5346
Lasso	1.0142	1.0272	0.9883	0.9913	1.0052
ElasticNet	1.0142	1.0272	0.9883	0.9913	1.0052
DT	0.8581	1.2273	1.4002	2.8004	1.5715
RFT	0.4335	0.9128	1.1147	1.1659	0.9067
XGBT	0.6600	0.7343	0.9158	1.2308	0.8852
RBM	1.0906	1.0373	1.7111	1.0610	1.2250
MLP	1.3927	1.9643	3.0129	2.6656	2.2589
RNN	0.6336	0.4512	1.5063	1.3315	0.9806
LSTM	0.3781	0.5830	0.9795	0.9520	0.7232
GRU	0.7722	0.6206	0.7658	0.9466	0.7763

Table 4.10 shows the ridge model trained on low economic data dimensions has the highest average *MAPE* value (at 2.5346), and the long-short term memory trained on low economic data dimensions has the lowest average *MAPE* value (at 0.7232).

The long-short term memory trained on low economic data dimensions performed 99% worse than the default vector autoregressive model, 99% worse than the highest-performing default regression model, 99% worse than the highest-performing default tree-based model, 12% worse than the highest-performing default neural network, and 99% worse than the highest-performing model trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by tree-based models.

Previous research, such as Ahangar *et al.* [2010] used the principal components analysis method to reduce dimensions of economic data, then forecast the reaction of stock market

returns to changing economic activities in Iran. Moreover, the research only trained a deep belief network with three *hidden* layers and 14 neurons, which achieved a *mean squared error* value for the deep belief network was 31.6.

4.6 Overall Performance Summary

Table 4.11 shows the top ten highest-performing machine learning models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.11: The top ten highest-performing machine learning models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error					
Model	**Selection**	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
XGBT	DT	0.0034	0.0059	0.0044	0.0068	0.0051
RFT	XGBT	0.0037	0.0056	0.0066	0.0053	0.0053
XGBT	XGBT	0.0033	0.004	0.007	0.0077	0.0055
RFT	NA	0.0027	0.0063	0.0072	0.0062	0.0056
DT	XGBT	0.004	0.0031	0.0103	0.0052	0.0056
RFT	RFT	0.0025	0.0079	0.0065	0.0065	0.0058
DT	NA	0.0052	0.005	0.0078	0.0067	0.0061
OLS	RFT	0.0043	0.0052	0.0074	0.0077	0.0061
Ridge	RFT	0.0045	0.0053	0.0073	0.0075	0.0062
RFT	DT	0.0049	0.007	0.0075	0.0059	0.0063

Figure 4.8 shows the top performance of each machine learning model over four periods.

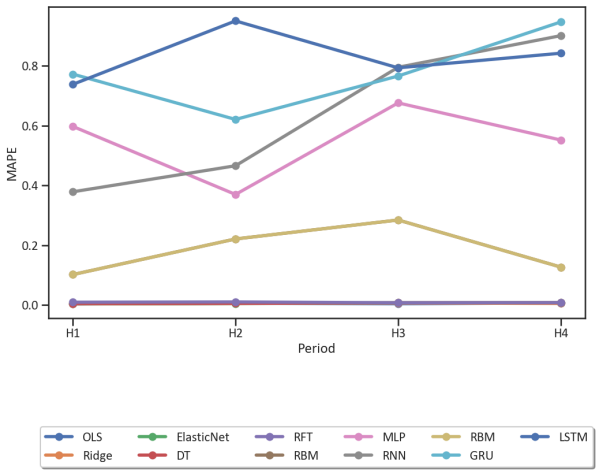


FIGURE 4.8: The top performance of each machine learning model over four periods

Table 4.12 shows the top ten worst-performing machine learning models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods.

TABLE 4.12: The top ten worst-performing machine learning models when forecasting the reaction of stock market returns to changing economic activities in South Africa over four periods: 3 months $(H1)$, 6 months $(H2)$, 12 months $(H3)$, and 24 months $(H4)$

	Mean absolute percentage error					
Model	**Selection**	**H1**	**H2**	**H3**	**H4**	**AVG MAPE**
RNN	XGBT	25.7889	2.0354	2.3727	1.1041	7.8253
RNN	RFT	2.8547	1.3562	7.3089	1.9658	3.3714
RNN	DT	7.7934	1.0464	1.0468	1.2549	2.7854
Ridge	PCA	1.0154	0.8492	4.2344	4.0395	2.5346
OLS	PCA	1.5017	0.9991	3.7365	3.2257	2.3657
MLP	PCA	1.3927	1.9643	3.0129	2.6656	2.2589
MLP	NA	0.9145	1.474	2.1006	3.8992	2.097
DT	PCA	0.8581	1.2273	1.4002	2.8004	1.5715
LSTM	XGBT	2.3909	1.9734	0.9071	0.7706	1.5105
LSTM	DT	2.389	1.043	1.1607	0.9286	1.3803

Table 4.11 shows the recurrent neural network trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the extreme gradient boosting tree model has the highest average *MAPE* value. The model performs worse than all candidate models.

Nengovhela [2022] found that the top five economic features when forecasting the reaction of stock market returns to economic activities in South Africa are: 1) gold, 2) Morgan Stanley Capital International (MSCI) emerging markets currency index, 3) MSCI all-country world index, 4) MSCI developing markets index, and 5) U.S. 10-year yield.

Table 4.13 shows the ranking economic features based on the *gini impurity* value calculated by each tree-based model.

TABLE 4.13: The ranking of economic features based on the *gini impurity* value calculated by each tree-based model

Decision Tree	**Random Forest Tree**	**Extreme Gradient Boosting Tree**
LL OECD: LL: CLI: normalized	LL OECD: LL: CLI: normalized	LL OECD: LL: CLI: amplitude adjusted
Return on Assets	LL OECD: LL: CLI: amplitude adjusted	Narrow money and components: M1 and components: M1
Domestic PPI: manufacturing	Return on assets	Total investments and bills discounted
COS: CI: composite indicators: OECD Indicator	Domestic PPI: manufacturing	Return on assets
Total loans and advances	Long-term government stock	Bank regulatory capital to risk-weighted assets

Table 4.13 shows *LL OECD: LL: CLI: normalized* was the most important economic feature in feature subsets selected based on the *gini impurity* value calculated by both the decision tree model and random forest tree model.

Figure 4.9 shows the top performance of each tree-based feature selection strategy over four periods.

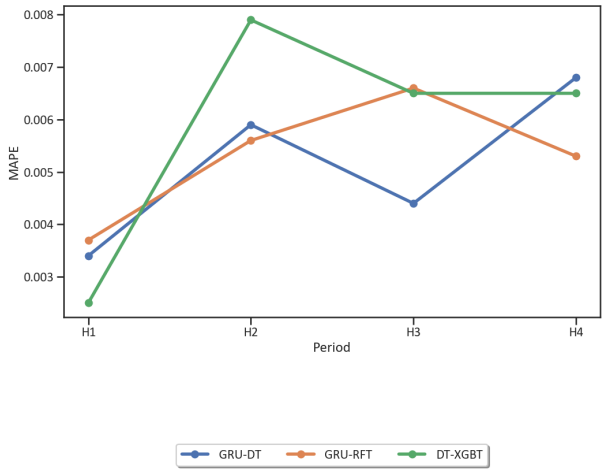


FIGURE 4.9: The top performance of each tree-based feature selection strategy

The optimal model is the extreme gradient boosting tree model trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the decision tree model.

