

https://dspace.org/

A machine learning approach to quantifying and relating the determinants of unemployment in South Africa

Mulaudzi, Rudzani

2021

https://hdl.handle.net/10539/32245

Mulaudzi, Rudzani (2021) A machine learning approach to quantifying and relating the determinants of unemployment in South Africa, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/32245>

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

MSc - Computer Science Research Dissertation

A Machine Learning Approach to Quantifying and Relating the Determinants of Unemployment in South Africa

Author:
Rudzani Mulaudzi
(0601737R)

Supervisor:
Dr Ritesh Ajoodha

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

April 23, 2021



A dissertation submitted in fulfilment of the requirements for the degree of MSc in Computer Science offered by the School of Computer Science and Applied Mathematics

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Masters of Computer Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



Rudzani Mulaudzi

22 April 2021

Date

Abstract

Unemployment is a significant problem that South Africa faces. The rate was 30.1% in the first quarter of 2020: placing it amongst the top ten worst unemployment rates in the world. Public policy is a typical instrument used by governments to address unemployment sustainably. It is normally informed by forecasts derived through economic (traditional statistical) models. These models are, however, suitable when the data is stationary and linear. The South African unemployment rate, on the other hand, is asymmetric, seasonal, upward trending, and nonstationary.

Vector autoregression (VAR), a traditional statistical model, was used to forecast the South African unemployment rate. It resulted in a mean absolute scaled error (MASE) of 41. Comparatively, twelve machine learning models were used to forecast the unemployment rate. The lowest MASE achieved was 0.39. Making the machine learning models 105 times more accurate than the VAR: the benchmark model. Additionally, through feature selection techniques, machine learning approaches enabled the identification of the most impactful features in forecasting the unemployment rate.

These features were used to construct a Dynamic Bayesian Network (DBN) to determine how they influence each other and the unemployment rate. The DBN was then used to perform do-Calculus, a data-driven scenario analysis technique. One scenario tested the impact of increasing the GDP on the unemployment rate. This positively impacts the unemployment rate. However, a decline in GDP has a greater negative impact. Therefore, policymakers should avoid, at all costs, a decline in the GDP.

This research, therefore, demonstrates the value of machine learning in forecasting the South African unemployment rate (a nonstationary macroeconomic variable) across the broad machine learning value chain: feature selection, forecasting, feature influence analysis, and do-Calculus scenario analysis. Previous research tends to only focus on one or two aspects of the value chain.

Publications

Some of the work contained in this dissertation appears in peer-reviewed IEEE, Springer, and Taylor & Francis conference publication:

- Mulaudzi and Ajoodha [2020b], "An exploration of machine learning models to forecast the unemployment rate of South Africa: a univariate approach," The International Multidisciplinary Information Technology and Engineering Conference, ISBN: 978-1-7281-9519-9, 2020.

- Mulaudzi and Ajoodha [2020a], "Application of Deep Learning to Forecast the South African Unemployment Rate: A Multivariate Approach," Seventh IEEE Computer Science and Data Engineering Conference, 2020.

- Mulaudzi and Ajoodha [2021a], "Demonstration that the Use of Feature Selection on High Dimensional South African Macroeconomic Data Results in Improved Performance with Lower Compute Requirements," Sixth International Congress on Information and Communication Technology Conference, ISBN: 2194-5357, 2021.

- Mulaudzi and Ajoodha [2021b], "Improving the Performance of Multivariate Forecasting Models Through Feature Engineering: A South African Unemployment Rate Forecasting Case Study," 2021 International Conference On Interdisciplinary Research in Technology and Management, 2021.

Dedication

To **Mushoni Selina Malatji** (*29 May 1971 to 13 January 2021*). Thank you for buying me my first laptop, which enabled me to transition to the university environment successfully.

"And I heard a voice from heaven saying, "Write this down: Blessed are those who die in the Lord from now on. Yes, says the Spirit, they are blessed indeed, for they will rest from their hard work; for their good deeds follow them!"
– *Revelation 14:13*

Contents

1 Introduction **9**
- 1.1 Problem Statement . . . 10
- 1.2 Purpose Statement . . . 12
- 1.3 Research Questions . . . 13
- 1.4 Research Contributions . . . 13
- 1.5 Research Methodology Overview . . . 14
- 1.6 Dissertation Structure . . . 14

2 Background & Related Work **15**
- 2.1 Traditional Unemployment Forecasting . . . 15
  - 2.1.1 Univariate Linear Models . . . 16
  - 2.1.2 Multivariate Models . . . 19
  - 2.1.3 Known Challenges with Traditional Forecasting Models . . . 20
- 2.2 Machine Learning Unemployment Forecasting . . . 21
  - 2.2.1 Regression . . . 21
  - 2.2.2 Kernel Methods . . . 24
  - 2.2.3 Neural Networks . . . 24
  - 2.2.4 Nearest Neighbour . . . 27
  - 2.2.5 Ensemble Learning . . . 28
  - 2.2.6 Model Comparison . . . 28
- 2.3 Significant Findings in the Related Work . . . 40

3 Methodology **42**
- 3.1 Research Strategy . . . 42
- 3.2 Data Sources . . . 45
- 3.3 Data Preprocessing . . . 46
  - 3.3.1 Data Cleaning . . . 46

3.3.2 Missing Values . . . 47
3.3.3 Data Transformations . . . 49
3.3.4 Feature Selection . . . 49
3.3.5 Statistical Description of the South African Unemployment Rate . . . 53
3.3.6 Performance Evaluation . . . 55
3.4 Ethical Considerations . . . 56
3.5 Summary . . . 56

4 Experimental Results & Discussions **58**
4.1 Vector Autoregression: the Benchmark Model . . . 58
4.2 Machine Learning Models . . . 62
4.2.1 Regression . . . 62
4.2.2 Neural Network . . . 65
4.2.3 Decision Trees . . . 69
4.2.4 Kernel . . . 70
4.3 Overall Performance Summary . . . 71
4.4 Feature Engineering . . . 74
4.5 Machine Learning vs Vector Autoregression . . . 77
4.6 The Causal Relationship of the Features . . . 79
4.7 Research Contributions . . . 86

5 Conclusion & Future Work **88**
5.1 Conclusion . . . 88
5.2 Policy Recommendations . . . 90
5.3 Future Work . . . 91

Bibliography **91**

Appendices **98**

A Feature List **99**

B Hyperparameter Settings **111**

C Regressions Models Results **113**

D Neural Networks Models Results **117**

E Decision Trees Models Results **120**

List of Figures

1.1 The South African unemployment rate has been trending upwards from 1998 to 2019. . . . 10

2.1 The South African unemployment rate is asymmetric. Between 2003 and 2006 unemployment decreased as GDP increased. However, between 1998 and 2001 unemployment increased as GDP increased. . . . 18

3.1 A visual representation of the missing data from the feature set. The black highlight represents observations without missing values whilst unfilled areas are those with missing values. . . . 48
3.2 Properties of the unemployment rate time series from 1970 to 2020. . . . 54
3.3 The probability distribution of the South African unemployment rate from 1970 to 2019. . . . 54

4.1 Features that were selected by the Elastic Net (ENET) embedded feature selection technique with their relative coefficient estimate. . . . 64
4.2 The loss (error) of the deep learning models through various epochs. The error is downward trending for both training and test data. The test is generally below the training error in most cases. . . . 66
4.3 Features that were selected by the Random Forest (RFR) embedded feature selection technique with their relative coefficient estimate. . . . 70
4.4 The performance ranking of the top machine learning models, as well as, the ranking of feature selection techniques. . . . 73
4.5 The Correlation Map for the lag values of the unemployment rate showing that the unemployment rate is highly correlated with the lag values. . . . 75
4.6 The comparison of the VAR model with a few machine learning models to show the difference in performance. . . . 78
4.7 A Bayesian network showing how key features that impact the South African unemployment rate influence each other. . . . 80

4.8 A Dynamic Bayesian network (DBN) showing how key features that impact the South African unemployment rate influence each other over time: with only a single lag value for each node. 81

List of Tables

2.1 A detailed comparison of machine learning approaches to forecasting unemployment. 30

3.1 Properties of features and targets for modelling the South African unemployment rate: The VAR has limited capacity to cater for the varying data properties whilst the machine learning models have varying areas of strength. 43
3.2 Publicly accessible data sources that were used for the research 45
3.3 A list of features used for this research. A combination of univariate and multivariate techniques were used. 51

4.1 P-values for the Granger-causality for 15 variables. A p-value less than 0.05 implies that the blue features influence the green features. 59
4.2 The performance (MASE) of the VAR model with different feature sets over the test period of 24 months. 61
4.3 The performance (MASE) of five regression machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 63
4.4 Five features that were selected by the recursive feature elimination (RFE) wrapper feature selection technique. 67
4.5 The performance (MASE) of four neural networks over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 68
4.6 The performance (MASE) of two decision trees, extreme gradient boosting and random forest, models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 69
4.7 The performance (MASE) of the support vector regression model over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 71
4.8 The top ten performing (MASE) machine learning models across all different model categories. The performance was assessed using an average MASE over the four horizons). 72

4.9 The top ten models as measured by R-squared. 74
4.10 The inclusion of lag values of the South African unemployment rate improves the R-squared by over 100% on average. 75
4.11 The comparison of how the inclusion of lag values changes the performance (MASE) of the top models. 76
4.12 The percentage distribution of the various states that the different nodes can have. 82
4.13 Summary of the classification metrics used to evaluate the DBN. 82
4.14 The conditional probabilities of the South African unemployment rate given that the other nodes have been observed. 83
4.15 Counter factuals based on the Dynamic Bayesian model, showing how the unemployment rate would change if its features are changed. 85

A.1 List of features that sourced from the South African Reserve Bank 100

B.1 The hyperparameter settings for the twelve machine learning models used for this research. Only key parameters are shown. 111

C.1 The performance (MASE) of all regression machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 113

D.1 The performance (MASE) of all neural network machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 117

E.1 The performance (MASE) of all neural network machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). 120

1 Introduction

The South African unemployment rate was 30.1% in the first quarter of 2020, this was the highest it had ever been since 1970 [Maluleke 2020]. This is viewed by economists and policymakers as one of the most significant problems in South Africa [Jones 2019; Belling 2020]. According to Aiken [1996]; Pelaez [2006], many see the unemployment rate as a proxy measure for the health of a country. By this metric, South Africa is notably unhealthy. The country is in the top ten countries with the highest unemployment rates across the world [TradingEconomics 2019; Meyer 2014].

Statistics South Africa is responsible for monitoring and reporting the unemployment rate. Their definition of unemployment is adopted for this research as it is the authoritative one in the country. Their definition being,

> *Unemployment refers to a period or state where a person is not currently in a paid job opportunity and has taken active steps, although unsuccessful, to look for a job or to start a business and is available to take up a job should it be offered [Statistics South Africa 2019].*

The consequences of this definition is that people without jobs but have not searched for a job are not classified as unemployed. Hence, over the COVID-19 lock-down period the unemployment of South Africa declined to 23%: it was impossible for people to look for jobs except through digital channels [Bloomberg 2020; Maluleke 2020]. Effectively by the end of the second quarter of 2020, the unemployment rate declined to 23%, but those without jobs increased by over 2 million [Bloomberg 2020; Maluleke 2020]. Bloomberg [2020] estimates that the unemployment rate will be 34% by the end of 2021. Unemployment remains a significant problem in South Africa.

The remainder of this chapter will discuss the problem and purpose statements for this research, as well as the research questions that this research intends to address.

1.1 Problem Statement

The unemployment rate increased by 1.1% to 30.1% in the first quarter of 2020 [Maluleke 2020]. Trending along the same trajectory, youth unemployment has also increased to 59.0% for 15 to 34-year-olds. The inability of South Africa to resolve this social challenge is a humanitarian crisis and is a significant problem in the country [Jones 2019; Belling 2020].

Figure 1.1 provides a visual depiction of the unemployment rate from 1998 to 2019. The figure shows that the unemployment rate has been persistently high since the dawn of democracy, with the rate hovering between 20 and 30% from 1998 to 2019.

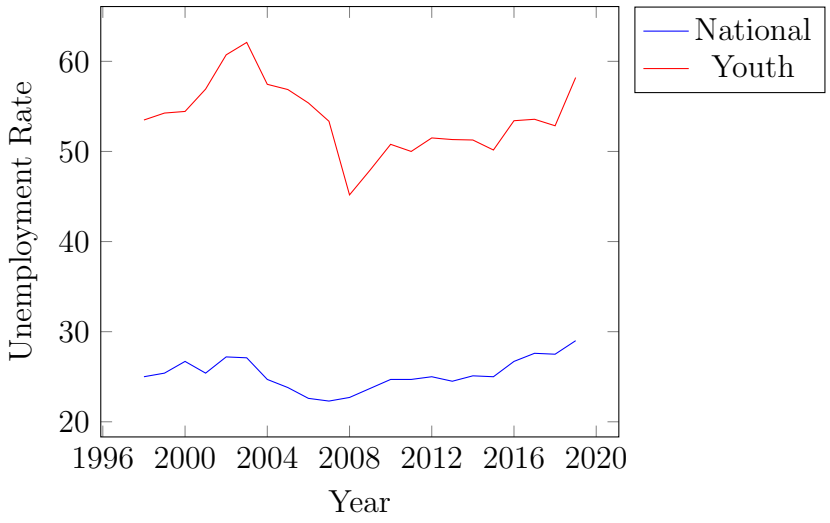


Figure 1.1: The South African unemployment rate has been trending upwards from 1998 to 2019.

There has been a wealth of research to understand and solve the South African unemployment problem [Fourie 2011; Meyer 2014; De Lannoy *et al.* 2018]. However, the problem still persists. Brynard [2011] states that a key reason for the persistence of the high unemployment is that policy making in the country is often too broad and not focused enough. Adding that, although there is a wealth of data relating to unemployment in South Africa, a limited amount is used in the policy making process.

Solutions to unemployment can either be programmatic or systemic. The programmatic being targeted at particular groups of unemployed persons. These are typically executed by companies, government municipalities, and non-profit organisations. Shankar *et al.* [2016] studied programmatic interventions and some of the programmes studied are the IT management company, EOH's youth job creation initiative and Mentec Foundation's train and place initiative. These programmes were found to be effective for unemployed youths placing over 100 000 of them into jobs.

Systemic interventions are those that aim to put in place measures that sustainably reduce unemployment across the board with the aim being to resolve the problem permanently [Levinsohn 2007; Brynard 2011]. Examples of this would be changes in the education system to make it aligned to employment opportunities or changes in government policies to prioritise job creation [De Lannoy *et al.* 2018]. One example of a policy intended to reduce unemployment is the wage subsidy policy in South Africa [Levinsohn 2007]. This policy enables more first-time job seekers to get employed because salaries for these employees are subsidised by the government.

Even though programmatic interventions are beneficial because they get a number of people into jobs, these have limited impact when compared to systemic interventions [De Lannoy *et al.* 2018]. Therefore, where possible, systemic interventions are preferred over programmatic ones. In order to effect systemic changes, policymakers are reliant on forecasts to determine which policy options to put in place [Makridakis 1988; Levinsohn 2007]. Currently, these forecasts are generated through traditional statistical methods [Makridakis *et al.* 2018; Brooks 2014; Hyndman and Athanasopoulos 2018].

Brynard [2011] states that in developing unemployment policy the labour market and its dynamics must be understood. In South Africa, this understanding comes from Statistics South Africa's quarterly labour force survey, which is a database that provides data on how unemployment changes yearly and quarterly. It also provides the demographic profile of the unemployed persons: their ages, location, race, gender, and education levels [Levinsohn 2007; Brynard 2011].

Once the labour market dynamics are understood previous policy interventions are studied for their successes or failures [Brynard 2011]. Thereafter, scenarios are run for new policy options with judgments from experts estimating possible reductions of the unemployment rate for each policy options being considered [Brynard 2011; Levinsohn 2007]. As such accurate forecasting is important throughout the policy making process [Brynard 2011].

Traditional statistical methods are historically the key methods for forecasting unemployment rates [Aiken 1996; Olmedo 2014; Hall 2018; Katris 2019ab]. However, these models have limitations as they primarily forecast data that is stationary, linear, and symmetric [Olmedo 2014; Aiken 1996; Brooks 2014; Hyndman and Athanasopoulos 2018; De Gooijer and Hyndman 2006]. Being only able to cater for nonlinear data with significant manipulation from the data analyst [Cook and Hall 2017; Hyndman and Athanasopoulos 2018; De Gooijer and Hyndman 2006].

These models are also biased towards the most recently seen data points. This is because the models are autoregressive which means that recent observations have a relatively larger weighing on the outcomes of the model [Makridakis 1988]. This is a challenge for nonstationary data as the data varies inconsistently over time. Hence, the models are not able to

accommodate economic shocks which cause sudden and unusual changes in the movement of unemployment rates such as economic recessions and booms [Aiken 1996; Olmedo 2014; Makridakis 1988; Makridakis *et al.* 2009].

De Gooijer and Hyndman [2006]; Clements *et al.* [2004] state that nonlinear modelling using traditional statistical approaches is still in its 'infancy'. They found no evidence in performance improvement – measured by error rate – when traditional statistical nonlinear models were compared to linear models on financial and macroeconomic data. This makes these traditional statistical methods limited when it comes to forecasting unemployment rates because unemployment data tends to be nonstationary, nonlinear, asymmetric, seasonal, and trendy [Hall 2018; Aiken 1996; Olmedo 2014; Katris 2019a].

Aiken [1996] was the first researcher to employ a multilayer perceptron (MLP) to address the challenges associated with forecasting unemployment rates with traditional statistical methods. The MLP model outperformed traditional statistical methods, as measured by the mean average error. Since then, several researchers have demonstrated that machine learning models have a higher accuracy than traditional statistical methods when forecasting unemployment rates [Katris 2019a; Cook and Hall 2017; Mahipan *et al.* 2013; Atsalakis *et al.* 2007; Kouziokas 2019; Sermpinis *et al.* 2014; Sharma and Singh 2016].