Figure 4.10 shows the learning curve of the optimal model that forecasts the reaction of stock market returns to changing economic activities in South Africa.

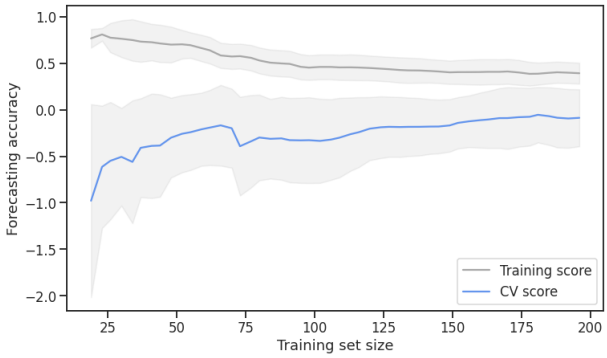


FIGURE 4.10: The learning curve of the optimal model that forecasts the reaction of stock market returns to changing economic activities in South Africa

Figure 4.10 shows that as the training data increases, the training score gradually diminishes and the cross-validation score gradually intensifies.

5 Conclusions & Future Research

5.1 Introduction

This research informs practitioners about the technical merits of machine learning models in forecasting the reaction of stock market returns to changing economic activities in the emerging economic context. Stock market returns in South Africa are erratic, thus widening the research on the machine learning approach is useful for those who seek to capture market complexities that define the financial district of Africa.

5.2 Literature Review

The inflexibility and incompetence of conventional statistical multivariate forecasting models in modeling non-linear relationships bring doubt to their inventiveness. The need for their alternatives has piqued the interest of researchers in investigating whether machine learning models perform better than them in forecasting the reaction of stock market returns to changing economic activities.

Equally, most research focused on whole features, few models for comparison, and barely focused on how different feature subsets and reduced dimensionality change model performance, a limitation this research addresses when considering the number of experiments.

This research established that machine learning models can be used to forecast the reaction of stock market returns to changing economic activities in the emerging economic context. It advances research by comparing various models trained on different feature subsets and economic data dimensions.

5.3 Research Methods

The initial task was to remove duplicate features before combining data sets for redundancy avoidance in analysis, rename features to their official names, concatenate data from different data sources, index, and format the data.

Subsequent steps were to specify the sample period, place feature values in decimal points, convert infinity values of a float type to *NaN* (missing feature values), impute missing feature values using the k nearest neighbor data imputation strategy, make features stationary, replace

outliers in the complete data set with *median* values, randomly partition data and standardize it, and reduce data dimensions using the principal components analysis method.

Last, this research considered various experiments, i.e., different feature subsets and data dimensions, to determine whether machine learning models perform better than the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns, as measured by the *mean absolute percentage error (MAPE)*.

5.4 Experimental Results

This research considered various experiments to compare the performance of machine learning models against the performance of the benchmark model (the vector autoregressive). Table 5.1 shows research questions, various experiments, and research findings:

TABLE 5.1: Research questions, various experiments, and research findings

Research Question	Experiment	Research Finding
How do stock market returns react to changing economic activities in South Africa? Do machine learning models perform better than the benchmark model (the vector autoregressive model) when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?	This research compared the performance of default supervised machine learning models to the performance of the benchmark model when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.	The default random forest tree model has the lowest average *MAPE* value (at 0.0056). The model performs better than the benchmark model and other default machine learning models, as measured by the average *MAPE* value.
Do models trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by tree-based models perform better than the models trained on whole features when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?	Presented feature subsets containing key economic features selected based on the *gini impurity* value calculated by tree-based models. Concurrently, compared the performance of models trained on a feature subset to the performance of models trained on whole features when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.	The extreme gradient boosting model trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the decision tree model has the lowest average *MAPE* value (at 0.0051).

Continued on next page

Table 5.1 – Continued from previous page

Research Question	Experiment	Research Finding
Is the performance of models trained on low economic data dimensions (found by the principal component analysis method) distinguishable from the performance of models trained on high economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric?	Produced low economic data dimensions found by the principal component analysis method. Subsequently, compared the performance of models trained on low economic data dimensions to the performance of models trained on high economic data dimensions when forecasting the reaction of stock market returns to changing economic activities in South Africa, as measured by the *MAPE* metric.	The long-short term memory trained on low economic data dimensions has the lowest average *MAPE* value (at 0.7232).

5.5 Policy Recommendations

Policymakers are interested in the dynamics of the economy, and the reaction of stock market returns to changing economic activities in South Africa, because it directs them to structure economic policies that reduce systematic market error and benefit the stock market.

The list below provides recommendations that policymakers must focus on to increase stock market returns in South Africa:

- ***Increase the coincident indicator:*** *The coincident indicator helps inform investors about the business cycle of an economy. It comprises core components like unemployment rate, industrial production turnover, and consumer savings..*

  Stock market returns decline when there is a decline in the coincident indicator in South Africa. Policymakers must promote economic initiatives that reduce the unemployment rate, increase industrial production or turnover, and promote consumer saving in the short run to increase stock market returns.

- ***Increase the domestic PPI: manufacturing:*** *The domestic PPI: manufacturing estimates the average variation over time in the price of goods from the perspective of domestic manufacturers.*

  Stock market returns decline when there is a decline in the domestic producer price index (PPI): manufacturing in South Africa. Policymakers must devise economic policies that increase the domestic PPI: manufacturing in the short run to increase stock market returns.

- ***Increase the value of wholesale trade:*** *Wholesale trade is the purchase and selling of large quantities of goods for resale to retailers.*

  Stock market returns decline when there is a decline in the value of wholesale trade in South Africa. Policymakers must incentive producers to increase the value of wholesale trade in the short run, which will increase stock market returns.

Research findings hold that policymakers must sustain economic stabilization and consider limited economic interventions, i.e., reduce the unemployment rate, increase industrial production or turnover, and promote consumer saving, increase the domestic PPI: manufacturing, and increase the value of wholesale trade, among other crucial economic interventions, in the short run, to enhance stock market returns in South Africa.

5.6 Learned Model Application

The extreme gradient boosting tree model trained on a feature subset containing key economic features selected based on the *gini impurity* value calculated by the decision tree model showed optimal model performance when forecasting stock market returns to changing economic activities in South Africa. The model performed better than the benchmark model and other machine learning models.

There are no hints of an improper or excessive model fit. The extreme gradient boosting tree model is expected to forecast the reaction of stock market returns to changing economic activities in other nations, and achieve sensible model performance when trained on a feature subset containing key economic features.