Although, traditional statistical methods are still preferred when it comes to forecasting unemployment, machine learning models are showing a lot of promise. There is currently an increase in research to improve the forecasting accuracy and adoption of machine learning models [Makridakis *et al.* 2018; Hall 2018; Cook and Hall 2017; Sharma and Singh 2016; Sermpinis *et al.* 2014]. Much of this work is in North America, Europe, and Asia and not much in Africa [Makridakis *et al.* 2018; Hall 2018; Cook and Hall 2017; Sharma and Singh 2016; Sermpinis *et al.* 2014].

At the writing of this paper, applications of machine learning models to forecast unemployment rates were still at experimental stages [Makridakis *et al.* 2018; Hall 2018; Cook and Hall 2017]. Researchers thus far have focused on forecasting unemployment t-periods in the future, with few using the data to determine which features drive unemployment [Kreiner and Duca 2019; Hall 2018; Katris 2019a]. The majority of these models are univariate models using past unemployment rates as the only predictor of future unemployment [Makridakis *et al.* 2018; Cook and Hall 2017; Hall 2018; Olmedo 2014; Kouziokas 2019; Katris 2019ab].

1.2 Purpose Statement

The intention of this research was to i) apply machine learning models to forecast the South African unemployment rate, as this could improve systemic interventions through policymakers, ii) use a multivariate approach to determine which features drive unemployment in South Africa, and iii) examine the causal links between key features that drive the unem-

ployment rate.
This kind of work is currently limited in the country, there is evidence of statistical methods but not machine learning models [Burger and Fourie 2015; Kyei and Gyekye 2017].

1.3 Research Questions

In order to address the research problem and achieve the purpose of this research the following research questions were explored:

1. Compared to the vector autoregression model (VAR), the benchmark multivariate forecasting model, how effective are machine learning models (regression, neural networks, and kernel methods) in forecasting the South African unemployment rate?

2. The South African Reserve Bank tracks 147 key macroeconomic variables, these are used as features in this research. What is the impact of feature selection on the unemployment forecasting accuracy?

3. What is the relationship between the various features of the unemployment rate and how do they influence the unemployment rate and each other?

1.4 Research Contributions

This research contributes to the growing body of knowledge on applying machine learning models in the context of unemployment rate forecasting. This research will make the following contributions:

1. Use machine learning (i.e. neural networks, regression, and kernel based models) algorithms to forecast the South African unemployment rate over various time horizons (i.e. 1, 2, 4, and 8 quarters ahead). The application of machine learning to forecast unemployment rates is concentrated in North America, Europe, and Asia. Only one paper was found that applied ordinary least squares to forecast unemployment in South Africa. Therefore, this research will contribute by localising such research beyond the OLS, which is a foundational machine learning technique.

2. Use feature selection techniques (i.e. information gain, ANOVA, Elastic Net, LASSO, and recursive feature selection) to determine which features contribute significantly in predicting unemployment rates. Economic forecasting typically employs arbitrary feature selection or feature selection based on expert input. This research demonstrates automatic feature selection which allows the discovery of unorthodox features influencing the unemployment rate.

3. Use probabilistic techniques (i.e. Dynamic Bayesian Networks) to discover how various features influence one another. Past research focuses on forecasting the unemployment rate t-periods into the future with none using the data to determine the influence

network of these features. This research will do so, demonstrating the value to policy makers of influence networks.

4. Advice policymakers on opportunities to reduce unemployment based on the findings of this research.

1.5 Research Methodology Overview

Economic data was used to forecast the South African unemployment rate. These economic variables are treated as the features whilst the South African unemployment rate itself is the label. The data was sourced from the South African Reserve Bank (SARB). An exploratory research approach was followed that used past research as a guideline of which models to experiment with.

Economic data generally has different magnitudes and, therefore, the data is first scaled to ensure features are treated equally by the models. After scaling these features, twelve machine learning models are applied to forecast the South African unemployment rate: five of which are regression models namely, support vector, bayesian, linear, ridge, LASSO, and elastic net regression; four are deep learning models namely, multilayer perceptron, long short-term memory, bidirectional long short-term memory, and gate recurrent unit; and the last two being random forrest and extreme gradient boosting of the decision tree type of models. The vector autoregression (VAR) model was used as a baseline traditional statistical model against which the machine learning models were compared. The VAR is the most commonly applied multivariate statistical method for unemployment rate forecasting. These models are future discussed in chapter 2.

1.6 Dissertation Structure

This chapter provided the introduction to the research paper. Discussing the research problem, research purpose, as well as the research questions. The rest of the research paper consists of four sections. The first being the background and literature review in chapter 2. The chapter will provide a background to the traditional statistical methods that are used to forecast unemployment rates, as well as the literature review on the use of machine learning techniques to forecast unemployment rates across the world. The third section, chapter 3, is the methodology section which discusses how this research was carried out, the data used, and the preprocessing steps taken to prepare the data for modelling. Chapter 4 discusses the results achieved by this research, with chapter 5 summarising the research, stating the research contributions, future work, and policy recommendations.

2

Background & Related Work

Unemployment is considered by many to be an indicator of the economic health of a nation [Aiken 1996; Pelaez 2006]. As such, forecasting it has been a priority activity of econometricians since the dawn of forecasting. However, in South Africa, there is a limited body of literature on quantitative unemployment forecasting techniques [Burger and Fourie 2015].

The use of these machine learning techniques to forecast unemployment has been of interest to researchers since the 1990s. Aiken [1996] was one of the first to demonstrate how neural networks could be used to forecast the unemployment rate. His research attempted to address issues associated with the accuracy and reliability of traditional statistical approaches when forecasting unemployment rates. He noted that the biggest problem is that traditional forecasting techniques require certain conditions to be met by the data before being applied: stationarity, linearity, constant variance, and constant mean.

Both traditional statistical univariate and multivariate models have extensive data requirements that need to be met before being effectively used. However, this is not the case with machine learning approaches as these techniques learn the underlying pattern in the data that map features to the target variable.

This section will provide an overview of traditional statistical models and machine learning approaches used to forecast unemployment rates. The section will start with univariate and multivariate traditional models, followed by regression techniques, neural networks, and ending with nearest neighbour and ensemble techniques.

2.1 Traditional Unemployment Forecasting

Previous forecasting research defines traditional statistical methods distinctly from machine learning methods. This tradition will be maintained throughout this paper.

Makridakis [1988]; Makridakis *et al.* [2018]; Cerqueira *et al.* [2019] define traditional statistical models as "models that assume constancy of patterns and/or constancy of relationships or predictability in the way that changes occur". This distinction and definition is well established in forecasting and statistical literature [Makridakis *et al.* 2018; Cerqueira *et al.* 2019].

Concretely, Makridakis *et al.* [2020 2018] refer to a number of models as traditional statistical models: naïve forecast, random walks, simple exponential smoothing (SES), Holt, Holt-Winters, damped exponential smoothing, theta method, autoregressive integrated moving average (ARIMA), vector autoregression (VAR), threshold autoregression (TAR), and generalized autoregressive conditional heteroskedasticity (GARCH) models.

This sub-section will discuss the traditional statistical approaches that are used to forecast unemployment rates across the world. Sub-section 2.1.1 and sub-section 2.1.2 will provide the literature review of univariate and multivariate approaches.

2.1.1 Univariate Linear Models

Univariate models are linear models that forecast a variable based on its lag values and associated error terms only [Brooks 2014]. The ones that are most commonly used for forecasting unemployment will be discussed in this section.

Naïve Forecasts

Naïve forecasts are those that assume what has happened in the past will continue happening in the future. Therefore, future unemployment is forecasted based on recently observed unemployment data only i.e. $u_{t+1} = u_t = u_{t-1}$, where, u is the unemployment rate. This technique is often used as a benchmark model, assisting in evaluating if proposed models have higher accuracy [Barnichon and Nekarda 2013].

Autoregressive Integrated Moving Average (ARIMA)

Montgomery *et al.* [1998] state that the most commonly used univariate statistical model is the ARIMA model. The model states that the features, u_t, can be determined using a linear combination of past and present values of the same feature i.e. u_{t-i}, where $t - i$ is a time period i times before t. The ARIMA model is made of a combination of three common time series models:

1. *Autoregressive (AR)*, which according to Brooks [2014] is the linear regression part of the model, where the current value of a feature, u_t, depends only on the past values of the feature and an error term. This can be presented as $\phi(L)u_t = \epsilon$, where $\phi(L)$ is a linear combination as shown in Equation 2.1, with L being the lag operator, which returns the previous element of a series, for example, L^2 and L^i would return two previous data points and i previous data points respectively. The ϵ is the error term.

2. *Integrated (I)*, Brooks [2014] defines the 'I' in ARIMA by stating that "if a nonstationary series, y_t must be differenced d times before it becomes stationary, then it is said to be integrated of order d." In other words, 'I', is the mechanism to ensure the data being considered is stationary.

3. *Moving Average (MA)*, are linear combinations of the error terms of a regression process. These terms are generated independently of each other, are not correlated with one another, and have a constant mean and variance [Brooks 2014]. MA is given by $y_t = \theta(L)u_t + \mu$, where $\theta(L)$ a linear combination as shown in Equation 2.2, although it is in a negative form in this context [Brooks 2014].

Several researchers and institutions use the ARIMA model to forecast the unemployment rate across the world [Montgomery *et al.* 1998; Barnichon and Nekarda 2013; Mahipan *et al.* 2013; Funke 1992; Katris 2019a; Hall 2018; Jelena *et al.* 2017]. In Equation 2.3, the ARIMA equation is shown, where u_t is the unemployment rate being estimated [Montgomery *et al.* 1998]. Variables p, d, and q are natural numbers used to set the autoregression order, the differencing order that makes the time series stationary, and moving average window.

$$\phi(L) = 1 - \phi_1 L - ... - \phi_2 L^p, \tag{2.1}$$

$$\theta(L) = 1 - \theta_1 L - ... - \theta_q L^q, \ and \tag{2.2}$$

$$\phi(L)(1-L)^d u_t = c + \theta(L)\epsilon + t. \tag{2.3}$$

Fractional ARIMA (FARIMA)/Generalised Autoregressive conditionally heteroscedastic (GARCH)

The ARIMA model can cater for different movements in time series data through modified versions of the ARIMA: Seasonal ARIMA (SARIMA) can cater for data with seasonality and Fractional ARIMA (FARIMA) allows for improved modelling of nonstationary data. Fractional ARIMA models are a generalised version of ARIMA that allows for long-term memory to be incorporated into the model.

According to Katris [2019ab]; Brooks [2014], FARIMA/GARCH extend the FARIMA and enable the modelling of data with differing variances (heteroskedastic time series). This model is preferred in modelling time series whose variance is not constant across its life span. The model can forecast unemployment even when its movements are affected by market shocks such as recessions with greater accuracy than ARIMA, SARIMA, and FARIMA [Katris 2019ab].

Holt-Winters

Hyndman and Athanasopoulos [2018] state that Holt is a model that takes trends into account in how it forecasts. Holt-Winters is an extension of Holt, allowing for seasonality

to be captured in models. The model enables trends and seasons to be modelled, by using exponential smoothing to represent and capture past data. Katris [2019a] states that Holt-Winters and ARIMA are standard benchmarks when modelling time series data. Showing that the model in some instances is more accurate than FARIMA/GARCH when modelling unemployment rates.

Threshold Autoregressive Models (TAR)

Although ARIMA has proven successful in the past, these models, along with other univariate models are unable to accurately forecast asymmetric variables. Unemployment data is typically asymmetric in that it sometimes declines when it is expected to increase, such as during the South African "jobless growth era", depicted in Figure 2.1 [Vermeulen 2017; Marinkov and Geldenhuys 2007]. Therefore, nonlinear models are preferred over them for such instances. TAR is such a model, that has been shown to be adequate in addressing the asymmetric nature of unemployment [Montgomery *et al.* 1998].

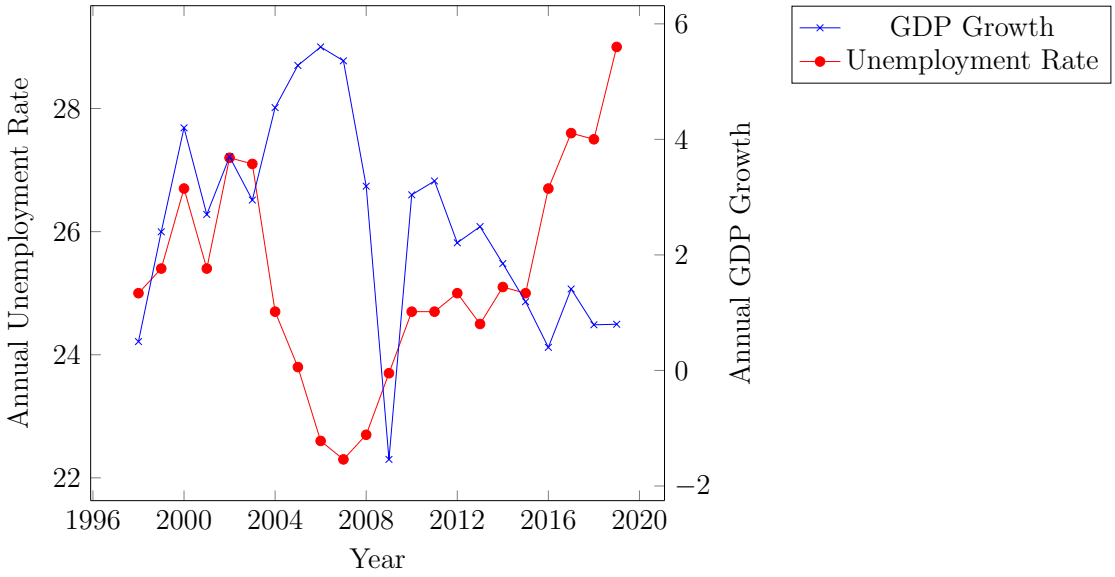


Figure 2.1: The South African unemployment rate is asymmetric. Between 2003 and 2006 unemployment decreased as GDP increased. However, between 1998 and 2001 unemployment increased as GDP increased.

TAR is able to accommodate asymmetries because the model allows for regime shifting [Montgomery *et al.* 1998; Fok *et al.* 2005; Fransesa *et al.* 2004]. Therefore, as the movement in unemployment rates change due to structural changes, these models can accommodate the changes. TAR can cater for multiple regime shifts through its modified versions: Smooth TAR (STAR) and Self-Exciting TAR (SETAR) [Fok *et al.* 2005; Pelaez 2006].

Fransesa *et al.* [2004] used root mean squared error ($RMSE$) and mean absolute percentage error ($MAPE$) as performance measures in an experiment to forecast unemployment using

two regime shifting models: TAR and Markov Switching Models. These models had lower error rates than the AR model in three of the four countries in the experiment. However, in Germany the AR models outperformed the regime shifting models. This is consistent with Clements *et al.* [2004]; De Gooijer and Hyndman [2006] who state that nonlinear models outside of machine learning, are still in their "infancy" and do not always produce better forecasts.

Univariate models are limited by the fact that the models only rely on the past values of a feature to determine its future value [Brooks 2014; Hyndman and Athanasopoulos 2018]. These models only model linear data and are unable to capture regime shifts (structural changes) reliably [Brooks 2014; Hyndman and Athanasopoulos 2018]. Nor are they able to model data that is asymmetric [Montgomery *et al.* 1998]. Regime shifts and asymmetries are common in movements of the South African unemployment rate, along with seasonality and an upward trend [Statistics South Africa 2019; Vermeulen 2017]. Furthermore, the ARIMA model requires extensive preprocessing when the underlying data is seasonal and trendy [Hyndman and Athanasopoulos 2018]. Therefore, multivariate models are explored as they address some of the limitations of univariate models

2.1.2 Multivariate Models

Multivariate models are generalisations of univariate models. They essentially allow for the forecasting of features and one or more targets simultaneously. The common model used to model unemployment will be discussed in this sub-section.

Vector Autoregressive Models (VAR)

The generalisation of ARIMA models from single dependent variables to multiple is vector autoregression models (VARs) [Montgomery *et al.* 1998]. A VAR is a regression model with more than one feature. The features are also assumed to equally influence each other [Hyndman and Athanasopoulos 2018]. A VAR model with two features can be written as,

$$y_{1,t} = c_1 + \phi_{11,1} y_{1,t-1} + \phi_{12,1} y_{2,t-1} + \epsilon_{1,t} \; and \tag{2.4}$$

$$y_{2,t} = c_2 + \phi_{21,1} y_{1,t-1} + \phi_{22,1} y_{2,t-1} + \epsilon_{2,t}, \tag{2.5}$$

where, $y_{i,t}$ is the i-th feature at time t, c_i is a constant, $\phi_{ij,t}$ captures the influence of feature i on feature j at time t. The ϵ terms are white noise generated errors [Hyndman and Athanasopoulos 2018].

In Equation 2.4 and Equation 2.5 it is also clear to see that each feature ($y_{i,t}$) is estimated by its own lags and the lags of all the other features in the model. The influence of each model is assumed to be equal across the different features [Hyndman and Athanasopoulos 2018]. This fact makes feature selection impossible with VAR.

This model was used by Montgomery *et al.* [1998] to model unemployment trends in the United States of America (USA) and found that the model had a higher accuracy when compared to ARIMA.

Cook and Hall [2017], however, noted that VARs have a key limitation in that if one wants to model complex data or nonlinearities, it would be dependent on the researcher being able to manipulate the data to suit the model which requires the researcher to have deep knowledge of the underlying data structure. The VAR moving average (VARMA) model, a generalised version of VAR, is an alternative to VAR that is also employed in unemployment forecasting [Pelaez 2006]. However, this model does not solve the challenges associated with VARs.

Yang [2007] attempted to overcome these challenges by introducing the seasonal additive nonlinear VAR model (SANVAR) and successfully demonstrated that it could model unemployment rates in the USA with a higher accuracy in most cases than univariate models. The model was able to draw insights around how people of different races and genders are affected by different unemployment drivers. The model, however, has drawbacks in that it requires extensive set up and expert knowledge from the researcher to successfully deploy.

2.1.3 Known Challenges with Traditional Forecasting Models

This sub-section provided a view of the traditional approaches to unemployment forecasting which have been in use for decades. These models have several limitations which the machine learning approaches, discussed in section 2.2, attempt to overcome. The key challenges being:

- Hall [2018]; Aiken [1996]; Pelaez [2006] state that the models are best catered to forecast linear data and often require a significant amount of tweaking to enable them to model nonlinear data. The tweaks themselves are complex as they require the researcher to have intimate knowledge of the data itself.

- The models assume that the data is symmetric and are prone to error when the data is asymmetric as unemployment data is, although threshold approaches try to address this [Marinkov and Geldenhuys 2007; Hyndman and Athanasopoulos 2018].

- The models are prone to error when there are economic shocks which cause the observed variable to behave in a manner which is not synonymous with more recent observations [Hall 2018; Aiken 1996; Olmedo 2014].

- The data is often required to meet certain data quality requirements such as white noise (randomly generated uncorrelated variables with zero mean and constant variance) and stationarity [Aiken 1996; Brooks 2014; Katris 2019a; Olmedo 2014].

- As there are hundreds of possible variables to consider in forecasting unemployment, the limitation in the method and tools demand that the researcher select a particular

subset of variables. This selection is often done arbitrarily or based on economic theory and therefore not data-driven [Katris 2019b; Davies *et al.* 2009; Liang 2005; Kyei and Gyekye 2017].

- The VAR model assumes that each feature is equally important in predicting the target, making it impossible to determine which features are most important [Montgomery *et al.* 1998; Hyndman and Athanasopoulos 2018].

This sub-section provided a review of traditional statistical methods that are used for forecasting unemployment across the world. Section 2.2, provides a literature review of the machine learning approaches that have thus far been employed to forecast unemployment. These models address the limitations of traditional unemployment forecasting model and offer an improvement in forecasting accuracy.

2.2 Machine Learning Unemployment Forecasting

Goodfellow *et al.* [2016] state that machine learning is a programmatic process of enabling computers to 'learn' from experience. Where 'learn' is defined as the ability to "improve over task T, with respect to performance measure P, based on experience E" [Mitchell 1997]. A task could be forecasting the unemployment rate, the measure being prediction accuracy measured by the mean absolute error, and the experience is the time series data.

Therefore, machine learning systems are those that are able to detect and generalize patterns from data [Mitchell 1997]. Improving the accuracy of the generalization as the data volume increases [Mitchell 1997; Goodfellow *et al.* 2016]. This sub-section will provide a literature review of the key machine learning methods that are used to forecast unemployment rates. Regression (subsection 2.2.1), kernel (subsection 2.2.2), neural network (subsection 2.2.3), nearest neighbour (subsection 2.2.4), and ensemble (subsection 2.2.5) machine learning approaches are discussed in this section.

2.2.1 Regression

Hyndman and Athanasopoulos [2018] state that regression is an approach for estimating the relationship between one or more variables. Where the variables are either features or a target. The models inherently assume that there is a relationship between features and target. As such, they are models that try to estimate the influence of features on the target. This section will discuss the common regression techniques used to model unemployment rates.

Least Absolute Shrinkage Selector Operator (LASSO) Regression

Linear regression is a fairly simple regression technique. This method is essentially the ordinary least squares (OLS) method shown in Equation 2.6. The method is useful for linear

contexts and is used to a limited extent in unemployment rate forecasting [Kyei and Gyekye 2017]. Linear regression also offers little avenues to avoid overfitting.

LASSO was introduced to extend the capabilities of the OLS method to enable overfitting to be addressed through regularization [Kreiner and Duca 2019]. The LASSO model minimizes prediction error through a regularization process. The model penalizes the reliance on particular variables for prediction, therefore, reducing overfitting. This regularization is known as L_1 regularization and is shown in Equation 2.7. Equation 2.7 shows the equation of the LASSO model:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 , \tag{2.6}$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 . \tag{2.7}$$

where, y is the actual value the algorithm is estimating, X is the input data, and λ_1 is the LASSO regularization term: $\lambda_1 \|\beta\|_1$ is referred to as L_1 regularization (the norm of the βs i.e. $\|\beta\|_1 = \Sigma_{i=0}^{n} |\beta_i|$).

Kreiner and Duca [2019] used LASSO regression to forecast unemployment rates in the USA, demonstrating that LASSO can improve the forecasting accuracy of the USA unemployment rate as well as identifying variables which are most important in forecasting unemployment. Interestingly, the model suggested that international data such as German job vacancies and Australian treasury rates were amongst the ten most important features in predicting the USA unemployment rate.

Elastic Net (ENET)

Hall [2018] states that core to machine learning algorithms is minimizing errors relating to bias and variance, which coincidentally have a trade-off relationship. The Elastic Net (ENET) model was designed to specifically minimize both these errors through regularization. The model has penalties for overfitting by penalizing the model if it is overly complex (too many variables) or has over-reliance on particular variables. Through this process, the ENET learns which features are most important in the data without a need for the researcher to make assumptions as is required in the traditional forecasting approaches.

Equation 2.8 and Equation 2.9 shows two mathematical representations of ENET. The unknown in the equation is β, which the ENET algorithm attempts to estimate such that the error $(y - X\beta)$ is minimized [Hall 2018]. It is also clear that Equation 2.8 is essentially Equation 2.7 with an additional regularizer, this is because ENET is an extension of LASSO.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \, and \tag{2.8}$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \alpha\rho \|\beta\|_1 + \frac{\alpha(1-\rho)}{2} \|\beta\|^2, \tag{2.9}$$

where, y is the actual value the algorithm is estimating, X is the input data, and the λs are a regularization term: $\lambda_1 \|\beta\|_1$ is referred to as L_1 regularization (the LASSO regularizer) and $\lambda_2 \|\beta\|^2$, the L_2 regularization (the euclidean norm of the βs).

In Equation 2.8, the λs penalizes the model for over-reliance on particular variables, therefore, ensuring β remain small whilst ensuring features with the lowest predictive powers are driven to zero. However, the λs are not bounded, although positive, and therefore can take any real positive value [Hall 2018]. Hence Equation 2.9 was introduced that replaces λ with α and ρ which are also regularization terms but are bounded between 0 and 1 [Hall 2018].

Hall [2018] demonstrated that the ENET can forecast unemployment over 3, 6, 9, 12, and 24-month horizons more accurately than traditional forecasting models and professional forecasters which were used as baseline models. This model was able to identify unemployment shifts over recessions and booms more accurately than the baseline models. Through the regularization process of ENET, the model was able to identify key variables that predict unemployment by setting all other coefficients to zero. In Hall [2018], the most important features identified by the ENET model for the USA were: housing, manufacturing, and interest rates.

Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is an extension of linear regression models, which enable the modelling of nonlinearities. The model is non-parametric and assumes no particular relationships amongst the features and target [Katris 2019b]. Katris [2019b] demonstrated that MARS models offer an improvement in forecasting although the improvement was not as significant as neural networks.

Diana *et al.* [2014] also used a multivariate regression model to forecast unemployment in Indonesia. They found similar results as Katris [2019a]: improvement in forecasting accuracy. Diana *et al.* [2014]'s implementation was, however, semi-parametric and it used Bayesian techniques to determine the forecasting confidence intervals. The researchers state that semi-parametric multivariable regression models offer an improvement over standard non-parametric MARS approaches by removing assumptions on how the regression function should look. The data is used to derive the regression function.

2.2.2 Kernel Methods

Support vector machines (SVMs) are an approach to supervised learning that enables nonlinear data to be classified using a hyperplane [Goodfellow *et al.* 2016]. The approach outputs a class identity, categorising the inputs into varying classes. These models are useful for modelling nonlinear data. There are two types of SVMs: Support Vector Classifiers (SVC) and Support Vector Regression (SVR).

To use SVMs for regression problems, an implementation of the ε-sensitive loss function is required which enables nonlinear regression modelling [Sermpinis *et al.* 2014]. The loss function is shown in Equation 2.10, where, y_i is the value being estimated, $f(x_i)$ is the predicted output, and ε is the error. Equation 2.11 shows the SVR estimator function [Ahmed *et al.* 2010]:

$$L_\varepsilon(x_i) = \begin{cases} 0, & \text{if } \mid y_i - f(x_i) \mid \leq \varepsilon \ , \ \varepsilon \geq 0 \\ \mid y_i - f(x_i) \mid - \varepsilon, & \text{if other} \end{cases}, \tag{2.10}$$

$$f(x) = \sum_{n=1}^{M} (\alpha_m^* - \alpha_m) x^T x + b, \ \textit{and} \tag{2.11}$$

$$J = C \sum_{n=1}^{M} |L_\varepsilon(x_i)| + \frac{1}{2} \|w\|^2 . \tag{2.12}$$

Where, α_m^* and α_m are Langrange multipliers, x is the input data, and b is the bias term. Langrange multipliers are used because the SVR is fundamentally a constraint optimization problem. These multipliers are useful in solving constraint optimization problem at the cost of adding additional variables to the model: the multipliers.

Equation 2.11 is optimized by minimizing the loss function shown in Equation 2.12. In which the first term is the ε-sensitive loss function and second a regularization term that penalizes the model if it overly relies on particular features.

Cook and Hall [2017]; Katris [2019b] demonstrated that the use of nonlinear kernels in SVRs makes them effective in unemployment forecasting tasks with an accuracy that is above linear regression. The kernel is a function that transformations the feature space into a higher one where a decision boundary can be more 'easily' fit by the support vector.

2.2.3 Neural Networks

Neural networks offer a significant advancement in time series forecasting because these techniques do not require particular data assumptions to be met before being used and they can function with incomplete or imperfect data, unlike other models that present issues when

data completeness is a requirement [Aiken 1996]. Therefore, neural networks are a preferred choice when forecasting nonlinear time series data.

Aiken [1996] was the first to demonstrate that neural networks can produce significantly better unemployment forecast estimates compared to traditional forecasting models. Table 2.1 provides more information on Aiken [1996]'s results along with other machine learning models discussed in this section. Since Aiken [1996], several other researchers have investigated various neural network architectures for forecasting unemployment, these are discussed in this sub-section.

Feed Forward Neural Networks (FFNN)

Feed Forward Neural Networks (FFNN) are neural networks where the data flows from input to output in a 'forward' direction without moving backwards. Their goal is to find a predictor function, f, which maps an input x to some output y i.e. $y = f(x)$ [Goodfellow *et al.* 2016]. FFNNs are foundations to more advanced neural networks and they are referred to by multiple names: Multilayer Perceptron (MLP) or Deep Feed Forward Networks. FFNNs enable nonlinear relationships to be effectively modelled, which is important for unemployment data.

Cook and Hall [2017] investigated the use of fully connected FFNNs to predict unemployment rates in the USA and they found them to perform better than Surveys of Professional Forecasters. Aiken [1996]; Kreiner and Duca [2019]; Pelaez [2006] were able to demonstrate that FFNNs offer a significant improvement in forecasting results when compared to traditional unemployment forecasting techniques.

Katris [2019a] found that in Mediterranean countries, FFNN were sometimes outperformed by FARIMA/GARCH and Holt-Winters models on preprocessed data that showed evidence of nonlinearity, heteroskedasticity, and non-normality. The key reason given for the underperformance by the neural network was the simplicity of the FFNN structure: a single hidden layer model with 1 to 10 nodes and a sigmoid activation function. Cerqueira *et al.* [2019] also observed that large data sets are required for neural networks to outperform traditional statistical methods. Therefore, it is clear that neural networks need sufficient data and an appropriate architecture to perform effectively.

Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) are a type of neural network architectures that enable effective modelling of sequences of data as they possess an internal memory structure [Goodfellow *et al.* 2016]. At their core they are essentially FFNNs with feedback loops [Goodfellow *et al.* 2016]. RNNs were developed for natural language processing requirements such as language translation.

Cook and Hall [2017] demonstrated that these models can forecast unemployment accurately

by using a variation of RNNs referred to as long short term memory (LSTM) networks. The model outperformed the ARIMA model, using only past unemployment rates in the USA as the input to the model.

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a class of neural networks that are used primarily for image classification [Goodfellow *et al.* 2016]. These networks employ a mathematical procedure called convolution. Through convolution, different filters and aspects of an image can be detected such as edges. Although CNNs are not typically used for time series data, they were used to predict unemployment rates and were found to perform poorly when compared to other neural network architectures but with higher accuracy rates than ARIMA [Cook and Hall 2017]. The results are consistent with the intention of the architecture which is to extract features from images and not to perform regression type activities.

Encoder Decoder (ED)

Encoder Decoder (EDs) LSTM networks are a specialised form of LSTM networks, that were designed for predicting words that are likely to be used within a particular sentence structure [Cook and Hall 2017]. These models have the ability to 'remember' things over a long period of time which makes them ideal to predict data over long horizons. Cook and Hall [2017] found the models to perform better than all other neural network models that were used as benchmarks (FFNN, CNN, and RNN). Part of the reason for this superior performance is that EDs are a combination of two LSTM models that have been tweaked to extend the long-term memory capacity.

Psi Sigma Neural Network (PNN)

Psi Sigma Neural Networks (PNN) are a specialised kind of FFNNs where the way in which weights are updated is modified to use a partial update methodology. Sharma and Singh [2016] used the PNN with six inputs, five hidden layers and one output, where the weight of the last layer was fixed to 1 and was never updated throughout the training process. The minimised function used was,

$$E(c, w_j) = \frac{1}{N} \sum_{n=1} N(y_n - f(w_k, c)^2) \tag{2.13}$$

where, y_n is the value being estimated, f is the estimated value, w_k are the adjustable weights, and c is an adjustable term. The adjustable term enables the model to provide better estimates: improved accuracy and reducing overfitting. The term can be compared to the traditional sigmoid functions in normal neural networks.

The model was used to forecast unemployment rates and was compared to FFNN and RNNs. The results showed that the PNN had a higher error rate which was five times higher than FFNNs and RNNs.

Radial Basis Function Neural Network (RBFNN)

A Radial Basis Function (RBF) is a function that is determined by the distance between the input value and a predetermined fixed point. According to Sharma and Singh [2016], RBFNN are neural networks that use the RBF as their activation function as opposed to the traditional sigmoids used in other neural networks. The output of these models is a linear combination of RBFs. This model, when compared to RNN and FFNN, had a mean average error that was four times higher [Sharma and Singh 2016].

Sharma and Singh [2016] found that although the mean average error of RBFNN was five times higher than RNNs and FFNNs when these models are combined with Learning Vector Quantization (LVQ) the error was reduced by close to 50%. Mathworks [2020] states that LVQs are classification neural networks that use unsupervised learning techniques to extract features which are then classified into individual classes. Adding that the unsupervised technique leverages competitive learning, where, all neurons compete to be activated resulting in only a single one being activated at each layer. Hence the approach is sometimes called a 'winner takes all' approach [Sharma and Singh 2016].

Adaptive Neural Fuzzy Inference System (ANFIS)

Atsalakis *et al.* [2007] define a Neural Fuzzy System as "a combination of Artificial Neural Networks (ANN) and Fuzzy Inference System (FIS) in such a way that neural network learning algorithms are used to determine the parameters of FIS". One such combined system is the Adaptive Neural Fuzzy Inference System (ANFIS). Where, a FIS is a system that uses fuzzy set theory to map inputs to outputs, it uses 'IF...THEN...' rules as well as logical operators AND and OR to achieve its mapping.

Atsalakis *et al.* [2007] investigated the use of ANFIS to forecast unemployment in Greece and these models performed better than ARIMA models. Therefore, demonstrating that these models are legitimate alternatives for forecasting time series data.

Yolcu and Bas [2016] used several fuzzy logic techniques to forecast unemployment rates in Turkey: one of which is the ANFIS (discussed in Table 2.1 along with their other models). Their results are similar to that of Atsalakis *et al.* [2007], in that their fuzzy time series forecasts had low error rates: affirming the suitability of these techniques to forecast unemployment rates. Yolcu and Bas [2016], therefore, demonstrate that the fuzzy logic approaches are appropriate for forecasting unemployment rates. Adding that these techniques offer advantages over others because they are able to perform optimally with small datasets.

2.2.4 Nearest Neighbour

Yang [2007] used a technique similar to K-nearest neighbour to predict unemployment rates in the USA. Their model is not named, however, it is a nonlinear and non-parametric near-

est neighbour method. The method works by splitting the unemployment time series into a subset referred to as the history of the time series and a subset referred to as the future of time series. This is similar to the train and test data splits that are common in machine learning.

The algorithm then estimates the future from the history, by lopping through the historical values and calculating their distance from the future value. A subset of n historical values that are closest to the future values are selected. If the n values form a simplex - a triangular shape - that contains the future value, then the n values can be used to forecast the future value. The forecasted future value is predicted using the weighted average of the nearest neighbours.

Yang [2007]'s nearest neighbour technique was able to forecast unemployment rates more accurately (as measured by MSE) than ARIMA, TAR, BVAR, and professional forecasters for multiple horizons: 3, 4, 5 ..., and 10 months ahead. The ARIMA was, however, more accurate in one and two months ahead forecasts. These results show that Yang [2007]'s technique can cater for uncertainly over longer horizons than traditional statistical methods. The technique, however, did not perform as well when it was predicting unemployment using quarterly unemployment rates as compared to monthly ones.