5.7 Future Research

This research expands knowledge about the usefulness of machine learning models in the research on the reaction of stock market returns to changing economic activities. Subsequent research should consider the points listed below to take advantage of research opportunities and make further advancements:

- This research considered only changing economic activities as predictor features. Forthcoming research should include company-specific indicators (i.e., profits, revenue, and the price-to-earnings ratio, among other company-specific indicators), along with technical indicators (i.e., moving averages, moving average convergence divergence, and relative strength index, among other indicators). This will enable researchers to better understand the dynamics of the economy and how stock market returns react to changing business and economic activities in South Africa.
- The *rectified linear unit* activation function is the main neural network activation function and the data standardization scaling strategy represents the primary data scaling strategy in this research, and the *batch size* and *epochs* selection criteria of networks were based on discretion. Subsequent research should consider other activation functions and data scaling strategies, including different combinations of *batch sizes* and *epochs*, .
- Another limitation of this study is that it does not address the matching of various time zones between South African economic and stock market index data, i.e., data from the United States of America and data from South Africa. Furthermore, the transmission method was fully uncovered–if economic data is published, it is already based on past occurrences. GDP production, for example, is a backward-looking variable. As a result, one must explain how backward-looking variables affect share returns today and in the future.

A

A List of Economic Features used Forecast the Reaction of Stock Market Returns in South Africa

Table A.1 shows a list of economic features used forecast the reaction of stock market returns in South Africa.

TABLE A.1: A list of economic features used forecast the reaction of stock market returns in South Africa

Economic Features	**Data Frequency**			**Date Range**	**Source**
	Monthly	**Quarterly**	**Annually**		
Business Tendency Surveys for Manufacturing: Confidence Indicators: Composite Indicators: OECD Indicator	*			2002 to 2022	FRED
Real Effective Exchange Rates Based on Manufacturing Consumer Price	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
National Currency to US Dollar Exchange Rate: Average of Daily Rates	*			2002 to 2022	FRED
National Currency to US Dollar Spot Exchange Rate	*			2002 to 2022	FRED
CPI: Total Food Excluding Restaurants	*			2002 to 2022	FRED
CPI: OECD Groups: Fuel, Electricity, and Gasoline	*			2002 to 2022	FRED
CPI: OECD Groups: All Items Non-Food and Non-Energy	*			2002 to 2022	FRED
Commercial Real Estate Loans to Total Loans	*			2002 to 2022	FRED
Consumer Opinion Surveys: Confidence Indicators: Composite Indicators: OECD Indicator	*			2002 to 2022	FRED
U.S. Exports of Goods by F.A.S. Basis to South Africa	*			2002 to 2022	FRED
U.S. Imports of Goods by Customs Basis from South Africa	*			2002 to 2022	FRED
Interest Rates, Government Securities, Government Bonds	*			2002 to 2022	FRED
Interest Rates, Government Securities, Treasury Bills	*			2002 to 2022	FRED
3-Month or 90-day Rates and Yields: Treasury Securities	*			2002 to 2022	FRED
Long-Term Government Bond Yields: Combined Terms	*			2002 to 2022	FRED
Long-Term Government Bond Yields: 10-year: Main (Including Benchmark)	*			2002 to 2022	FRED
Immediate Rates: Less than 24 Hours: Central Bank Rates	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
Immediate Rates: Less than 24 Hours: Call Money/Interbank Rate	*			2002 to 2022	FRED
Liquid Assets to Short Term Liabilities	*			2002 to 2022	FRED
Liquid Assets to Total Assets (Liquid Asset Ratio)	*			2002 to 2022	FRED
M1	*			2002 to 2022	FRED
M0	*			2002 to 2022	FRED
M2	*			2002 to 2022	FRED
M3	*			2002 to 2022	FRED
Broad Effective Exchange Rate	*			2002 to 2022	FRED
Producer Prices Index: Economic Activities: Domestic Mining and Quarrying Activities	*			2002 to 2022	FRED
Producer Prices Index: Economic Activities: Domestic Manufacturing	*			2002 to 2022	FRED
Real Broad : Effective Exchange Rate	*			2002 to 2022	FRED
Bank Regulatory Capital to Risk-Weighted Assets	*			2002 to 2022	FRED
Return on Assets	*			2002 to 2022	FRED
Return on Equity	*			2002 to 2022	FRED
Regulatory Tier 1 Capital to Risk-Weighted Assets	*			2002 to 2022	FRED
Total Reserves excluding Gold	*			2002 to 2022	FRED
Goods, Value of Exports	*			2002 to 2022	FRED
Goods, Value of Imports	*			2002 to 2022	FRED
Ratio of Exports to Imports	*			2002 to 2022	FRED
Exports: Value Goods	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
Imports: Value Goods	*			2002 to 2022	FRED
Net Trade: Value Goods	*			2002 to 2022	FRED
CPI: Food and non-Alcoholic beverages (COICOP 01): Total	*			2002 to 2022	FRED
CPI: Clothing and footwear (COICOP 03): Total	*			2002 to 2022	FRED
CPI: Housing, water, electricity, gas and other fuels (COICOP 04): Actual rentals for housing: Total	*			2002 to 2022	FRED
CPI: Alcoholic beverages, tobacco and narcotics (COICOP 02): Total	*			2002 to 2022	FRED
CPI: Housing, water, electricity, gas and other fuels (COICOP 04): Maintenance and repairs of the dwellings: Total	*			2002 to 2022	FRED
CPI: Housing, water, electricity, gas and other fuels (COICOP 04): Water supply and miscellaneous services relating to the dwelling: Total	*			2002 to 2022	FRED
CPI: Housing, water, electricity, gas and other fuels (COICOP 04): Electricity, gas and other fuels: Total	*			2002 to 2022	FRED
CPI: Furnishings, household equip. and routine household maintenance (COICOP 05): Total	*			2002 to 2022	FRED
CPI: Health (COICOP 06): Total	*			2002 to 2022	FRED
CPI: Transport (COICOP 07): Total	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
CPI: Transport (COICOP 07): Fuels and lubricants for personal transport equipment: Total	*			2002 to 2022	FRED
CPI: Communication (COICOP 08): Total	*			2002 to 2022	FRED
CPI: Recreation and culture (COICOP 09): Total	*			2002 to 2022	FRED
CPI: Education (COICOP 10): Total	*			2002 to 2022	FRED
CPI: Restaurants and hotels (COICOP 11): Total	*			2002 to 2022	FRED
CPI: Miscellaneous goods and services (COICOP 12): Total	*			2002 to 2022	FRED
CPI: All items: Total	*			2002 to 2022	FRED
CPI: All Items Excluding Food and Energy	*			2002 to 2022	FRED
CPI: Energy	*			2002 to 2022	FRED
Component series: BTS - Demand or orders inflow: Normalised	*			2002 to 2022	FRED
Component series: BTS - Business situation: Normalised	*			2002 to 2022	FRED
Component series: Construction: Normalised	*			2002 to 2022	FRED
Component series: Construction: Original series	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
Component series: Interest rate spread: Normalised	*			2002 to 2022	FRED
Component series: Interest rate spread: Original series	*			2002 to 2022	FRED
Component series: Car registration - sales: Normalised	*			2002 to 2022	FRED
Leading indicators: CLI: Amplitude adjusted	*			2002 to 2022	FRED
Leading indicators: CLI: Normalised	*			2002 to 2022	FRED
Leading indicators: CLI: Trend restored	*			2002 to 2022	FRED
Reference series: Gross Domestic Product (GDP): Normalised	*			2002 to 2022	FRED
Reference series: Gross Domestic Product (GDP): Ratio to trend	*			2002 to 2022	FRED
Reference series: Gross Domestic Product (GDP): Trend	*			2002 to 2022	FRED
Monetary aggregates and their components: Broad money and components: M3: M3	*			2002 to 2022	FRED
Monetary aggregates and their components: Narrow money and components: M1 and components: M1	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
Orders: Construction: Permits issued: Dwellings / Residential buildings	*			2002 to 2022	FRED
Permits Issued for Dwelling	*			2002 to 2022	FRED
Producer Prices Index: Economic activities: Mining and quarrying activities: Domestic	*			2002 to 2022	FRED
Producer Prices Index: Economic activities: Manufacturing: Domestic	*			2002 to 2022	FRED
Producer Prices Index: Economic activities: Industrial activities: Domestic	*			2002 to 2022	FRED
Domestic Producer Prices Index: Manufacturing	*			2002 to 2022	FRED
Production: Construction: Total construction: Total	*			2002 to 2022	FRED
SZAFPRENEL01MLM" : "Production: Energy: Electricity: Total	*			2002 to 2022	FRED
Production: Mining: Total mining: Total	*			2002 to 2022	FRED
Production: Manufacturing: Total manufacturing: Total manufacturing	*			2002 to 2022	FRED
Production of Total Construction	*			2002 to 2022	FRED
Production in Total Manufacturing	*			2002 to 2022	FRED
OECD based Recession Indicators from the Period following the Peak through the Trough	*			2002 to 2022	FRED
OECD based Recession Indicators from the Peak through the Trough	*			2002 to 2022	FRED