2.2.5 Ensemble Learning

Cook and Hall [2017] noted that when comparing forecasting models to Surveys of Professional Forecasters (SPF) it is important to realise that their forecast is in an average forecast of a number of different forecasters who use different methods and approaches. On that basis, Cook and Hall [2017] demonstrated that training four different neural networks independently, these could then be combined to form an Ensemble model (EM) which combines predictions from the individual architectures to a new improved model.

The combined model was able to out perform each individual model significantly, therefore, showing that such approaches can be leveraged to improve accuracy of unemployment rate forecasts. Makridakis *et al.* [2018] also noted that ensemble techniques have grown in popularity and accuracy over the past few years and are now the best performing forecasting techniques in forecast competitions.

2.2.6 Model Comparison

Table 2.1 shows a performance comparison of machine learning models used to forecast unemployment rates across the world. Most the models were discussed in the chapter 2. This comparison gives a comprehensive review of machine learning models used to forecast unemployment, their performance, and implementation specification. Most of these models are applied to univariate data, benchmarked against the ARIMA model with few being multivariate.

In cases where multivariate data is used in the research, dimension reduction techniques such as PCA are used as well as feature selection techniques such as LASSO [Sermpinis *et al.* 2014; Kreiner and Duca 2019; Hall 2018]. These methods are employed to reduce the computational requirements of the model.

Katris [2019a] shows that model architecture makes a difference in performance outcomes. Their finding is that simple neural network architectures can be outperformed by traditional statistical models by preprocessing the data to meet the statistical requirements. Furthermore, feature selection is fairly arbitrary and often based on the domain knowledge of the researcher [Aiken 1996; Atsalakis *et al.* 2007; Kouziokas 2019; Diana *et al.* 2014]. Therefore, Table 2.1, serves as a foundation on which this research is built on.

The table also shows that FFNN are the most commonly used machine learning forecasting techniques. Other methods are not used as often.

Table 2.1: A detailed comparison of machine learning approaches to forecasting unemployment.

	Models															Evaluation *			
Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Aiken [1996]	■															0,12			A FFNN was used with a genetic algorithm. Training was run for millions of iterations until the error rate was 3%. The researcher considered 250 macroeconomic indicators over a period of 50 years, after analysis seven features where chosen from economic theory and statistical analysis of the data. The features included: Money growth (money supply Ml,money supply M2), interest rates, exports, consumer expectations, consumer sentiment, the composite index of 11 leading indicators, and the Center for International Business Cycle Research's composite index.
Katris [2019a]	■															0,12			A FFNN was used with a single hidden layer, with variations in the number of nodes from 1 to 10. The sigmoid activation function used. The R package AMORE was used in the implementation with 500 epochs. The data was preprocessed to ensure nonlinearity, heteroskedasticity and non-normality, which is a prerequisite for FARIMA/GARCH model (the baseline). Only past unemployment rates were used as input to the model. The data was accessed from the Eurostat database.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Cook and Hall [2017]	■	■	■				■									45,50			Data was split into training, validation and testing set, where, the training was from 1963 to 1996 with every tenth observation added to the validation set. The remaining data was used for testing. A stochastic gradient descent was applied with dropout for regularization. The researchers opted for a univariate approach, where only past unemployment rates were used as the features.
Mahipan *et al.* [2013]	■															65,35			Four different FFNNs were considered as candidate models each using the tangent activation function, with changes in the number of hidden layers and the size of the network which varied as follows: 12 input variables, 11 hidden layers and 1 output; 2 input variables, 2 hidden layers and 1 output; and 6 input variables, 11 hidden layers and 1 output. 30% of the data was used for testing purposes. In the preprocessing, the data was normalized to ensure each value was in the range [0.1 - 0.9]. The learning process was stopped after 50 000 epochs were reached. The data was accessed from the UNational Statistical Office of Thailand database, where only past unemployment rates were used. It is important to note that MAPE was used instead of MAE.
Olmedo [2014]	■																	0,02	FFNN with backpropagation method was employed with past unemployment rates being the only input to the model. The data was accessed from Eurostat database.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Atsalakis *et al.* [2007]								■										0,34	FFNN techniques combined with Fuzzy Logic techniques were run over 250 000 epochs. Features used where: population, population change, education, popularity of education courses, types of jobs by education, pay, percentage of labour force in government job, past unemployment, private consumption expenditure, government expenditure, exports of goods and services, current unemployment rate and the gross national product. Trial and error was used to determine the final feature list. The data was accessed from Bureau of Labor Statistics in Greece.
Yolcu and Bas [2016]								■	■								0,67		Four different fuzzy logic techniques were used. The standard steps in their fuzzy logic algorithms are fuzzification, determination of fuzzy relations, and defuzzification. They employed four different approaches for the 'determination of fuzzy relations' step: standard matrix operations, linguistically defined relationships, feed forward neural networks, and a clustering algorithm. The data used was Turkey's unemployment rate data from January 2005 to December 2013 with 11 different age groups.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Kouziokas [2019]	■																	0,25	The Levenberg-Marquardt algorithm was used as the learning algorithm for the FFNN, it is fast and iterative algorithm. It combines Gauss-Newton algorithm and the steepest descent algorithm used for solving nonlinear least squares problems. The data collected was split into training (60%), validation (20%) and testing (20%). Through trial and error, it was found that a FFNN with 14 neurons was most effective with the linear activation function for the first hidden layer and 9 neurons and Tanh-Sigmoid transfer function in the second hidden layer as the activation function. 8 epochs were used. The past unemployment rates, GDP growth, exports of goods and services were used as the model's features. The data was accessed from the UK Office for National Statistics.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Sermpinis *et al.* [2014]	■											■	■			0,01			A Genetic Algorithm algorithm with high cross over probability of 0,85 was used to determine the feature set. The Support Vector Regression was then used to forecast unemployment rates. The initial population size was 400 chromosomes with maximum generations set at 5 000 with roulette wheel selection and the best member of each generation is maintained (elitism). Essentially, the best features are discovered through evolution. 11 predictors were used: HOUSE, INDP, M1, EMPL, PCE, PI, WAGE, and DJIA with WAGE and INDP being the consistently preferred features. The data was accessed from the Federal Reserve Bank of St Louis (FRED) and Bloomberg. The preprocessing step included seasonal adjustments to the data.
Hall [2018]											■					0,66			The ENET was used with 50% of the data reserved for training, leveraging a rolling forecast framework, which enables period $t + 1$ to be forecasted using data until period t. The model started with 138 macroeconomic variables drawn from a number of economic categories in the FRED-MD database from March 1959 to April 2017. The ENET was able to discard features through its regularization process. The data was accessed from the Bureau of Labor Statistics, Federal Reserve Bank of St. Louis FRED.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Kreiner and Duca [2019]	■																0,13		Size of FFNN was selected through trial and error, with one to five hidden layers being considered each with 1 to 150 neurons. The model started with 600,000 variables in the FRED database, structured as a recursive tree with data binned into eight categories. PCA was then applied to reduce the variables to 185. 3 hidden layers were used each with 97 neurons per layer with a logistic activation function. The data was accessed from FRED.
Kreiner and Duca [2019]															■		0,17		Through regularisation LASSO was able to set all that coefficients to zero with only 10 remaining. The 10 most important features were housing starts, the change in non-financial corporate commercial mortgages, unfilled German job vacancies, retail vehicle registrations, weekly aggregate payrolls, household net lending, 3-month Australian Treasury rates, and producer prices for pharmaceuticals and tractors. The data was accessed from FRED.
Sharma and Singh [2016]	■															0,34			The FFNN had 8 input neurons, 6 hidden layer and 1 output layer with a learning rate of 0,005 and 60 000 epochs. The data was accessed from FRED.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Sharma and Singh [2016]			■													0,5			The RNN had 7 input neurons, 6 hidden layer and 1 output layer were used with a learning rate of 0,003 and 50 000 epochs. The data was accessed from FRED.
Sharma and Singh [2016]				■												2,62			The PSN had 6 input neurons, 5 hidden layers and 1 output layer with a learning rate of 0,002 and of 75 000 epochs. The weight of the first layer was fixed to 1, whilst the rest are updated through the iterations. The data was accessed from FRED.
Sharma and Singh [2016]					■											2,14			The RBFN had 8 input neurons, 4 hidden layers and 1 output layer with a learning rate of 0,003 and 45 000 epochs. The data was accessed from FRED.
Sharma and Singh [2016]						■										1,22			In each epoch, the neurons 'compete' to be activated and only 1 is activated at each iteration. The data was access from FRED.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Katris [2019a]						■										0,21			The FFNN was implemented in the R package AMORE. A single hidden layer was used and trial and error was used to determine number of nodes: from 1 to 10. The model used an adaptive gradient descent with adaptive backpropagation for 500 epochs with the sigmoid activation function. The model was implemented with past unemployment rates being the only input to the model. The data was accessed from the Eurostat database.
Katris [2019a]										■						0,25			A standard SVR was implemented from the R programming language without any modifications. The model was implemented with past unemployment rates being the only input to the model. The data was accessed from the Eurostat database.
Katris [2019a]														■		0,24			A standard MARS was implemented from the R programming language. The model with lagged 1, 2, 3 and 4 variables was implemented. The model was implemented with past unemployment rates being the only input to the model. The data was accessed from the Eurostat database.

Table 2.1 continued from previous page

Author	FFNN	CNN	RNN	PNN	RBFNN	LVQ	ED	ANFIS	FL	SVR	ENET	GA	GA-SVR	MARS	LASSO	*MAE*	*RMSE*	*MSE*	Model specifications
Diana *et al.* [2014]														■				0.32	A semi-parametric multivariable regression model was used with Generalized Maximum Likelihood to estimate the smoothing spline function. Features were predetermined for the model: percentage of educated above high school, economic growth, population density, the ratio of investment to labor force, regional minimum wage, and ratio of large and medium industries to labor force. The data was accessed from the Indonesian Central Bureau of Statistics.

Legend:

- FFNN - Feed Forward Neural Network
- CNN - Convolutional Neural Network
- RNN - Reccurent Neural Network
- PNN - Psi Sigma Neural Network
- RBFNN - Radial Basis Function Neural Network
- LVQ - Learning Vector Quantization
- ED - Encoder Decoder
- ANFIS - Adaptive Neural Fuzzy Inference System
- FL - Fuzzy Logic / Fuzzy Time Series
- SVR - Support Vector Regression
- ENET - Elastic Net
- GA - Genetic Algorithm
- MARS - Multivariate Adaptive Regression Splines

* The performance measures are in most cases average performance as the researchers often had error rates calculated over multiple horizons. In other cases models are run across multiple countries, in such cases the best country's results are displayed: the results were often significantly close to one another.

2.3 Significant Findings in the Related Work

Chapter 2 shows that several traditional statistical forecasting models have been applied to unemployment time series data. These models, however, are limited when modelling asymmetric and nonlinear data, as well as, data that was not preprocessed to meet a particular standard such as white noise data [Marinkov and Geldenhuys 2007; Mahipan *et al.* 2013; Katris 2019a].

Furthermore, the models performed poorly when there were economic shocks due to recessions and booms that make unemployment rates behave inconsistently with past observations, this is largely because these models do not have long-term memory [Olmedo 2014; Makridakis 1988].

To address these challenges, machine learning models, mostly regression models and neural networks, were employed to unemployment forecasting problems. Their performance has proved that they are more successful than traditional approaches in forecasting unemployment rates. Most of the current literature on the topic compares the forecasting performance, as measured by root mean squared error ($RMSE$) or mean average error (MAE) of machine learning models against traditional statistical models.

Few researchers provide insight to the key features driving unemployment and with the researchers who did, key insights were uncovered such as how German job vacancies and Australian treasury rates affected the USA unemployment rates [Kreiner and Duca 2019; Kyei and Gyekye 2017]. Which is an odd finding as there is little action that the USA can take to mitigate this finding.

The neural network machine learning models implementation were often simple variations of a particular architecture [Katris 2019a; Cook and Hall 2017; Hall 2018]. For example, most of the neural network implementations only had one to three hidden layers and in most cases the input layer was made of past unemployment rates only. Katris [2019a] observed that in some cases simple architectures could be outperformed by more sophisticated traditional statistical models such as FARIMA/GARCH.

It is also worth noting that the vast majority of the literature on modelling unemployment rates using machine learning was in North America, followed by Europe and Asia. There were few to no similar models found for the South African context. In both the United States and Europe, key economic models such as Okun's Law ('as GDP increases unemployment should drop') and Philips Curve ('inflation and unemployment have a trade off relationship, increasing inflation reduces unemployment rates') hold strongly resulting in GDP and inflation being standard features when forecasting unemployment [Aiken 1996; Jelena *et al.* 2017]. However, Ball *et al.* [2019] observed that in developing nations and particularly South Africa, these models are often weak (low R^2) in explaining unemployment. Meaning that South Africa is well suited for machine learning approaches as traditional economic models

do not always provide the required explanations.

Machine learning models are often preferable if the model is to be inferred from the data, as is needed in South Africa. In the country, the frequency of unemployment forecasting is currently quarterly, whilst key macroeconomic indicators are monthly and weekly. This is unlike other countries where machine learning models have been applied to forecast unemployment, where unemployment rates are reported in similar frequencies as other macroeconomic indicators. Therefore, this is a motivation for the importance of South Africa having local versions of machine learning-based unemployment forecasts. It is an opportunity to use these models in forecasting contexts with varying data frequencies.

Concretely, three key literature gaps where identified by reviewing related literature:

- localization - the application of machine learning to forecast unemployment rates is concentrated in North America, Europe, and Asia. One paper was found that applied machine learning to forecast unemployment in South Africa: employing OLS. Therefore, this is a gap that needs to addressed;
- automatic feature selection - economic forecasting typically employs arbitrary feature selection or feature selection based on expert input. Some machine learning researchers have started using feature selection techniques, however, only embedded feature selection methods and genetic algorithms have been used. Therefore, it is unknown from the literature if wrapper and filter feature selection methods might result in better unemployment forecasts;
- feature influence determination - past research focuses on forecasting the unemployment rate t-periods into the future with none using the data to determine the influence network of these features. Closing such a gap would enable policymakers to perform data-driven scenario testing.

This chapter provided a literature review of the traditional statistical and machine learning approaches that have thus far been employed to forecast unemployment across the world. Table 2.1 provides a table that compares the performance and architecture of the machine learning models discussed. In chapter 3 the methodology used in this research is discussed, as well as, the data used and a statistical description of the South African context.

3 Methodology

The purpose of the research was to examine how effective machine learning models are at forecasting the South African unemployment rate. Along with identifying key features that impact the unemployment rate and their influence on each other. Past research used neural networks, regression, and kernel techniques to try and achieve similar goals.

Based on desktop research conducted on unemployment rate forecasting, machine learning techniques are yet to be applied in South Africa for forecasting the unemployment rate. Some traditional statistical approaches have been applied but to a limited extent [Burger and Fourie 2015; Kyei and Gyekye 2017]. Hence, an exploratory approach was taken for the research. Prior research was used as a guide for which models to attempt. This method was selected because the models that would perform best were not known upfront. This is further discussed in this chapter.

This chapter discusses the research methodology that was employed in carrying out this research. In section 3.1 the research strategy is provided. In section 3.2 an overview of the data sources that were used for this research is provided. In section 3.3 the data wrangling and preprocessing techniques used are discussed. The chapter ends with section 3.4, which discusses ethical considerations that were important for this research.

3.1 Research Strategy

Unemployment is an important macroeconomic indicators as it is used to signal the economic health of a country [Aiken 1996; Pelaez 2006]. In South Africa, the unemployment rate has been trending upwards for the past 12 years and there is no indication that it will stop doing so [Statistics South Africa 2019]. Numerous policies have been explored in an attempt to address this issue but none have yielded the necessary fruit [Levinsohn 2007; Brynard 2011; De Lannoy *et al.* 2018]. Therefore, accurately forecasting the unemployment rate is important.

This research followed an exploratory approach with guidance from previous literature, where the current and top performing machine learning models were explored for their ability to forecast the South African unemployment rate.

The vector autoregression (VAR) model was used as a baseline traditional statistical model against which the machine learning models were compared. The VAR is the most commonly applied multivariate statistical method for unemployment rate forecasting.

Table 3.1 shows the proprieties of the data discussed further in section 3.3. Various classes of machine learning models that have been used in previous research were analysed to determine their suitability in forecasting data with various properties observed in the South African unemployment rate data. Each class had varying strengths and weaknesses. Therefore, machine learning classes that were previously employed in similar prior research were used for this research.

Table 3.1: Properties of features and targets for modelling the South African unemployment rate: The VAR has limited capacity to cater for the varying data properties whilst the machine learning models have varying areas of strength.

Data Properties	**Regression**	**Neural Networks**	**Trees**	**Kernel**	**Vector Autoregression (VAR)**	**Comments**
Continuous	***	***	***	***	***	
Missing data (Mixed frequencies)		*	**			Because of the missing data, data imputation techniques are required in order for other models to cater for this.
High dimensional (147 features)	***	***	**	**		VAR is suitable for low dimensional data.

Data Properties	Regression	Neural Networks	Trees	Kernel	Vector Autoregression (VAR)	Comments
Different magnitudes	***	***	***	***	***	Normalization is used to ensure data is within reasonable ranges i.e. 0 to 1 or -1 to 1.
Non-stationary	***	***	***	***		VAR can only work with stationary data and data can be differenced to make it stationary prior to modelling.
Limited observations (794 observations)	***	*	***	**	***	Neural networks require large datasets to perform optimally.
Non-linear	*	***	**	***		VAR was designed for linear data. So too the OLS method in machine learning.
Feature: multi-modal	**	***	***	***		

Legend:
* Somewhat capable of handling type of data.
** Moderately capable of handling type of data.
*** Highly capable of handling type of data

The analysis of the properties of the data in Table 3.1 affirms that the selected classes of machine learning models are suitable for forecasting the South African unemployment rate. Therefore, although the approach is exploratory the analysis increases the confidence in that the models should be able to forecast the South African unemployment rate. The South African unemployment rate exhibits nonstationarity which makes it more difficult to

forecast. In other areas where machine learning models have been employed for forecasting unemployment rates, the data tends to be stationary i.e., the USA and Europe. The analysis, however, affirms that machine learning models should be able to forecast the South African unemployment rate.

3.2 Data Sources

The data used in this research is all publicly available, regularly reported, and accessible without cost. Data from the South African Reserve Bank (SARB) and Bureau of Economic Research's (BER) macroeconomic databases that track the South African macroeconomic indicators was used for this research. Table 3.2 gives more information about these databases. The list of features that was extracted from these databases is listed in Appendix A.

Table 3.2: Publicly accessible data sources that were used for the research

Data Source	**Description**	**Features**	**Preprocessing needed**
South African Reserve Bank (SARB)	"The South African Reserve Bank is the central bank of the Republic of South Africa. The primary purpose of the Bank is to achieve and maintain price stability in the interest of balanced and sustainable economic growth in South Africa. Together with other institutions, it also plays a pivotal role in ensuring financial stability" [South African Reserve Bank 2020].	There are 149 macroeconomic indicators such as: interest rates, exchange rates, gross domestic expenditure, production volumes, and other macroeconomic indicators. The data is available in quarterly and monthly frequencies from 1922 to date.	The data is available in different frequency formats: monthly and quarterly. Therefore, there is a requirement to align these dates. Furthermore, some of the data is only available for the last ten years, whilst other is available from 1960, so the missing data needs to be addressed.

Table 3.2 continued from previous page

Data Source	Description	Features	Preprocessing needed
Bureau of Economic Research (BER)	The Bureau of Economic Research is part of the University of Stellenbosch in South Africa. The database covers a range of macroeconomic variables, including the South African unemployment rate.	The database was used as a source of the South African unemployment rate as it provided the rate from 1960 to December 2019. Every other database in South Africa typically only provides the rate for the last twenty years.	The data is reported on a quarterly basis.

3.3 Data Preprocessing

The data used for this research was sourced from the SARB and BER. The unemployment rate is the target to be forecasted and it was sourced from the BER database. The rate is reported quarterly and was available from BER from January 1960 to December 2020: a total 209 observations. Data preprocessing techniques were used to prepare this data for modelling.

3.3.1 Data Cleaning

Features were acquired from the SARB database: 147 features in total. These features were accessed through the SARB online database, where macroeconomic data can be downloaded in an excel format. The list of features is listed in Appendix A. Each feature was in a separate excel document, which had its relevant time series data: observations ordered by date. The excel documents had fairly consistent formatting and, therefore, downloading and extracting of information could be automated.

However, the date formats were not consistently applied across the SARB database. Features that were reported quarterly had a format 'Q1/99' meaning 'quarter 1 of 1999'. Therefore, these dates were converted to full dates that aligned with the quarterly reporting schedule of the South African unemployment rate. The date 'Q1/99' became 1 March 1999, whilst 'Q2/99' became 1 June 1999. The date format for months had two formats, one being 'Jun, 2000' and the other 'Jun/00'. Both were cleaned to be 1 June 2000.

The names of the features were extracted from the same excel document where the name was often located in top rows of the excel document. Where features had names over multiple lines, only the first line was retained as the name of the feature. The feature names were

consistently in either of two locations in the excel document. An 'if statement' was sufficient to ensure that the names were successfully acquired for all 147 features.

Once the dates were in the right format, the data was transferred from the SARB excel documents into a dataframe. A dataframe being an in-memory database that is available in most machine learning development environments. Each feature (a single excel document) was appended to the dataframe as a column. The associated values of these were stored as the 'float' data type, which is appropriate for macroeconomic data.

3.3.2 Missing Values

The data that was used for this research was in three different frequencies: annual, quarterly, and monthly. When this data was merged into a single dataframe, the data that was in annual and quarterly frequencies automatically had missing values for every month that the monthly data had a value but annual and/or quarterly did not. The percentage of this missing data varied from 50% to 90%: the unemployment rate initially had 209 observations after being merged with the feature dataframe it had 794 observations, a 73% missing data rate (shown in Figure 3.1). This is to be expected when dealing with mixed frequency data.

Figure 3.1 shows observations from the first to the last observation: observation 1432. Observations with black highlighting are those without missing data, and those without highlighting are those with missing data. The right-most graph shows the percentage of observations without missing data at each observation.

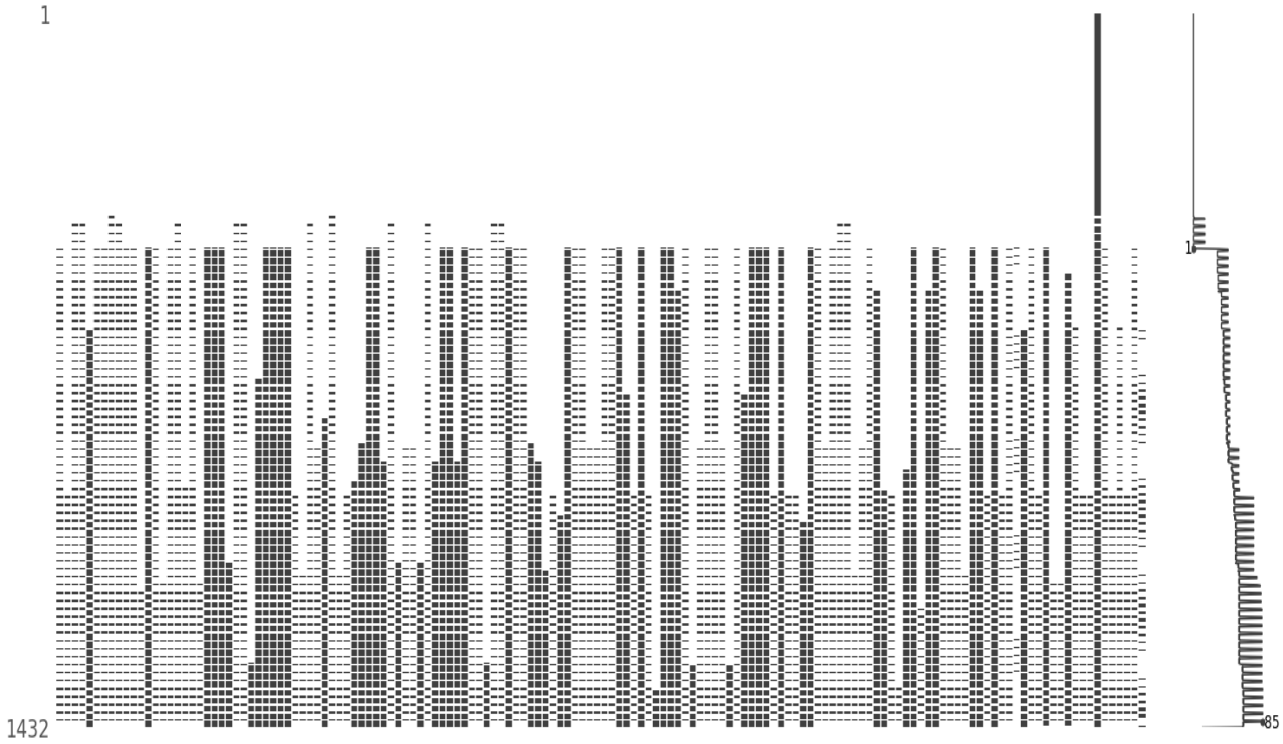


Figure 3.1: A visual representation of the missing data from the feature set. The black highlight represents observations without missing values whilst unfilled areas are those with missing values.

Prior to exploring techniques to deal with this missing data, it was noted that some of the features from the SARB database had observations from 1922. However, unemployment rates were not reported in South African back then. Therefore, the data from 1922 to 1959 was deleted to ensure that the data was kept was in date range in which unemployment rates were reported.

Missing data imputation techniques were employed to address the missing data. Six data imputation strategies were explored [Bertsimas *et al.* 2018]:

- Constant imputation - which replaces all the missing values with a constant value;
- Last known value imputation - which replaces missing data with the last valid data observation;
- Forward imputation - is the same as last known value but in reverse;
- Mean value imputation - replaces missing values with the average value of the series;
- Multivariate Imputation by Chained Equations (MICE) - which uses the values of other features to determine the possible value of the missing data; and

- k-Nearest Neighbor (kNN) imputation - replaces missing values with the nearest one as determined by the kNN algorithm.

The last known value imputation strategy resulted in lower error rates than the other imputation strategies. This strategy makes sense from an observational point of view. In months where Statistics South Africa does not report the unemployment rate, the last quarterly rate is carried over to those months i.e. the last known value is reported.

The data in this case is missing due to the collection and reporting methods. Whilst imputation methods such as constant, kNN, or mean imputation introduce new data into the feature set, last known value does not do so, because it is consistent with the reporting procedures used in macroeconomics. Therefore, this strategy is in actual fact the ground truth for the missing data. Economists always reference the last known economic value as the current value of a feature until a more recent observation is available.

3.3.3 Data Transformations

The data accessed from the SARB covers various aspects of the economy and were in different magnitudes. Some data was in tens whilst others was in billions. For example CPI is typically between 3% and 6%, population is between 40 and 60 million, and GPD between 100 and 360 billion. Therefore, in order for machine learning models to not be skewed towards features with large magnitudes, the data was normalised through scaling.

The Minimum-Maximum scaler was used for this research because it is more appropriate for data with such large differences in magnitudes as well as has different data ranges within the features. The scaler is robust to outliers and it ensures all features are within the same range [Patro and Sahu 2015]. For this research the range was 0.00001 to 1. With 0.00001 being close enough to zero but not zero to avoid any possibility of a division by zero.

Equation 3.1 and Equation 3.2 show the equation that scales features according to the Minimum-Maximum approach [Patro and Sahu 2015].

$$Feature_{standardised} = \frac{Feature - minimum\ value\ of\ Feature}{maximum\ value\ of\ Feature - minimum\ value\ of\ Feature},\ and \tag{3.1}$$

$$Feature_{scaled} = Feature_{standardised} * (1 - 0.00001) + 0.00001. \tag{3.2}$$

3.3.4 Feature Selection

The final dataframe that was used for modelling had 794 observations with 147 features. In similar high dimensional research, embedded methods and genetic algorithms were used to select a subset of features to be used for modelling. Therefore, this research, which is high

dimensional, employed feature selection techniques.

According to Chandrashekar and Sahin [2013], there are three types of feature selection methods: filter, wrapper, and embedded. Filter methods rank features based on statistical scores representing their relative significance in predicting the target variable. Embedded methods are feature selection methods that select a subset of the feature set and evaluates the performance (measured by error rate) of the subset. The feature subset with the lowest error rate is then selected as the most impactful feature set.

Wrapper methods are similar to embedded methods but are not as computationally efficient. Wrapper methods create subsets of the entire dataset first then evaluates each subset, using a greedy search strategy. Greedy search strategies are not computationally efficient: hence embedded methods are more computationally efficient relative to the wrapper methods.

Four filter feature selection methods were used in this research, as well as two embedded methods. The four were;

- Pearson correlation coefficient (removal of correlated features);
- Analysis of variance (ANOVA);
- Mutual information gain (MIG); and
- Variance threshold (removal of variables with low variance).

Along with the four feature selection methods, the deletion of duplicated features was treated as a feature selection technique. The four feature selections were also combined to form a chain of filter methods e.g. 'deletion of duplicate features and removal of correlation' or 'removal of low variance feature, deletion of duplicate features, and removal of correlation'. There was, therefore, a total of 17 different feature selection combinations that were applied on the data. This included three wrapper methods and the two embedded methods which were LASSO and Elastic Net (ENET) (discussed in chapter 2).

The recursive feature selection wrapper method was also used. However, as it is a computationally inefficient, the method easily leads to combinatorial explosion, i.e. choosing 5 features from the 147 requires evaluating over 500 million different options and choosing 6 requires over 12 billion. Hence, for this paper it was restricted to 5 features. Six or more proved to be computationally impossible on the infrastructure used for this research.

The RFE method requires a machine learning model to be used to evaluate the performance of each feature subset. Various models were explored. LASSO, ENET, and ridge proved to be the most computationally efficient. Other models like support vector regression and decision trees were not computationally efficient i.e., the infrastructure used for the

research would run out of resources [1].

Table 3.3 shows a list of feature selection techniques that were used for the research. Filter, wrapper, and embedded methods were used, covering univariate and multivariate feature selection methods.

Table 3.3: A list of features used for this research. A combination of univariate and multivariate techniques were used.

Feature Selection Technique	**Type of Feature Selection**	**Comments**
Removal of duplicates (Unique)	Univariate - Filter	Removal of duplicate features: Two features removed, 145 features retained.
Variance Threshold	Univariate - Filter	Removal of features with a low variance: 5% threshold: Two features removed, 145 retained.
Pearson Correlation Coefficient (No Corr)	Univariate - Filter	Removal of highly correlated features: 55 features retained.
Mutual Innovation Gain (MIG)	Univariate - Filter	Removal of features with the lowest information gain: 36 features retained.
Analysis of Variance (ANOVA)	Univariate - Filter	Removal of features with low F-score: 36 features retained.
Unique No Corr	Univariate - Filter	Removal of duplicate features as well as features that are highly correlated: 54 features retained.
Unique No Corr Variance	Univariate - Filter	Removal of duplicate features, removing features that are highly correlated, as well as features with low variance: 53 features retained.

[1] A macbook with 16 Gb RAM, 512 Gb SSD, 4 CPU cores, and 4 GPU cores was used. This was additionally supplemented by Microsoft Azure Virtual Machine with 32 Gb RAM, 4 cores, and 200 Gb hard drive.

Table 3.3 continued from previous page

Feature Selection Technique	Type of Feature Selection	Comments
Variance Threshold, Unique	Univariate - Filter	Remove duplicated features and removal of features with a low variance: 5% threshold: Three features removed, 144 retained.
MIG No Corr	Univariate - Filter	Removal of features with the lowest information gain as well as removing features that are highly correlate: 14 features retained.
Extreme Gradient Boosting (XGBoost)	Multivariate - Embedded	Feature importance determined by the decision tree: 5 features were selected by the model.
Random Forest (RFR)	Multivariate - Embedded	Feature importance determined by the decision tree: 8 features were selected by the model.
Elastic Net (ENET)	Multivariate - Embedded	Feature shrinkage and reduction through L1 and L2-norm regularization: 5 features retained.
LASSSO	Multivariate - Embedded	Feature reduction through L1-norm regularization: 4 features retained.
Recursive Feature Elimination (RFE) Ridge	Multivariate - Wrapper	Feature determination using iterative sub-set error scoring: the model was set to select 5 features.
Recursive Feature Elimination (RFE) LASSO	Multivariate - Wrapper	Feature determination using iterative sub-set error scoring: the model was set to select 5 features.
Recursive Feature Elimination (RFE) ENET	Multivariate - Wrapper	Feature determination using iterative sub-set error scoring: the model was set to select 5 features.

Table 3.3 continued from previous page

Feature Selection Technique	Type of Feature Selection	Comments
Principal Components Analysis (PCA)	Multivariate - Dimension Reduction	Reduction from 147 to 50 components using dimension reduction technique.

3.3.5 Statistical Description of the South African Unemployment Rate

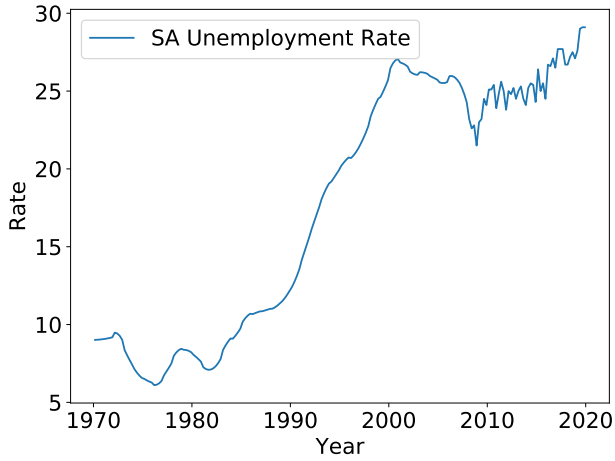
Some of content that appear in this chapter have been published at an IEEE conference:

- *Mulaudzi and Ajoodha [2020b], "An exploration of machine learning models to forecast the unemployment rate of South Africa: a univariate approach," The International Multidisciplinary Information Technology and Engineering Conference, ISBN: 978-1-7281-9519-9, 2020.*

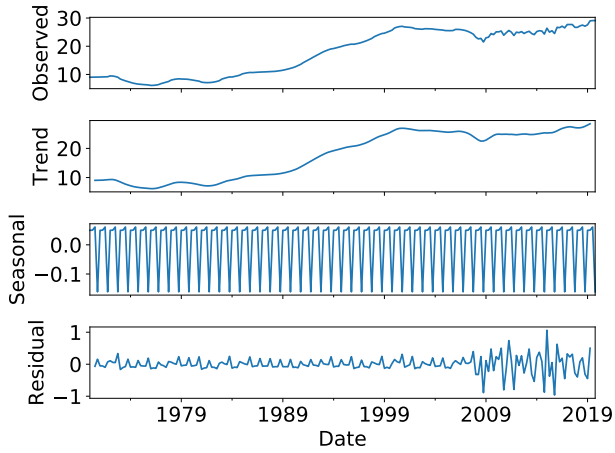
The unemployment rate used for this research was accessed from the Bureau of Economic Research (BER). The time series had a total of 209 quarterly observations of the South African unemployment rate from January 1960 to January 2020. Figure 3.2 (a) depicts this data. It can be seen that the unemployment has an upward trend and high variance over the last 10 years. The average unemployment rate over the time period is 17.7%, the lowest was 6.1% in the 1970s, and the highest rate was 29.1% which is the December 2019 unemployment rate.