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
OECD based Recession Indicators from the Peak through the Period preceding the Trough	*			2002 to 2022	FRED
Passenger Car Registrations	*			2002 to 2022	FRED
Total Retail Trade	*			2002 to 2022	FRED
Sales: Manufacturing: Total manufacturing: Value	*			2002 to 2022	FRED
Sales: Retail trade: Car registration: Passenger cars	*			2002 to 2022	FRED
Sales: Retail trade: Total retail trade: Volume	*			2002 to 2022	FRED
Sales: Retail trade: Total retail trade: Value	*			2002 to 2022	FRED
Sales: Wholesale trade: Total wholesale trade: Volume	*			2002 to 2022	FRED
Sales: Wholesale trade: Total wholesale trade: Value	*			2002 to 2022	FRED
Deposits by residents	*			2002 to 2022	SARB
Deposits by non-residents	*			2002 to 2022	SARB
Total deposit liabilities	*			2002 to 2022	SARB
Loans received under repurchase agreements	*			2002 to 2022	SARB
Total liabilities to the public, Total equity	*			2002 to 2022	SARB
Cash reserve balances with the Reserve Bank	*			2002 to 2022	SARB
Treasury bills discounted	*			2002 to 2022	SARB
Land Bank bills and promissory notes discounted	*			2002 to 2022	SARB
Other bills discounted, including bankers acceptances	*			2002 to 2022	SARB
Advances to the domestic private sector	*			2002 to 2022	SARB
Total deposits, loans and advances	*			2002 to 2022	SARB
Short-term government stock	*			2002 to 2022	SARB
Long-term government stock	*			2002 to 2022	SARB

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
Total investments and bills discounted	*			2002 to 2022	SARB
Total assets	*			2002 to 2022	SARB
Required liquid assets	*			2002 to 2022	SARB
Actual liquid asset holdings	*			2002 to 2022	SARB
Credit card purchases processed during the period	*			2002 to 2022	SARB
Cash credit advances	*			2002 to 2022	SARB
Total loans and advances	*			2002 to 2022	SARB
Value of shares traded (Rm)	*			2002 to 2022	SARB
Number of shares traded (millions)	*			2002 to 2022	SARB
Net purchases of shares	*			2002 to 2022	SARB
Net purchases of bonds	*			2002 to 2022	SARB
Total net purchases of shares and bonds	*			2002 to 2022	SARB
Net issues of marketable debt securities (Rm)	*			2002 to 2022	SARB
Nominal value of bonds traded (Rm)	*			2002 to 2022	SARB
Consideration of bonds traded (Rm)	*			2002 to 2022	SARB
Number of transactions (actual number)	*			2002 to 2022	SARB
Yield on long-term government bonds (%)	*			2002 to 2022	SARB
Number of deals (actual number)	*			2002 to 2022	SARB
Number of contracts (actual number)	*			2002 to 2022	SARB
Underlying value (Rm)	*			2002 to 2022	SARB
Open interest (actual number)	*			2002 to 2022	SARB
Manufacturing (constant prices, 2019=100)	*			2002 to 2022	SARB
Wholesale trade (constant prices, 2019=100)	*			2002 to 2022	SARB
Retail trade (constant prices, 2019=100)	*			2002 to 2022	SARB

Continued on next page

Table A.1 – Continued from previous page

Economic Features	Data Frequency			Date Range	Source
	Monthly	Quarterly	Annually		
Manufacturing (2019=100)	*			2002 to 2022	SARB
Coincident indicator (2015=100)	*			2002 to 2022	SARB
Lagging indicator (2015=100)	*			2002 to 2022	SARB
Leading indicator (2015 = 100)	*			2002 to 2022	SARB
Cash-flow revenue (nsa)	*			2002 to 2022	SARB
Cash-flow expenditure (nsa)	*			2002 to 2022	SARB
Cash-flow balance (deficit -/surplus +)	*			2002 to 2022	SARB
Plus: Cost/profit on revaluation of foreign debt at redemption	*			2002 to 2022	SARB
Plus: Accrual adjustment	*			2002 to 2022	SARB
Net borrowing requirement (nsa)	*			2002 to 2022	SARB
Treasury bills and short-term loans	*			2002 to 2022	SARB
Domestic government bonds	*			2002 to 2022	SARB
Foreign bonds and loans	*			2002 to 2022	SARB
Other financing	*			2002 to 2022	SARB
Change in cash balances	*			2002 to 2022	SARB
Total net financing (nsa)	*			2002 to 2022	SARB

B

A Full Index that Offers Insight Into the Structure of the Economy in South Africa

Table B.1 shows a full index that offers insight into the structure of the economy in South Africa.