Time series data can be decomposed into three key parts: the trend, seasonality, and residuals [Hyndman and Athanasopoulos 2018]. It was observed that the South African unemployment rate trends upwards and it is seasonal, where repeated movements in the data are observed each year: depicted in Figure 3.2 (b). The change in variation is visible by studying the residuals, where it can be seen that between 2009 and 2019 there was greater variability in the data than all the previous periods.

From Figure 3.3, it is clear that the South African unemployment rate does not follow a normal distribution. Kolmogorof ($kstest$) and Lillietest ($lillie fors$) tests were also conducted, which are well-known tests for normality, and they both confirmed that the unemployment rate is not normally distributed. Figure 3.3 shows a kernel distribution function whose shape suggests that the unemployment rate is bi-modal.



(a) The unemployment rate from January 1970 to January 2020.



(b) The decomposition of the unemployment rate into trend, seasonal, and residual.

Figure 3.2: Properties of the unemployment rate time series from 1970 to 2020.

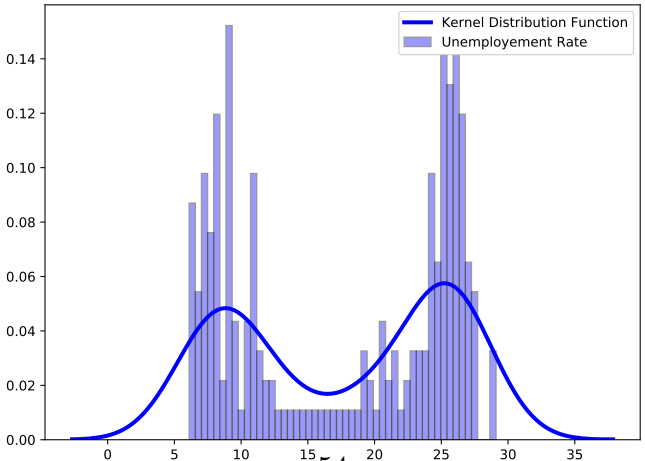


Figure 3.3: The probability distribution of the South African unemployment rate from 1970 to 2019.

It was also observed that the unemployment rate is also nonstationary over the observed period. Stationarity was tested through the Augmented Dickey–Fuller (ADF) test, whose null-hypothesis is that the series is nonstationary. The $ADF\ test$ returned a $p - value$ of 0.923 which means that we accept the null-hypothesis that the South African unemployment rate is nonstationary. This result was confirmed with two other tests: Dickey-Fuller GLS and Phillips Perron test, both of which confirmed the ADF results. These results differs from Burger and Fourie [2019], who found the unemployment rate to be stationary. However, Burger and Fourie [2019]'s data was only between a select period from 2008 to 2017, whilst these results are from 1960 to 2020. The differing results indicate that there are regime shifts within the South African unemployment: in some regimes the South African unemployment rate is stationary whilst in others it is not.

3.3.6 Performance Evaluation

The most common performance measure in time series analysis and unemployment forecasting is the mean absolute percentage error (MAPE) and root mean squared error (RMSE) [Hyndman and Athanasopoulos 2018]. RMSE is used to compare different models run on the same dataset. Whereas MAPE is used to compare models run on the same or different datasets: this is possible because MAPE is not a scale-dependent model [Hyndman and Athanasopoulos 2018; Cook and Hall 2017]. Therefore, for this research, MAPE would have been an appropriate performance measure as it allows models run on different datasets to be compared.

The challenge, however, with MAPE is that MAPE is not symmetric. It penalizes cases where the actual value is less than the forecast value heavier than when the actual value is greater than the forecast value. For example if the actual value is 10 and forecast value is 15 then MAPE would be (10-15)/10 * 100 = 50%, as opposed to the case where the actual is 15 and the forecast is 10, MAPE would be (15-10)/15 * 100 = 33% [Hyndman 2006; De Gooijer and Hyndman 2006]. Therefore, Hyndman [2006]; Hyndman and Athanasopoulos [2018], propose mean absolute scaled error (MASE) as an alternative to MAPE. MASE resolves issues associated with MAPE because it is symmetric.

The MASE equation is show in Equation 3.3:

$$MASE = \frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{y_t - \hat{y}_t}{\frac{1}{n-1}\Sigma_{i=2}^{n} \mid y_i - \hat{y}_i \mid}\right) \tag{3.3}$$

where, y is the actual value and $\hat{y}$ is the forecast value. Another valuable feature of MASE above MAPE is that it is interpretable: MASE above 1 is a MASE that is worse than a naïve model and MASE below 1 is better than a naïve model.

Time series cross validation was also employed. The "evaluation on a rolling forecasting origin" approach was selected for this research [Hyndman and Athanasopoulos 2018]. This

approach grows the size of the cross validation data over time with each test i.e. on test 1 the first 100 observations are used, on the next test the first 200, then the first 300, and so forth. The alternative approach is the 'moving window evaluation' [Hyndman and Athanasopoulos 2018]. However, this requires a random selection or trial and error selection of the window size. Hence, the first was chosen to avoid this.

The unemployment time series data was split into five sequential cross validation datasets, the MASE that was achieved was then validated by this approach. Because the data is a time series, the cross validation maintained the sequential order: there was no random shuffling of observations. Thus, essentially, in this approach the model is tested for performance on datasets of varying lengths to determine its consistency. The cross validation average error was compared against the MASE of the model to determine if the model was indeed robust and consistent in its performance.

3.4 Ethical Considerations

The data used for this research is publicly available and widely used for economic research and business reporting. Therefore, because of the orthodox nature of this data there was no ethical clearance required to access the selected data sets.

3.5 Summary

The methodology used for this research enables the research questions to be addressed. Each question is linked to an aspect of the selected methodology:

1. Compared to the vector autoregression model (VAR), the benchmark multivariate forecasting model, how effective are machine learning models (regression, neural networks, and kernel methods) in forecasting the South African unemployment rate?

   Twelve machine learning models are employed to forecast the South African unemployment rate. The vector autoregression was used as a benchmark model. The selection of the twelve was through an exploratory process guided by past research.

2. The South African Reserve Bank tracks 147 key macroeconomic variables, these are used as features in this research. What is the impact of feature selection on the unemployment forecasting accuracy?

   Seventeen feature selection techniques are employed to assess the impact of using different feature subsets. Enabling the detection of a subset of features that aid the most in forecasting the unemployment rate.

3. What is the relationship between the various features of the unemployment rate and how do they influence the unemployment rate and each other?

Dynamic Bayesian networks are employed to show how the different features relate to each other. Do-Calculus is also performed to analyse different counterfactuals i.e. scenario analysis. These options were discovered through an exploratory process.

This chapter provided the methodology that was employed in this research as well as the data used and the data preprocessing that was undertaken through the research. Chapter 4, discusses the experimental results achieved in the research.

4 Experimental Results & Discussions

In order to address the research objective, 1522 experiments were conducted for this research. The aim being to i) forecast the South African unemployment rate using machine learning models, ii) identify features that are most impactful when forecasting the unemployment rate, and iii) determine how these features influence each other.

An exploratory research method was used, with twelve machine learning models, 17 feature selection techniques, and one Dynamic Bayesian network. This chapter discusses the findings of the research. The following url provides the source code which shows detailed implementation information:
https://drive.google.com/drive/folders/1XkEJBryoZd3rh-9VwRElw34sh6nJuwT1.

The chapter is structured as follows. In section 4.1 the MASE achieved by the benchmark model, VAR, is discussed. In section 4.2 the results of twelve machine learning models are discussed. The discussions are grouped according to the type of algorithm i.e. regression models, neural networks, decision trees, and kernel-based models. Section 4.3 shows how the VAR performance compares to the machine learning models. In section 4.4 the benefits of feature engineering when forecasting the unemployment rate are discussed. In section 4.5 the VAR results and machine learning results are compared. In section 4.6 the causal relationship of five most impactful features are discussed, showing how they relate to each other and the unemployment rate. The chapter closes with section 4.7 which discusses the contributions of this research.

4.1 Vector Autoregression: the Benchmark Model

The vector autoregression model (VAR) was used as a benchmark model for this research. This selection was made because the VAR is the orthodox multivariate traditional statistical technique that is used in unemployment forecasting. There are modified versions such as VARMA and SANVAR, however, these are used in limited instances in literature. Because

of the limited use of these, this research focused on the orthodox model: VAR.

The VAR implementation approach that was implemented by Stock and Watson [2001] was used for this research. As part of the data analysis that they recommend, two VAR tests were conducted: Granger-causality and stationarity tests.

Granger-causality test Core to the functioning of the VAR is that it assumes that each feature is equally influenced by the others in the model. The Granger-causality (GC) test assess this assumption by checking if the lag values of one feature can assist in predicting another variable.

The GC test tests for a null hypothesis that the lag values of one features do not influence another feature. Therefore, if the p-value is less than 0.05, the significance level, the null hypothesis can be rejected. Table 4.1 shows the p-value for a sample of 15 randomly selected features. The blue columns are predictors of the green rows. Therefore, the first row is influenced by column 2 and 11 as their p-values are below 0.05.
There is an even split in the number of features that influence others and those that do not. Therefore the Granger-causality holds in some cases.

Table 4.1: P-values for the Granger-causality for 15 variables. A p-value less than 0.05 implies that the blue features influence the green features.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	1.00	0.01	0.15	0.65	0.22	0.54	0.13	0.69	0.71	0.27	0.00	0.58	0.20	0.12
2	0.29	1.00	0.21	0.82	0.02	0.56	0.03	0.18	0.02	0.01	0.23	0.00	0.35	0.54
3	0.00	0.00	1.00	0.00	0.04	0.18	0.31	0.00	0.79	0.45	0.07	0.02	0.79	0.93
4	0.27	0.10	0.00	1.00	0.61	0.24	0.83	0.86	0.53	0.67	0.08	0.89	0.50	0.53
5	0.00	0.00	0.02	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.13	0.02	0.14	1.00	0.01	0.00	0.00	0.00	0.03	0.00	0.01	0.07
7	0.00	0.00	0.78	0.23	0.00	0.00	1.00	0.00	0.00	0.00	0.45	0.00	0.02	0.02
8	0.00	0.11	0.01	0.18	0.00	0.00	0.00	1.00	0.00	0.08	0.28	0.00	0.10	0.03
9	0.01	0.23	0.08	0.30	0.00	0.00	0.00	0.00	1.00	0.00	0.39	0.00	0.00	0.13
10	0.00	0.00	0.26	0.06	0.00	0.00	0.05	0.25	0.00	1.00	0.10	0.00	0.00	0.04
11	0.00	0.28	0.00	0.00	0.96	0.90	0.99	0.98	0.91	0.85	1.00	1.00	0.06	0.49
12	0.22	0.00	0.04	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.38	1.00	0.06	0.47
13	0.04	0.00	0.60	0.32	0.11	0.14	0.10	0.01	0.01	0.00	0.01	0.01	1.00	0.71
14	0.00	0.02	0.20	0.01	0.02	0.00	0.42	0.31	0.00	0.00	0.59	0.76	0.43	1.00

Legend:

1. Final consumption expenditure by general government
2. Domestic output: All groups
3. Change in inventories
4. Foreign exchange rate: SA rand per USA dollar
5. The net borrowing requirement of national government
6. Total outstanding domestic marketable bonds
7. Changes in the Exchequer and Paymaster-General Account balances with the South African Reserve Bank and other banks
8. Cash received from the issuance of Treasury bills less cash payments made to redeem Treasury bills
9. Total outstanding foreign marketable bonds by unexpired maturity: Exceeding 1 but not 3 years
10. Total outstanding foreign marketable bonds by unexpired maturity: Not exceeding 1 year Regression
11. Total outstanding domestic non-marketable loan levies
12. Net foreign assets after valuation adjustment
13. Value added at basic prices of mining and quarrying
14. Capital account (nsa)

Stationarity test The VAR model and its variants only work with data that is stationary. Each feature was tested for stationarity using the Augmented-Dickey Fuller method (AD-Fuller) whose null hypothesis is that the series is not stationary. If a feature was not stationary the entire dataset was differenced and tested again for stationarity, this process was repeated until the entire dataset was stationary. There was a total of eighteen different datasets (feature sets) i.e. the original dataset and feature sets with different feature selection strategies. Each of these was tested for stationarity and each one had nonstationary features. The feature sets were typically differenced twice to ensure that the data was stationary.

After these tests were conducted it was clear that the VAR model should be capable of modelling the data as the tests were all affirmative of this fact. Nine of seventeen datasets were passed through the VAR model: Pearson correlation coefficient with duplicates removed; mutual information gain; mutual information gain with Pearson correlation; ANOVA; recur-

sive feature elimination, with LASSO and ENET as the black box estimator; random forest; extreme gradient boosting (XGBoost); and principal components analysis (PCA).

Not all eighteen data sets could be used for the VAR model because VAR fundamentally estimates features by solving for the simultaneous equations using matrix algebra. To solve such equations it is required the matrix be invertible. This condition was only met by nine datasets. The other datasets did not work because they resulted in matrices that were singular or ones that were not positive definite, which is required when implementing the VAR model.

The VAR model centered all its predictions around 8.5. This suggests that the model assumes that the unemployment rate will revert to some mean. However, this is not the case because of the nonlinearity of the data: the unemployment rate itself is also nonlinear. Table 4.2 shows the mean absolute scaled error (MASE) that the VAR achieved on the different feature sets. The number of features in each model is also shown. The lag order was found through grid search and was set to 30 across the different experiments. Different VAR models were run with different lag orders and the one with the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) had a lag order of 30.

Table 4.2: The performance (MASE) of the VAR model with different feature sets over the test period of 24 months.

Model	**Feature Selection**	**Number of Features**	**MASE**
VAR	MIG	37	41.8194
VAR	Unique No Corr	54	44.2132
VAR	MIG No Corr	15	41.6990
VAR	ANOVA	37	41.0985
VAR	RFE LASSO	6	41.6118
VAR	RFE ENET	6	42.3765
VAR	XGBoost	7	43.5895
VAR	Random Forest	9	44.1892
VAR	PCA	51	50.5239

The MASE achieved by the VAR shows that the model is significantly worse than the naïve model when forecasting the South African unemployment rate. A MASE less than 1 is better than the naïve model and MASE above 1 is worse than the naïve model. The VARMA was also implemented and a similar MASE was achieved, it did not help in improving the forecasting accuracy of the VAR model. This is high because the South Africa unemployment rate is nonlinear and VAR model is best suited for modelling linear data.

4.2 Machine Learning Models

Some of contents that appear in this chapter have been published at IEEE and Springer conferences:

- *Mulaudzi and Ajoodha [2020a], "Application of Deep Learning to Forecast the South African Unemployment Rate: A Multivariate Approach," Seventh IEEE Computer Science and Data Engineering Conference, 2020.*
- *Mulaudzi and Ajoodha [2021a], "Demonstration that the Use of Feature Selection on High Dimensional South African Macroeconomic Data Results in Improved Performance with Lower Compute Requirements," Sixth International Congress on Information and Communication Technology Conference, ISBN: 2194-5357, 2021.*

Regression, neural network, decision tree, and kernel-based techniques were employed for this research. These techniques have been employed in past literature and have demonstrated that they generally improve the forecasting accuracy, as well as offer additional benefits such as feature importance determination. Various machine learning models were applied to eighteen different datasets: with and without feature selection methods. The hyperparameter settings are shown in Appendix B Table B.1. The source code with detailed implementation information for each machine learning model can be found on this link: https://drive.google.com/drive/folders/1XkEJBryoZd3rh-9VwRElw34sh6nJuwT1.

4.2.1 Regression

Five machine learning regression techniques were employed for this research: linear regression (ordinary least squared - OLS / LR), ridge regression (ridge), least absolute shrinkage and selection operator (LASSO), Elastic Net (ENET), and Bayesian ridge regression (Bayes ridge). OLS, LASSO, and ENET have been applied in previous similar research. The additional two, ridge and Bayes ridge, were added because these complement the other three: ridge uses L2-norm which is one of the two regularization techniques used by ENET and it is comparable to the L1-norm used by LASSO; the Bayes ridge uses the same regulization as ridge but tunes the L2-norm parameter to the data instead of tuning this by 'hand'.

Table 4.3 shows the top models across four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4). In H1, 33% of the models were better than the naïve model i.e. with a MASE below 1. This percentage increased to 57% in H2 but decreased to less than 1% in H3. This movement is due to the nonstationary nature of the South African unemployment rate. In H4, 87% of the models were above the naïve model: showing how the models over a longer horizon are able to capture the nonstationary nature of the South African unemployment rate. This observation is consistent with past research where ARIMA would generally be the top model in H1 but decline as machine learning models improve by H4.