TABLE B.1: A full index that offers insight into the structure of the economy in South Africa

Rank	Economic Features	EV Ratio
1	Business Tendency Surveys for Manufacturing: CI: Composite Indicators: OECD Indicator	20.2738
2	Real Effective Exchange Rates Based on Manufacturing Consumer Price	14.0164
3	National Currency to US Dollar Exchange Rate: Average of Daily Rates	12.9776
4	National Currency to US Dollar Spot Exchange Rate	8.0563
5	CPI: Total Food Excluding Restaurants	6.3136
6	CPI: OECD Groups: Fuel, Electricity, and Gasoline	5.8097
7	CPI: OECD Groups: All Items Non-Food and Non-Energy	4.6073
8	Commercial Real Estate Loans to Total Loans	4.3957
9	COS: CI: Composite Indicators: OECD Indicator	3.8083
10	U.S. Exports of Goods by F.A.S. Basis to South Africa	3.1414
11	U.S. Imports of Goods by Customs Basis from South Africa	3.1017
12	Interest Rates, Government Securities, Government Bonds	2.7323
13	Interest Rates, Government Securities, Treasury Bills	2.6233
14	3-Month or 90-day Rates and Yields: Treasury Securities	2.3466
15	Long-Term Government Bond Yields: Combined Terms for South Africa	2.1716
16	Long-Term Government Bond Yields: 10-year: Main (Including Benchmark)	2.0238
17	Immediate Rates: Less than 24 Hours: Central Bank Rates	1.8959
18	Immediate Rates: Less than 24 Hours: Call Money/Interbank Rate	1.7775
19	Liquid Assets to Short Term Liabilities	1.7419
20	Liquid Assets to Total Assets (Liquid Asset Ratio)	1.5729
21	Broad Money, Index for South Africa (DISCONTINUED)	1.5164
22	M1	1.4719
23	Narrow Money, Index for South Africa (DISCONTINUED)	1.4472
24	M0	1.4313

Continued on next page

Table B.1 – Continued from previous page

Rank	Economic Features	EV Ratio
25	M2	1.3202
26	M3	1.264
27	Broad Effective Exchange Rate	1.2225
28	PPI: EC: Domestic Mining and Quarrying Activities	1.1954
29	PPI: EC: Domestic Manufacturing	1.1414
30	Real Broad: Effective Exchange Rate	1.0687
31	Bank Regulatory Capital to Risk-Weighted Assets	1.051
32	Return on Assets	1.0289
33	Return on Equity	0.974
34	Regulatory Tier 1 Capital to Risk-Weighted Assets	0.9466
35	Total Reserves excluding Gold	0.9389
36	Goods, Value of Exports	0.883
37	Goods, Value of Imports	0.8495
38	Ratio of Exports to Imports	0.8132
39	Exports: Value Goods	0.8036
40	Imports: Value Goods	0.7775
41	Net Trade: Value Goods	0.7706
42	CPI: Food and non-Alcoholic beverages: Total	0.7281
43	CPI: Clothing and footwear: Total	0.7069
44	CPI: Housing, water, electricity, gas and other fuels	0.672
45	CPI: Housing, water, electricity, gas and other fuels: Actual rentals for housing: Total	0.6632
46	CPI: Alcoholic beverages, tobacco and narcotics (COICOP 02): Total	0.6449
47	CPI: Housing, water, electricity, gas and other fuels: Maintenance and repairs of the dwellings	0.6209
48	CPI: Housing, water, electricity, gas and other fuels: Water supply and msc.: Total	0.6077
49	CPI: Housing, water, electricity, gas and other fuels: Electricity, gas and other fuels	0.597
50	CPI: Furnishings, household equip. and routine household maintenance: Total	0.5832
51	CPI: Health (COICOP 06): Total	0.5651
52	CPI: Transport: Total	0.5519
53	CPI: Transport: Fuels and lubricants for personal transport equipment: Total	0.5306
54	CPI: Communication: Total	0.5089
55	CPI: Recreation and culture: Total	0.5037
56	CPI: Education: Total	0.4754
57	CPI: Restaurants and hotels: Total	0.4671
58	CPI: Miscellaneous goods and services: Total	0.4629
59	CPI: All items: Total	0.4498
60	CPI: All Items Excluding Food and Energy	0.4439
61	CPI: Energy	0.4155
62	LL OECD: CS: BTS - Demand or orders inflow: Normalised	0.409
63	LL OECD: CS: BTS - Business situation: Normalised	0.3965
64	LL OECD: CS: Construction: Normalised	0.3908
65	LL OECD: CS: Construction: Original series	0.3784
66	LL OECD: CS: Interest rate spread: Normalised	0.3695

Continued on next page

Table B.1 – Continued from previous page

Rank	Economic Features	EV Ratio
67	LL OECD: CS: Interest rate spread: Original series	0.3355
68	LL OECD: CS: Car registration - sales: Normalised	0.3223
69	LL OECD: LL: CLI: Amplitude adjusted	0.3202
70	LL OECD: LL: CLI: Normalised	0.3156
71	LL OECD: LL: CLI: Trend restored	0.3065
72	LL OECD: Reference series: GDPNormalised	0.2995
73	LL OECD: Reference series: GDPRatio to trend	0.2888
74	LL OECD: Reference series: GDPTrend	0.2742
75	Broad money and components: M3	0.2699
76	Narrow money and components: M1 and components: M1	0.265
77	Orders: Construction: Permits issued: Dwellings / Residential buildings	0.2524
78	Permits Issued for Dwelling	0.2438
79	PPI: EC: Mining and quarrying activities: Domestic	0.2401
80	PPI: EC: Manufacturing: Domestic	0.2332
81	PPI: EC: Industrial activities: Domestic	0.227
82	Domestic PPI: Manufacturing	0.2203
83	Production: Construction: Total construction: Total	0.2123
84	Production: Energy: Electricity: Total	0.2057
85	Production: Mining: Total mining: Total	0.2019
86	Production: Manufacturing: Total manufacturing: Total manufacturing	0.1815
87	Production of Total Construction	0.1767
88	Production in Total Manufacturing	0.1732
89	OECD based Recession Indicators from the Period following the Peak through the Trough	0.164
90	OECD based Recession Indicators from the Peak through the Trough	0.1602
91	OECD based Recession Indicators from the Peak through the Period preceding the Trough	0.1557
92	Passenger Car Registrations	0.1522
93	Total Retail Trade	0.1465
94	Sales: Manufacturing: Total manufacturing: Value	0.1385
95	Sales: Retail trade: Car registration: Passenger cars	0.1309
96	Sales: Retail trade: Volume	0.1264
97	Sales: Retail trade: Value	0.1199
98	Sales: Wholesale trade: Volume	0.1148
99	Sales: Wholesale trade: Value	0.112
100	Deposits by residents	0.1084
101	Deposits by non-residents	0.1079
102	Total deposit liabilities	0.0999
103	Loans received under repurchase agreements	0.0981
104	Total liabilities to the public	0.0954
105	Total equity	0.0902
106	Cash reserve balances with the Reserve Bank	0.0884
107	Treasury bills discounted	0.0851
108	Land Bank bills and promissory notes discounted	0.0832
109	Other bills discounted, including bankers' acceptances	0.0773
110	Advances to the domestic private sector	0.074
111	Total deposits, loans and advances	0.0717

Continued on next page

Table B.1 – Continued from previous page

Rank	Economic Features	EV Ratio
112	Short-term government stock	0.0688
113	Long-term government stock	0.0626
114	Total investments and bills discounted	0.0606
115	Total assets	0.0577
116	Required liquid assets	0.0551
117	Actual liquid asset holdings	0.0536
118	Credit card purchases processed during the period	0.0513
119	Cash credit advances	0.0473
120	Total loans and advances	0.0454
121	Value of shares traded (Rm)	0.0415
122	Number of shares traded (millions)	0.0405
123	Net purchases of shares	0.0376
124	Net purchases of bonds	0.0359
125	Total net purchases of shares and bonds	0.0354
126	Net issues of marketable debt securities (Rm)	0.0313
127	Nominal value of bonds traded (Rm)	0.0302
128	Consideration of bonds traded (Rm)	0.0262
129	Number of transactions (actual number)	0.025
130	Yield on long-term government bonds (%)	0.023
131	Number of deals (actual number)	0.0222
132	Number of contracts (actual number)	0.0203
133	Underlying value (Rm)	0.0186
134	Open interest (actual number)	0.0183
135	Manufacturing (constant prices, 2019=100)	0.0158
136	Wholesale trade (constant prices, 2019=100)	0.0142
137	Retail trade (constant prices, 2019=100)	0.0135
138	Manufacturing (2019=100)	0.0112
139	Coincident indicator (2015=100)	0.0105
140	Lagging indicator (2015=100)	0.0078
141	Leading indicator (2015 = 100)	0.0063
142	Cash-flow revenue (nsa)	0.0051
143	Cash-flow expenditure (nsa)	0.0037
144	Cash-flow balance (deficit -/surplus +)	0.0027
145	Plus: Cost/profit on revaluation of foreign debt at redemption	0.0003
146	Plus: Accrual adjustment	0.0001
147	Net borrowing requirement (nsa)	0
148	Treasury bills and short-term loans	0
149	Domestic government bonds	0
150	Foreign bonds and loans	0
151	Other financing	0
152	Change in cash balances	0
153	Total net financing (nsa)	0

C Coefficients of the Optimal Model

Table C.1 shows coefficients of the optimal model (the ridge model).

TABLE C.1: Coefficients of the the optimal model (the ridge model)

Economic Features	Coefficients
Coincident indicator (2015=100)	0.230761
Domestic PPI: Manufacturing	0.143402
Sales: Wholesale trade: Value	0.133152
Number of shares traded (millions)	0.105246
Total liabilities to the public	0.0523371
LL OECD: LL: CLI: Normalised	0.0517331
Return on Assets	0.0489404
Long-term government stock	0.0447829
LL OECD: Reference series: GDPTrend	0.0283969
COS: CI: Composite Indicators: OECD Indicator	0.0270618
Total loans and advances	0.0135386
Required liquid assets	-0.00465482
Broad money and components: M3	-0.0974683
Real Effective Exchange Rates Based on Manufacturing Consumer Price	-0.124286

Bibliography

[Adejayan and Oke 2022] Adeola Oluwakemi Adejayan and Michael O. Oke. Sensitivity of macroeconomic variables to exchange rate shock in nigeria and south africa: A structural vector auto regression approach. *European Journal of Economic and Financial Research*, 2022.

[Ahangar *et al.* 2010] Reza Gharoie Ahangar, Mahmood Yahyazadehfar, and Hassan Pournaghshband. The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in tehran stock exchange. *ArXiv*, abs/1003.1457, 2010.

[Alhomadi 2021] Abraham Alhomadi. Forecasting stock market prices: A machine learning approach. 2021.

[Andrei *et al.* 2018] Daniel Andrei, Julien Cujean, and Mungo Ivor Wilson. The lost capital asset pricing model. *S&P Global Market Intelligence Research Paper Series*, 2018.

[Anscombe and Guttman 1960] F. J. Anscombe and Irwin Guttman. Rejection of outliers. *Technometrics*, 2(2):123–147, 1960.

[Ataman and Kahraman 2021] Görkem Ataman and Serpil Kahraman. Stock market prediction in brics countries using linear regression and artificial neural network hybrid models. *The Singapore Economic Review*, 2021.

[Aye *et al.* 2020] Goodness C. Aye, Mehmet Balcilar, and Rangan Gupta. The effectiveness of monetary policy in south africa under inflation targeting: Evidence from a time-varying factor-augmented vector autoregressive model. *The Journal of Developing Areas*, 54:–, 2020.

[Babu and Reddy 2014] C. Narendra Babu and B. Eswara Reddy. A moving-average filter based hybrid arima-ann model for forecasting time series data. *Appl. Soft Comput.*, 23:27–38, 2014.

[Balusik *et al.* 2021] Adam Balusik, Jared de Magalhaes, and Rendani Mbuvha. Forecasting the jse top 40 using long short-term memory networks. *ArXiv*, abs/2104.09855, 2021.

[BenSaïda and Litimi 2013] Ahmed BenSaïda and Houda Litimi. High level chaos in the exchange and index markets. *Chaos, Solitons Fractals*, 54:90–95, 2013.

[Bordo and Jeanne 2002] Michael David Bordo and Olivier D Jeanne. Monetary policy and asset prices: Does 'benign neglect' make sense? *Banking & Financial Institutions*, 2002.

[Burdenko 2017] Iryna Burdenko. Fair value: history, assessment and financial crisis. 2017.

[Carter *et al.* 2017] Bradley Carter, Chris Muller, and Mike Ward. The applicability of black's variation of the capital asset pricing model (capm) in the south african context. *ERN: Asset Pricing Models (Topic)*, 2017.

[Chifurira and Chinhamu 2019] R. Chifurira and K. Chinhamu. Evaluating south africa's market risk using asymmetric power auto-regressive conditional heteroscedastic model under heavy-tailed distributions. *Journal of Economic and Financial Sciences*, 12:11, 2019.

[Chitenderu *et al.* 2014] Tafadzwa Chitenderu, Andrew Maredza, and Kin Sibanda. The random walk theory and stock prices: Evidence from johannesburg stock exchange. 2014.

[Chung *et al.* 2014] J. Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014.

[Crack and Grieves 2017] Timothy Crack and Robert B. Grieves. The practical implications of modern portfolio theory. *ERN: Asset Pricing Models (Topic)*, 2017.

[da Costa and Gebbie 2020] Joel da Costa and Tim Gebbie. Learning low-frequency temporal patterns for quantitative trading. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1091–1099, 2020.

[Dietterich 1996] Thomas G. Dietterich. Machine learning. *ACM Comput. Surv.*, 28:3, 1996.