Table 4.3: The performance (MASE) of five regression machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)				
Model	**Feature Selection**	**H1**	**H2**	**H3**	**H4**	**CV**
Bayes Ridge	MIG	1.6291	1.0824	1.2523	0.3846	0.4423
Bayes Ridge	EM ENET	0.3108	0.2546	1.0929	0.4422	0.3410
Bayes Ridge	VARIANCE UNIQUE	1.6836	0.9916	1.0082	0.5226	0.3863
Bayes Ridge	RFE ENET	0.8246	0.5103	0.5920	0.6192	0.7577
Bayes Ridge	EM LASSO	0.3201	0.3048	1.4005	0.7787	0.5713
ENET	MIG NO CORR	0.9204	0.8261	1.2535	0.4831	0.3770
ENET	NO CORR	0.9204	0.8261	1.2535	0.4831	0.3831
ENET	MIG	0.3537	0.2747	1.2919	0.5005	0.3635
ENET	RFE ENET	0.8086	0.6312	0.7500	0.5580	0.3856
LASSO	MIG NO CORR	0.9553	0.8386	1.2581	0.4694	0.4950
LASSO	NO CORR	0.9553	0.8386	1.2581	0.4694	0.4950
LASSO	UNIQUE NO CORR	0.9553	0.8386	1.2581	0.4694	0.4950
LASSO	VARIANCE UNIQUE NO CORR	0.9553	0.8386	1.2581	0.4694	0.4950
LASSO	MIG	0.4415	0.2990	1.2940	0.5153	0.4950
LASSO	RFE ENET	0.8264	0.6542	0.7757	0.5594	0.4950
LASSO	RFE LASSO	0.5130	0.4904	1.5839	0.5845	0.4950
LR	EM ENET	0.3097	0.2531	1.1031	0.4434	0.4123
LR	RFE ENET	0.8211	0.5086	0.5899	0.6189	0.2744
LR	VARIANCE UNIQUE NO CORR	1.1843	0.6598	0.7350	0.6933	7.80E+09
LR	UNIQUE NO CORR	1.1843	0.6598	0.7350	0.6933	7.75E+10
LR	NO CORR	1.1843	0.6598	0.7350	0.6933	1.38E+09
Ridge	EM ENET	0.4493	0.2887	1.0046	0.4523	0.3410
Ridge	RFE LASSO	0.4789	0.4983	1.0626	0.5949	0.3365
Ridge	RFE ENET	0.8828	0.5383	0.6269	0.6242	0.3819
Ridge	RANDOM FOREST	0.4591	0.4621	1.0995	0.7387	0.3450

Over the first horizon, OLS (LR in the Table 4.3) was the top performing model. This model sustained this performance across the horizons: being in the top five through the horizons. However, this model overfitted as the cross validation error was much larger than the various errors achieved through the horizons.

The other top models were ridge, Bayes ridge, ENET, and LASSO. These models address the overfitting of the OLS model by using regularization techniques. Bayes ridge and ridge regression consistently had amongst the lowest MASE over the four horizons without overfitting. These models achieved these results with ENET selected features: a multivariate

feature selection technique.

The shaded areas in the Table 4.3 are the lowest ten MASE results for each horizon. It is clear to see that the feature selection technique, EM ENET (Elastic Net embedded method), had the highest number of shaded values. The wrapper and filter methods were not as consistently top performing. Therefore, this shows that there is a handful of features that inform the South African unemployment rate. Furthermore, this subset of features includes features that are highly correlated with each other: they have group effects. This is evident because ENET feature selection takes into account group effects in its feature selection [Guyon and Elisseefl 2006].

For this research the ENET selected five features: Final consumption expenditure by general government; Domestic output: all groups; Foreign exchange rate: SA rand per USA dollar; Total outstanding domestic non-marketable loan levies; and Total foreign non-marketable debt, as the most impactful features when predicting the South African unemployment rate. Figure 4.1 shows the estimated coefficients.

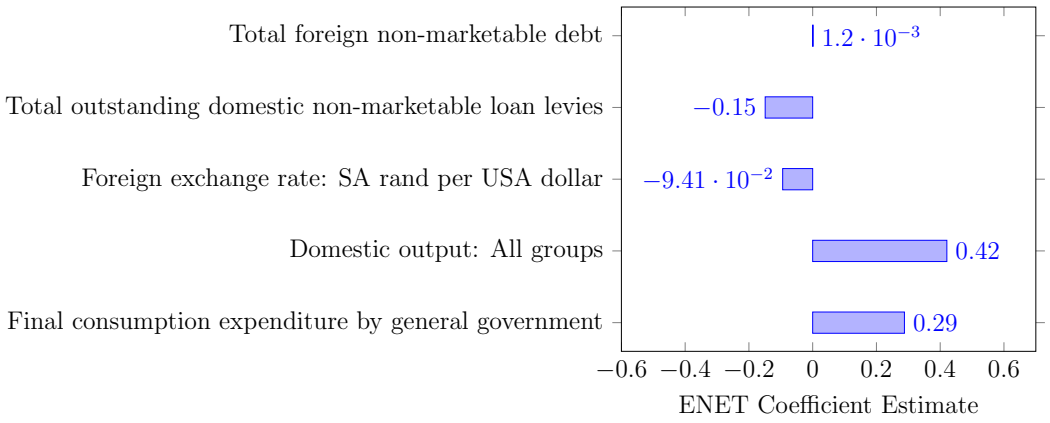


Figure 4.1: Features that were selected by the Elastic Net (ENET) embedded feature selection technique with their relative coefficient estimate.

The machine learning pipeline that used ENET for feature selection and then a machine learning model, such as Bayes ridge, benefits from the ENET regularization process. The process shrinks the number of features and ensures that the coefficient of each feature was not high, therefore, reducing the chances of overfitting. Bayes ridge benefited from this and further reduced the chances of overfitting by reducing the reliance of the model on any individual features.

It was also observed that the OLS was the worst performing model across all horizons: Appendix C, Table C.1 shows the full results. The performance of the model was improved

in cases where multivariate feature selection techniques were employed. This is because the OLS model is better suited for data that is linear and macroeconomic data is not typically linear. In this case the data was nonstationary and multi-modal thus not linear.

The OLS overfitted across horizons regardless of feature selection techniques employed. The removal of correlations had a positive impact on the MASE score and the cross validation error. Embedded (ENET and XGBoost) and wrapper (recursive feature elimination (RFE) with LASSO and ENET) feature selection techniques had the lowest MASE across the horizons: these methods reduced the number of features to between four to nine features and generally removed correlated features as well. Therefore, there was less noise for the OLS to overfit when these models were employed.

4.2.2 Neural Network

Four neural networks were deployed: multilayer perceptron (MLP), gated recurrent unit (GRU), long short-term memory (LSTM), and bidirectional LSTM (BiLSTM). The MLP was the only neural network that was not deep, it only had one layer with 43 nodes.

The LSTM, BiLSTM, and GRU all had four hidden layers. The models used the adaptive momentum optimizer, with mean squared error as the loss function, and the rectified linear unit as the activation function. In order to avoid overfitting, dropout was applied as well as early stopping. Ten percent of the data was also set apart for cross validation. Figure 4.2 shows the model loss over the training and test data.

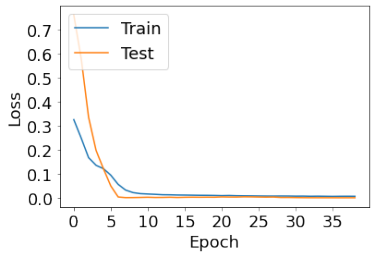
In most cases the model did not overfit and the error over the test data was lower than the error over the train data. The graphs follow a health trajectory, where they start high but decline over time. The GRU with PCA and LSTM or BiLSTM with RFE ridge are the only model where there is evidence of overfitting.

The LSTM had the lowest MASE over the longest horizon: H4. The LSTM generally performed above the naïve model when it was used on data that had the wrapper feature section method applied. The LSTM was in the top three deep learning models as shown in Table 4.5. The model is able to capture the nonstationary and autoregressive nature of the South African unemployment rate over the long run.

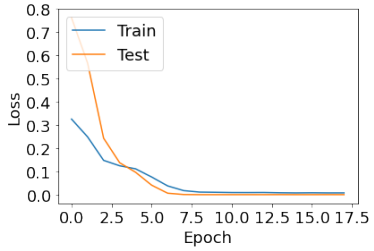
Because of the similarities between LSTM, GRU, and BiLSTM, these models are often compared. The LSTM is fundamentally a GRU with an additional gate and BiLSTM is an LSTM that is not just uni-directional. It is observed that the LSTMs were the best performing of the deep learning models. Therefore, it would not be worthwhile to employ the BiLSTM model. Although GRU is a simpler model, LSTMs are preferable in this case because of a lower MASE.

Figure 4.2: The loss (error) of the deep learning models through various epochs. The error is downward trending for both training and test data. The test is generally below the training error in most cases.

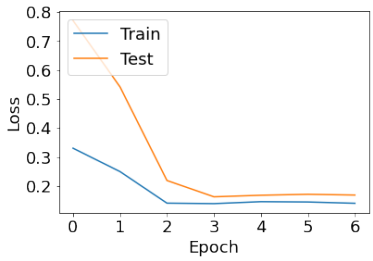
(a) LSTM on RFE ENET



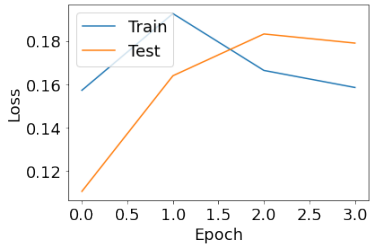
(b) LSTM on RFE LASSO



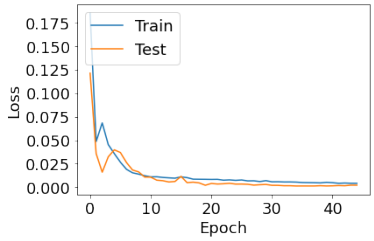
(c) LSTM on RFE Ridge



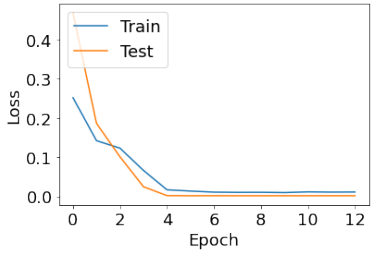
(d) GRU on PCA



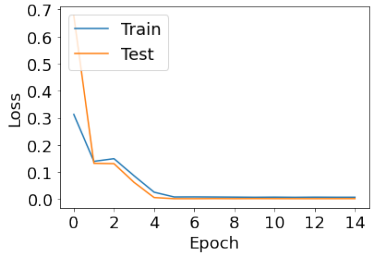
(e) GRU on Random Forests



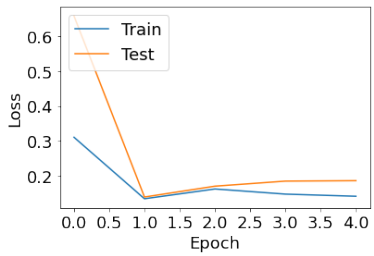
(f) GRU on RFE LASSO



(g) BiLSTM on RFE LASSO



(h) BiLSTM on RFE Ridge



The recursive feature elimination (RFE) with LASSO, ridge, and ENET had the lowest MASE across all four horizons. The features that each of the models used are shown in Table 4.4. These features resulted in the lowest MASE regardless of the neural network architecture used.

Table 4.4: Five features that were selected by the recursive feature elimination (RFE) wrapper feature selection technique.

LASSO	**ENet**	**Ridge**
Foreign exchange rate: SA rand per USA dollar	Foreign exchange rate: SA rand per USA dollar	Foreign assets
Total cash-flow revenue	Foreign assets: Total direct investment	Consolidated general government: Liabilities
Total outstanding domestic non-marketable loan levies	Total claims on the private sector	Consolidated general government: Liabilities: Domestic
Total South African population	Total outstanding domestic non-marketable loan levies	Consolidated general government: Liabilities: Foreign
The nominal effective exchange rate	The nominal effective exchange rate	Loans

Table 4.5 shows the MASE achieved over the four horizons by all four neural networks. The MLP was the top performing model across the first, second, and third horizon. The performance of the MLP was robust across different feature selection techniques. The wrapper method was top performing, followed by mutual information gain (MIG) and extreme gradient boosting (XGBoost).

Table 4.5: The performance (MASE) of four neural networks over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)			
Model	**Feature Selection**	**H1**	**H2**	**H3**	**H4**
LSTM	RFE LASSO	0.5244	0.4961	1.2502	0.5187
GRU	PCA	1.2858	1.1915	1.3790	0.5269
MLP	RFE ENET	0.3506	0.3357	0.3103	0.5593
MLP	MIG	0.4745	0.6219	1.3744	0.5680
MLP	RFE LASSO	0.6188	0.5982	1.1594	0.5895
GRU	RFE LASSO	0.7624	0.7359	1.1550	0.6091
BiLSTM	RFE RIDGE	1.8084	0.9594	1.1880	0.6101
LSTM	RFE RIDGE	1.8254	0.9712	1.1929	0.6137
BiLSTM	RFE LASSO	0.7704	0.7528	1.1957	0.6160
LSTM	RFE ENET	0.4532	0.5024	0.4684	0.6563
GRU	RANDOM FOREST	1.5507	1.8060	2.4502	0.6716
MLP	XGBOOST	1.3883	1.7957	2.4384	0.8079
GRU	VARIANCE UNIQUE	4.7377	3.0134	3.3968	0.9091
GRU	NO CORR	4.8002	3.0407	3.4419	0.9113
LSTM	Unique	4.7918	3.0524	3.4422	0.9124
LSTM	NO FS	4.8146	3.0665	3.4570	0.9146
BiLSTM	Unique	4.8402	3.0807	3.4696	0.9179
GRU	NO FS	4.8445	3.0848	3.4743	0.9187
LSTM	VARIANCE	4.8848	3.1171	3.5127	0.9207
GRU	UNIQUE NO CORR	4.8942	3.1155	3.5165	0.9221
BiLSTM	VARIANCE	5.0066	3.1962	3.5953	0.9329

The bottom performing models cut across the different deep learning models, except for the MLP. All three bottom performing models, LSTM, GRU, and BiLSTM, used either no feature selection or filter methods. This shows that the South African unemployment rate is predicted by features with group interactions. Selection based on statistical significance alone is insufficient because such an approach ignores group dynamics of features. Appendix D Table D.1 shows the full results for the neural networks.

Previous research such as Hall [2018] used simple neural networks, similar to the MLP implementation in this research. It was observed from this research that the simple MLP was able to outperform the deep models over three of the four horizons. This result differs from Cerqueira *et al.* [2019] who suggested that deeper and bigger is better. This paper finds that over 'very long' horizons, perhaps bigger is better, but over shorter horizons simple seems to be a better solution.

4.2.3 Decision Trees

Two decision trees were implemented: extreme gradient boosting (XGBoost) and Random forest (RF). These are very similar decision trees: both are ensemble models. The decision trees were all above the naïve model over the forth horizon. RFR performed above the XGBoost through all four horizons. Table 4.6 shows the MASE of these model.

Table 4.6: The performance (MASE) of two decision trees, extreme gradient boosting and random forest, models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)				
Model	**Feature Selection**	**H1**	**H2**	**H3**	**H4**	**CV**
RFR	RFE LASSO	0.4243	0.4859	0.3843	0.4828	0.1917
RFR	RFE ENET	0.4491	0.6190	0.4831	0.5629	0.1889
RFR	EM ENET	0.4683	0.7914	1.2908	0.5654	0.1683
XGB	RFE ENET	1.5102	0.8208	0.9413	0.5999	0.4662
RFR	MIG NO CORR	0.5481	1.2367	1.8036	0.6277	0.1831
XGB	RFE LASSO	1.7982	1.0217	1.1609	0.6309	0.4622
RFR	ANOVA	1.1570	1.2212	1.5935	0.6393	0.2049
RFR	EM LASSO	0.5893	1.1111	1.7803	0.6406	0.1544
XGB	EM ENET	1.8226	1.0635	1.2598	0.6460	0.4521
XGB	RANDOM FOREST	1.9640	1.2263	1.5530	0.6677	0.4487
XGB	ANOVA	1.9029	1.2378	1.5373	0.6814	0.4524
XGB	UNIQUE NO CORR	2.0054	1.2961	1.5582	0.6923	0.4662
XGB	EM LASSO	1.9607	1.2845	1.5534	0.6943	0.4534
RFR	NO CORR	0.5195	1.4603	2.1599	0.7090	0.2006
RFR	UNIQUE NO CORR	0.5391	1.4637	2.1643	0.7099	0.2041
RFR	VARIANCE UNIQUE NO CORR	0.5279	1.4650	2.1700	0.7144	0.1980

It is normally expected that XGBoost will produce better accuracy than RFR. However, this is not the case for this research. The XGBoost was tuned through a grid search approach: the loss function was the squared loss, columns were sampled by tree and were set to a 0.3 ratio, learning rate was 0.1, maximum tree depth was 6, L1 regularization parameter was set to 45, and the number of estimators was 4. These hyperparameters were selected through grid search with trial and error fine tuning: Appendix B Table B.1 shows all the hyperparameters for the decision trees and all the machine learning models.

The XGBoost is a specialised ensemble of decision trees that improves regression accuracy by combining multiple trees [Krauss *et al.* 2016]. RFR works in a similar fashion but uses a random subset of features to build its multiple trees [Krauss *et al.* 2016]. This subset selection enables the model to avoid noise as it uses ensemble trees that are not highly correlated

[Krauss *et al.* 2016]. The XGBoost on the other hand can overfit when the data is high dimensional and full of noise.

The tree models are able to score the importance of different features. Figure 4.3 shows the features that were highlighted by RFR as important features in forecasting the South African unemployment rate.

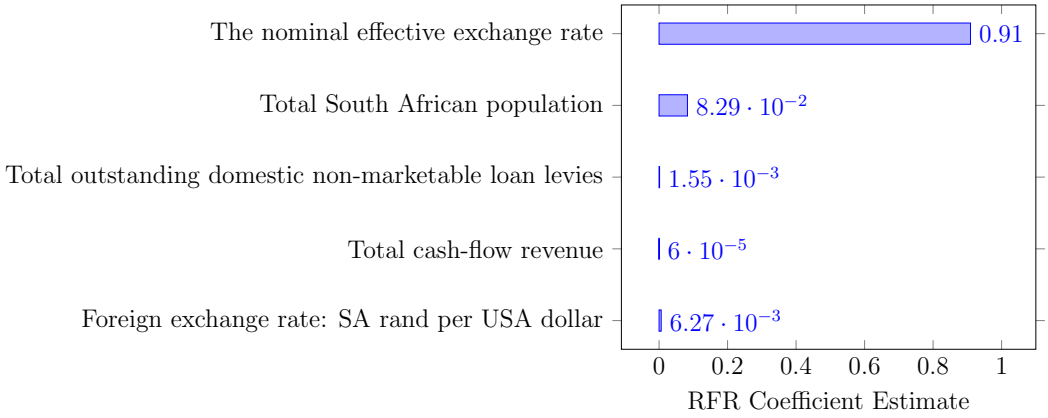


Figure 4.3: Features that were selected by the Random Forest (RFR) embedded feature selection technique with their relative coefficient estimate.