[Elton and Gruber 2018] Edwin J. Elton and Martin J. Gruber. Modern portfolio theory and investment analysis / edwin j. elton, martin j. gruber. 2018.

[Fama and French 1993] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56, 1993.

[Fama 1970] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[Fu *et al.* 2018] XingYu Fu, JinHong Du, YiFeng Guo, MingWen Liu, Tao Dong, and XiuWen Duan. A machine learning framework for stock selection. *ArXiv*, abs/1806.01743, 2018.

[Gouriéroux *et al.* 2017] C. Gouriéroux, A. Monfort, and Jean-Paul Renne. Statistical inference for independent component analysis: Application to structural var models. *Journal of Econometrics*, 196:111–126, 2017.

[Gruber 1998] Marvin H. J. Gruber. Improving efficiency by shrinkage: The james–stein and ridge regression estimators. 1998.

[Hastie *et al.* 2009] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Boosting and Additive Trees*, pages 1–51. 04 2009.

[Hedging and Umoetok 2013] Futures Hedging and Ekerete Umoetok. Evidence from south africa. 2013.

[Hochreiter and Schmidhuber 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[Hou *et al.* 2017] Zehan Hou, Zeyu Li, and Yang Zhou. Review on the modern portfolio theory and optimization model. 2017.

[Huberman 2005] Gur Huberman. Arbitrage pricing theory. *Staff Reports*, pages 72–80, 2005.

[Hyndman and Athanasopoulos 2013] R. Hyndman and G. Athanasopoulos. Forecasting: Principles and practice. 2013.

[Ilesanmi and Tewari 2020] Kehinde Damilola Ilesanmi and Devi Datt. Tewari. Financial stress index and economic activity in south africa: New evidence. *Economies*, 8:110, 2020.

[Jasra *et al.* 2012] Javed Mahmood Jasra, Rauf I Azam, and Akram Ahmad Muhammad. Impact of macroeconomic variables on stock prices: Industry level analysis. *Actual Problems of Economics*, pages 403–412, 2012.

[JSELtd 2021] JSELtd. *INTEGRATED ANNUAL REPORT*. INTEGRATED ANNUAL REPORT for the year ended December 2021, Dec. 31, 2021.

[Kamalov 2020] Firuz Kamalov. Forecasting significant stock price changes using neural networks. *Neural Computing and Applications*, pages 1–13, 2020.

[Kennedy *et al.* 2020] Edward H. Kennedy, Sivaraman Balakrishnan, and Larry A. Wasserman. Discussion of "on nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning". *Statistical Science*, 35:540–544, 2020.

[Khudoykulov 2017] Khurshid Khudoykulov. The analysis of the arbitrage pricing model on the stock return: a case of athens stock market. *American J. of Finance and Accounting*, 5:51–63, 2017.

[Klibanov *et al.* 2021] Mikhail V. Klibanov, Kirill V. Golubnichiy, and Andrey V. Nikitin. Application of neural network machine learning to solution of black-scholes equation. 2021.

[Leke and Marwala 2019] Collins Leke and T. Marwala. Deep learning and missing data in engineering systems. *Studies in Big Data*, 2019.

[Lintner 1965] John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47:13–37, 1965.

[Liu *et al.* 2020] Lin Liu, Rajarshi Mukherjee, and James M. Robins. Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 2020.

[Liu *et al.* 2021] Yuanji Liu, Boyang Zhang, and Fengchen Xu. Research on forecasting volatility rate of financial market based on garch and neural networks. *Journal of Physics: Conference Series*, 1802, 2021.

[Liu 1996] Huan Liu. Book review: Machine learning, neural and statistical classification edited by d. michie, d.j. spiegelhalter and c.c. taylor (ellis horwood limited, 1994). *Intelligencesigart Bulletin*, 7:16–17, 1996.

[Macfarlane 2011] Andrew Macfarlane. Do macroeconomic variables explain future stock market movements in south africa. 2011.

[Makatjane and Moroke 2021] Katleho Makatjane and Ntebogang D Moroke. Predicting extreme daily regime shifts in financial time series exchange/johannesburg stock exchange—all share index. *International Journal of Financial Studies*, 9:18, 2021.

[Malamud 2015] Semyon Malamud. A dynamic equilibrium model of etfs. *ERN: Asset Pricing Models (Topic)*, 2015.

[Mallikarjuna and Rao 2019] M. Mallikarjuna and R. Prabhakar Rao. Evaluation of forecasting methods from selected stock market returns. *Financial Innovation*, 5, 2019.

[Marozva 2020] Godfrey Marozva. The effects of monetary policy on stock market returns and volatility: Evidence from south africa. *Academy of Accounting and Financial Studies Journal*, 24, 2020.

[Mason *et al.* 1999] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. pages 512–518, 01 1999.

[Mbuvha *et al.* 2021] Rendani Mbuvha, P. T. Zondo, Aluwani Mauda, and T. Marwala. Predicting higher education throughput in south africa using a tree-based ensemble technique. *ArXiv*, abs/2106.06805, 2021.

[Meier *et al.* 2008] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70:53–71, 02 2008.

[Mitchell 1997] Tom Michael Mitchell. Machine learning, international edition. In *McGraw-Hill Series in Computer Science*, 1997.

[Moore II 1987] Dan H. Moore II. Classification and regression trees, by leo breiman, jerome h. friedman, richard a. olshen, and charles j. stone. brooks/cole publishing, monterey, 1984,358 pages, $27.95. *Cytometry*, 8(5):534–535, 1987.

[Mossin 1966] Jan Mossin. Equilibrium in a capital asset market. *Econometrica*, 34:768, 1966.

[Mpofu 2011] Raphael Tabani Mpofu. An analysis of the south african ftse/jse all-share index and the macroeconomic variables during the period 2002 to 2010. *Corporate Ownership and Control*, 8, 2011.

[Mukoyi and Ogujiuba 2022] Lenia Mukoyi and Kanayo K. Ogujiuba. Comparison of multifactor asset pricing models in the south african stock market [2000–2016]. *Journal of Risk and Financial Management*, 2022.

[Mulaudzi and Ajoodha 2020] Rudzani Mulaudzi and Ritesh Ajoodha. Application of deep learning to forecast the south african unemployment rate: A multivariate approach. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6, 2020.

[Munk 2013] Claus Munk. Financial asset pricing theory. 2013.

[Muthama *et al.* 2014] Anthony K. Muthama, Macharia Munene, and Ondabu Ibrahim Tirimba. Empirical tests of capital asset pricing model and its testability for validity verses invalidity. *International Journal of Innovation and Applied Studies*, 9:600–614, 2014.

[Muzindutsi and Niyimbanira 2012] Paul-Francois Muzindutsi and Ferdinand Niyimbanira. The exchange rate risk in the johannesburg stock market: An application of the arbitrage pricing model. *Journal of Global Business and Technology*, 8:60, 2012.

[Myttenaere *et al.* 2016] Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016.

[Ndikum 2020] Philip Ndikum. Machine learning algorithms for financial asset price forecasting. *ArXiv*, abs/2004.01504, 2020.

[Nengovhela 2022] Murendeni Nengovhela. *Forecasting stock market returns using machine learning models*. PhD thesis, University of Johannesburg, 2022.

[Nokeri 2021] Tshepo Chris Nokeri. Tree modeling and gradient boosting with scikit-learn, xgboost, pyspark, and h2o. *Data Science Solutions with Python*, 2021.

[Olayode *et al.* 2021] Fatoki Olayode, Adeleye Najeem Friday, and Afolabi Nosimot Omowunmi. Application of error correction model in assessing the impact of macroeconomic variables on stock market performance in nigeria. *International Journal of Data Science and Analysis*, 2021.

[Otuteye and Siddiquee 2017] Eben Otuteye and Mohammad Siddiquee. A critique of modern portfolio theory and asset pricing models based on behavioral insights from benjamin graham's value investing paradigm. *FEN: Experimental Finance (Topic)*, 2017.

[Pahlawan *et al.* 2021a] M. Reza Pahlawan, Edwin Riksakomara, Raras Tyasnurita, Ahmad Muklason, Faizal Mahananto, and Retno A. Vinarti. Stock price forecast of macroeconomic factor using recurrent neural network. *IAES International Journal of Artificial Intelligence*, 10:74–83, 2021.

[Pahlawan *et al.* 2021b] Muhammad Pahlawan, Edwin Riksakomara, Raras Tyasnurita, Ahmad Muklason, Faizal Mahananto, and Retno Aulia Vinarti. Stock price forecast of macro-economic factor using recurrent neural network. *IAES International Journal of Artificial Intelligence*, 10:74–83, 2021.

[Pillay 2020] Suren Pillay. Determining the optimal arima model for forecasting the share price index of the johannesburg stock exchange. *Journal of Management Information and Decision Sciences*, 23, 2020.

[Pincus 1995] R. Pincus. Barnett, v., and lewis t.: Outliers in statistical data. 3rd edition. j. wiley & sons 1994, xvii. 582 pp., £49.95. *Biometrical Journal*, 37(2):256–256, 1995.

[Pratiwi *et al.* 2021] Siska Gita Pratiwi, Robiyanto Robiyanto, and S Martono. The influence of the fundamental and macroeconomic factors on manufacturing companies' stock returns. 2021.

[Ramos-Pérez *et al.* 2021] Eduardo Ramos-Pérez, Pablo J. Alonso-González, and José Javier Núñez-Velázquez. Multi-transformer: A new neural network-based architecture for forecasting s&p volatility. *ArXiv*, abs/2109.12621, 2021.

[Reddy and Thomson 2011] T. L. Reddy and Robert John Thomson. The capital-asset pricing model : the case of south africa : refereed paper. *South African Actuarial Journal*, 11:43–84, 2011.

[Rubin 1975] Donald B. Rubin. Inference and missing data. *Psychometrika*, 1975:19, 1975.

[Sako *et al.* 2022] Kady Sako, Berthine Nyunga Mpinda, and Paulo Canas Rodrigues. Neural networks for financial time series forecasting. *Entropy*, 24, 2022.

[Saxena and Bhadauriya 2021] Swami Prasad Saxena and Sonam Bhadauriya. Causal interactions between macroeconomic variables and stock market returns in india with special reference to nse. *Economics of Networks eJournal*, 2021.

[Shamsudin *et al.* 2021] Norazidah Shamsudin, Mohd Izuan Rosmi, and Suhana Mohamed. Macroeconomic variables influence on stock market performance. *International Journal of Academic Research in Business and Social Sciences*, 2021.

[Shen 2009] Heng Tao Shen. Principal component analysis. In *Encyclopedia of Database Systems*, 2009.

[Sinha 2016] Paritosh Chandra Sinha. Systematic risk and liquidity risk in the nse market: A critique on r2. *The IUP Journal of Applied Finance*, 22:5, 2016.

[SSEInitiative 2021] SSEInitiative. *Johannesburg Stock Exchange (JSE)*. Johannesburg Stock Exchange (JSE), Dec. 31, 2021.

[Stadtmann and Dunsch 2018] Georg Stadtmann and Sophie Dunsch. The taylor rule in the year 2015: Still a benchmark for monetary policy? 2018.

[Swamidass 2000] Paul Swamidass. Encyclopedia of production and manufacturing management. 2000.

[Tabak 2004] John Tabak. Probability and statistics : the science of uncertainty. 2004.

[Taljaard and Maré 2021] B. H. Taljaard and Eben Maré. If the equal weighted portfolio is so great, why isn't it working in south africa? *Investment Analysts Journal*, 50:32 – 49, 2021.

[Tank *et al.* 2021] Alex Tank, Ian Covert, Nicholas J. Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021.

[Thomson and Reddy 2013] Robert John Thomson and T. L. Reddy. The capital-asset pricing model reconsidered : tests in real terms on a south african market portfolio comprising equities and bonds. *South African Actuarial Journal*, 13:221–263, 2013.

[Thulare *et al.* 2021] Evans Thulare, Ritesh Ajoodha, and Ashwini Jadhav. An empirical analysis and application of the expectation-maximization and matrix completion algorithms for varying degrees of missing data. *2021 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, pages 1–7, 2021.

[Tibshirani 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[Tsepang Patrick 2013] Mzamane Tsepang Patrick. *GARCH Modelling of Volatility in the Johannesburg Stock Exchange Index*. Master's thesis, 2013. Mzamane10232.

[Vijayakumar and Cheung 2018] R. Vijayakumar and M. Cheung. Replicability of machine learning models in the social sciences: A case study in variable selection. *Zeitschrift für Psychologie*, 226:259–273, 2018.

[Wong *et al.* 2020] Steven Y. K. Wong, Jennifer So-Kuen Chan, Lamiae Azizi, and Richard Yi Da Xu. Time-varying neural network for stock return prediction. *arXiv: Computational Finance*, 2020.

[Xiong *et al.* 2015] Ruoxuan Xiong, Eric P. Nichols, and Yuan Shen. Deep learning stock volatility with google domestic trends. *arXiv: Computational Finance*, 2015.

[Young and Shellswell 1972] P. Young and S. Shellswell. Time series analysis, forecasting and control. *IEEE Transactions on Automatic Control*, 17:281–283, 1972.

[Young 2020] G. Alastair Young. High-dimensional statistics: A non-asymptotic viewpoint, martin j.wainwright, cambridge university press, 2019, xvii 552 pages, £57.99, hardback isbn: 978-1-1084-9802-9. *International Statistical Review*, 88:258–261, 2020.

[Zhang 2012] S. Zhang. Nearest neighbor selection for iteratively knn imputation. *J. Syst. Softw.*, 85:2541–2552, 2012.

[Zimek and Filzmoser 2018] Arthur Zimek and Peter Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, 2018.

[Zimmerman 1994] Donald W. Zimmerman. A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology*, 121:391–401, 1994.

[Zou and Hastie 2005a] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[Zou and Hastie 2005b] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.