It is also worth noting that both the RFR and XGBoost had the lowest MASE when combined with the wrapper feature elimination techniques. This confirms that indeed only a subset of features informs the South African unemployment rate. Appendix E Table E.1 shows the full results for the decision tree methods.

4.2.4 Kernel

Support vector regression (SVR) machine learning kernel techniques were employed for this research. The model had MASE significantly above 1 when the filter feature selection techniques were used. This was particularly true when correlated features were removed. This is because the model relies on being able to linearly separate the features along a plane. When the correlated features are removed placing the kernel becomes harder, hence, the MASE was above 1. The model performed the best when the wrapper feature selection techniques were applied as well as filter methods that did not remove correlated features.

The hyperparameters for the SVR model were tuned through grid search. The optimal hyperparamaters being those with the lowest error. The following proved to be most effective, the regularization was set to 1.37, ε was set to 0.02, and a linear kernel was used.

The kernel is used to determine a higher dimension that the feature set should be translated to in order to create a decision boundary. The linear kernel keeps the data in the same

dimension. Therefore, the unemployment rate is linearly separable when some of the feature selections are used such as ANOVA, EM ENET, or RFE.

Table 4.7: The performance (MASE) of the support vector regression model over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)				
Model	**Feature Selection**	**H1**	**H2**	**H3**	**H4**	**CV**
SVR	MIG	1.5727	0.9904	1.0405	0.3229	0.3997
SVR	ANOVA	0.8342	0.5405	0.5359	0.3692	0.2714
SVR	EM ENET	0.3556	0.2930	1.1201	0.4487	0.3343
SVR	RFE LASSO	0.5875	0.6166	1.1709	0.6189	0.3021
SVR	RFE ENET	0.7256	0.4374	0.5007	0.6210	0.2677
SVR	Unique	1.1363	0.8091	1.0208	0.6505	0.4028
SVR	VARIANCE UNIQUE	1.1363	0.8091	1.0208	0.6505	0.4028
SVR	NO FS	1.1499	0.8149	1.0261	0.6529	0.4023
SVR	VARIANCE	1.1499	0.8149	1.0261	0.6529	0.4023
SVR	RANDOM FOREST	0.5868	0.6413	1.2138	0.7007	0.5856
SVR	EM LASSO	0.2938	0.3420	1.3585	0.7031	0.3209
SVR	PCA	2.8183	1.9450	2.3707	0.7179	0.2782
SVR	XGBOOST	0.4659	0.9118	1.4016	0.8843	0.4521
SVR	RFE RIDGE	4.6596	2.9485	3.9336	1.6363	0.6405
SVR	NO CORR	5.8350	4.4346	4.7082	1.7819	0.3463
SVR	UNIQUE NO CORR	5.8350	4.4346	4.7082	1.7819	0.3463
SVR	VARIANCE UNIQUE NO CORR	5.8350	4.4346	4.7082	1.7819	0.3463
SVR	MIG NO CORR	3.1131	2.1279	2.9351	1.7887	0.3259

4.3 Overall Performance Summary

The MLP was the top performing overall model: measured by the average MASE over the four horizons. MLP had an average MASE of 0.3990, which was 14% higher than the model that followed it: the RFR. Compared to deep learning models, the model did not achieve an error as low their H4 error, but was higher in the first three horizons.

Table 4.8 shows the top ten models. The different machine learning model categories - regression, deep learning, kernel, and decision trees - are all represented in the top performing models. Although, the OLS (LR) is in the top ten, it overfits and should not be considered in the top ten. Hence, eleven models are shown.

The SVR is represented more than the other models in the top ten. Particularly when

it is paired with ENET feature selected data. This implies that the grouping effect of the ENET results in a feature set that is linear separable.

Table 4.8: The top ten performing (MASE) machine learning models across all different model categories. The performance was assessed using an average MASE over the four horizons).

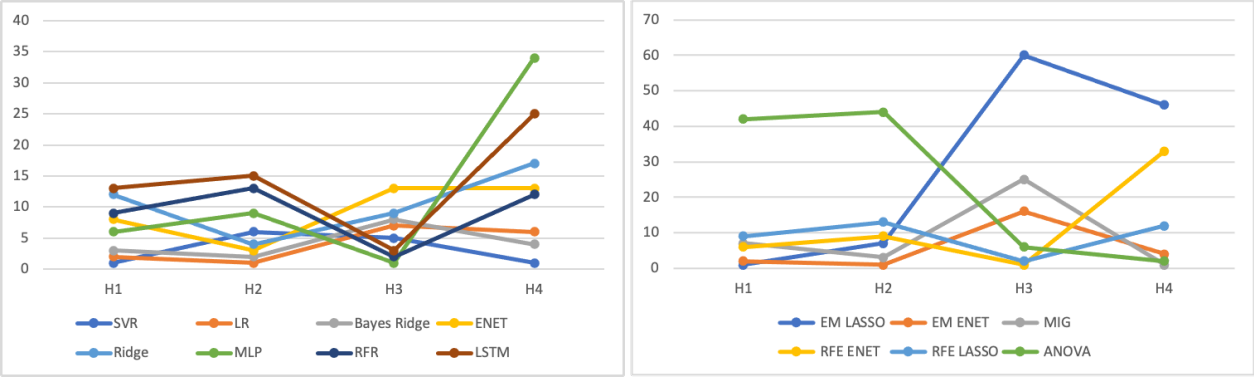
		Mean Absolute Scaled Error (MASE)				
Model	**Feature Selection**	**H1**	**H2**	**H3**	**H4**	**AVG**
MLP	RFE ENET	0.3506	0.3357	0.3103	0.5593	0.3890
RFR	RFE LASSO	0.4243	0.4859	0.3843	0.4828	0.4443
LSTM	RFE ENET	0.4532	0.5024	0.4684	0.6563	0.5201
Bayes Ridge	EM ENET	0.3108	0.2546	1.0929	0.4422	0.5251
LR	EM ENET	0.3097	0.2531	1.1031	0.4434	0.5273
RFR	RFE ENET	0.4491	0.6190	0.4831	0.5629	0.5285
Ridge	EM ENET	0.4493	0.2887	1.0046	0.4523	0.5487
SVR	EM ENET	0.3556	0.2930	1.1201	0.4487	0.5543
SVR	ANOVA	0.8342	0.5405	0.5359	0.3692	0.5700
SVR	RFE ENET	0.7256	0.4374	0.5007	0.6210	0.5712
ENET	RFE ENET	0.3537	0.2747	1.2919	0.5005	0.6052

It is worth noting that only one of the top feature sets is not a multivariate feature selection method: the SVR ANOVA. The rest of the models are multivariate and all used either ENET and/or LASSO as the estimator for the selection of the feature subsets. Both embedded (ENET and LASSO) and wrapper (recursive feature elimination) methods are almost equally represented.

The inclusion of the ANOVA in the top models is interesting because although ANOVA is univariate it has multivariate proprieties because it estimates how the variation in a feature compares to the variation amongst all other features. It, therefore, considers the variation across the entire feature set.

Figure 4.4 (a) shows the ranking of the models through the various horizons. Whilst (b) shows the ranking of different feature selection techniques. These show how multivariate feature techniques are generally consistent in their performance through horizons whilst ANOVA is not.

Figure 4.4: The performance ranking of the top machine learning models, as well as, the ranking of feature selection techniques.



(a) Machine Learning ranking over four horizons.

(b) Feature selection techniques

The subset of features that influence the South African unemployment rate are discussed in subsection 4.2.1 and subsection 4.2.2. Fundamentally, there are distinct key features that influence the South African unemployment rate. These features are identified by ENET and LASSO, the top performing feature selection techniques, as key features that impact the South African unemployment rate:

1. Foreign exchange rate: SA rand per USA dollar
2. Total outstanding domestic non-marketable loan levies
3. Final consumption expenditure by general government
4. Total outstanding domestic non-marketable loan levies
5. Domestic output: all groups

It was further observed that the model that achieved the highest $R-squared$ was the SVR model with an R-squared of 0.6705 and Bayes ridge with an R-squared of 0.5821. Table 4.9 shows the R-squared for the top performing models. Only three models achieved a R-squared above 0.5. The R-squared being below 0.5 was a cause for concern because it signaled that the models were only able to explain less than 50% variance in the data.

R-squared measures how much variation in the data the model is able to explain relative to a naïve model i.e. a model with just the intercept [Hyndman and Athanasopoulos 2018]. Feature engineering was considered as a mechanism to improve this: discussed in section 4.4. Although, R-squared was not the primary metric for this research, it is wise to also assess it as it provides additional information not captured by MASE.

Table 4.9: The top ten models as measured by R-squared.

Model	Feature Selection	R squared
SVR	ANOVA	0.6705
SVR	MIG	0.6700
Bayes Ridge	MIG	0.5821
Ridge	ANOVA	0.4536
ENET	PCA	0.4411
LASSO	PCA	0.4310
Ridge	MIG	0.4118
Ridge	EM ENET	0.3679
Bayes Ridge	Variance	0.3438
Bayes Ridge	NO FS	0.3438

4.4 Feature Engineering

Some of the contents that appear in this chapter have been published at a Taylor & Francis conference:

- *Mulaudzi and Ajoodha [2021b], "Improving the Performance of Multivariate Forecasting Models Through Feature Engineering: A South African Unemployment Rate Forecasting Case Study," 2021 International Conference On Interdisciplinary Research in Technology and Management, 2021.*

In order to improve the R-squared, lag variables of the unemployment rate were added as features to forecasting the South African unemployment rate. Two lag variables were added as this was found as the most optimal (lowest error) number of lag variables. Therefore, the number of features increased from 147 to 149.

The number of lags were selected by trial and error. Lags between 2 and 7 were tried. These lags were treated as a parameter that was set and updated based on the performance of the model. Two lags were the best performing across the models and, therefore, this was used across the models. Figure 4.5 shows that the unemployment rate has a high correlation with its lag values. Demonstrating that there is evidence of temporal dependencies in the South African unemployment rate.

The addition of lag variables improved the R-squared by over 100% on average across all the experiments. The highest R-squared without the lag values was 0.6705 and with lag values it was 0.7875. Only 87 experiments had a positive R-squared without addition of lags, this improved to 132 after the lags were added. Furthermore, 73 experiments achieved an R-squared above 0.5, an improvement from 3. Table 4.10 shows the top ten R-squared changes.

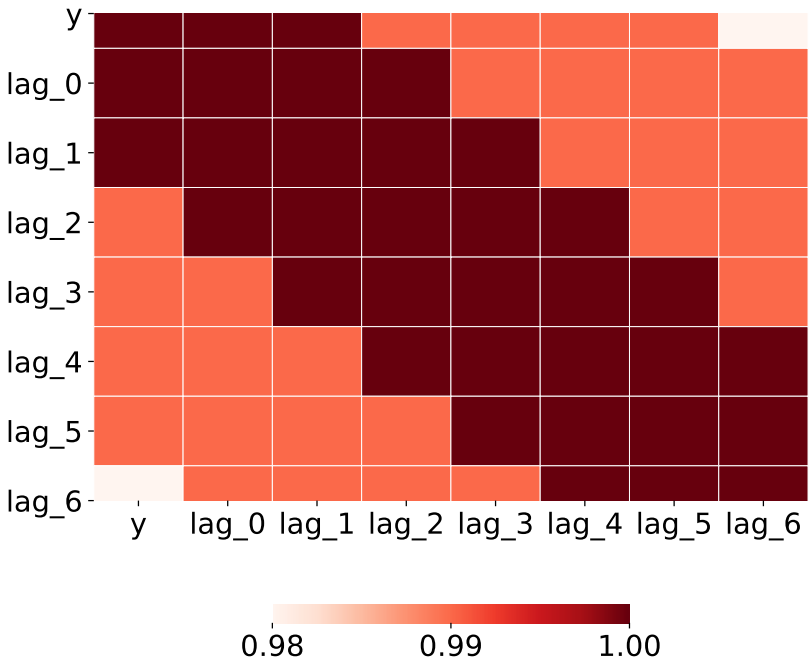


Figure 4.5: The Correlation Map for the lag values of the unemployment rate showing that the unemployment rate is highly correlated with the lag values.

Table 4.10: The inclusion of lag values of the South African unemployment rate improves the R-squared by over 100% on average.

		Without Lags	**With Lags**	
Model	**Feature Selection**	**R-squared**	**R-squared**	**% Change**
Ridge	MIG	0.4118	0.7875	91.2%
ENET	MIG	0.0171	0.7875	4510.8%
ENET	PCA	0.4411	0.7872	78.5%
LASSO	XGBoost	0.3097	0.7872	154.2%
LASSO	PCA	0.4310	0.7849	82.1%
LASSO	MIG	0.0511	0.7849	1437.0%
ENET	XGBoost	0.2911	0.7842	169.4%
LASSO	Random Forest	0.2244	0.7842	249.5%
RFR	RFE LASSO	0.0092	0.7750	8285.6%
ENET	Unique	0.1513	0.7731	410.9%
Bayes Ridge	NO FS	0.3438	0.7649	122.5%

The inclusion of lag values improves the R-squared of the models significantly. Some models, which had negative R-squared, improved by values much larger than 100%. An example of this the MLP which with RFE ENET, the model achieved the top MASE score when lags were not included but the R-squared was -0.2127. After lags the R-squared improved to 0.7273: a 439% improvement.

The accuracy (reduction in MASE) also improved slighly. This was measured using the average over the four horizons. The average MASE was 0.2950 without lag values and was 0.2832 after the lag values were added. Therefore, the inclusion of lag values not only improved the percentage of variation that the model can explain, but it also improves the error achieved by the model.

After the inclusion of lag values, the deep learning models were all in the top models as measured by the average MASE over the the four horizons. The LSTM, BiLSTM, and GRU improved their MASE by more than 20%. This improvement in the recurrent neural network models is because these models have a long term memory structure, therefore, introducing lag variables enables these models to take advantage of this capability. The results are shown in Table 4.11, which shows the change in order amongst the top models.

Table 4.11: The comparison of how the inclusion of lag values changes the performance (MASE) of the top models.

Without Lags				With Lags		
Model	**Feature Selection**	**AVG**		**Model**	**Feature Selection**	**AVG**
MLP	RFE ENET	0.3890		SVR	ANOVA	0.3827
RFR	RFE LASSO	0.4443		MLP	RFE ENET	0.3879
LSTM	RFE ENET	0.5201		GRU	RFE ENET	0.3894
Bayes Ridge	EM ENET	0.5251		LSTM	RFE ENET	0.4067
LR	EM ENET	0.5273		BiLSTM	EM LASSO	0.4101
RFR	RFE ENET	0.5285		MLP	ANOVA	0.4125
Ridge	EM ENET	0.5487		Ridge	VARIANCE	0.4176
SVR	EM ENET	0.5543		Ridge	NO FS	0.4176
SVR	ANOVA	0.5700		Ridge	VARIANCE UNIQUE	0.4179
SVR	RFE ENET	0.5712		Ridge	Unique	0.4179
ENET	MIG	0.6052		BiLSTM	RFE ENET	0.4324

The overall top model after the inclusion of the lags was the SVR. Showing that inclusion of lags also improves the linear separability of the dataset selected through ANOVA. The model was eighth without lags. The MLP moved from first place to second after lags were added. The addition of lags had minimal impact on the MLP with RFE ENET feature selection.

This shows the robustness of the MLP with RFE ENET: this result increases confidence in the model.

It is also worth noting that ANOVA is represented twice in the table with lags: one with SVR and another with MLP. The recursive feature selection with ENET as the black-box model remains well represented in the top models.

Table 4.11 shows that 50% of the top models were all multivariate feature selection techniques. Most of these multivariate models used ENET as an estimator. This suggests that the group effect should be considered or retained regardless of the inclusion of exclusion of lags.

The increase in the number of univariate feature selection techniques is interesting and shows that when the lag values are added the variation within the feature has a big impact on the ability of the model to forecast the unemployment rate. The removal of low variance features improved the accuracy, this can be seen by observing that four out of the six univariate features in the top models are concerned with variance: variance thresholds removals, the ratio of the variance within features and between features, or removal of low variance features.

Feature engineering improved the performance (reduced MASE and increase R-squared) of most machine learning models because of the temporal nature of the South African unemployment rate. The improvements were across all the different kinds of machine learning models used. Models with memory structures were able to improve a lot more than models without such structures. Hence, the deep learning models all improved because of the temporal nature of the recurrent neural networks.

The inclusion of lag values improves the confidence in the results because the inclusion of lags not only results in improved forecasting accuracy, but it also improves the percentage of variation in the data that the models can explain.

4.5 Machine Learning vs Vector Autoregression

The VAR was used as a benchmark model, Table 4.2 shows that the VAR performed poorly compared to a naïve model. The application of VAR moving average (VARMA) did not improve this. Grid search was used to find the parameters of the model i.e. lag order. Differencing was also required in order to make the data stationary.

The VAR model centered all its predictions around 8.5; this suggests that the model assumes that unemployment reverts to some mean. However, this is not the case because of the nonlinearity of the data: the unemployment rate itself is also nonlinear. The MASE of the VAR model was over 100 times higher than the machine learning models: VAR was 40 times worse than the naïve models.

As the VAR model was the benchmark model, the results of the model were compared to the machine learning models to explore how the performance differed. Figure 4.6 shows the difference between a number of the machine learning models and VAR. It is clear that the machine learning models outperform the benchmark model. None of the VAR predictions were below naïve model, whilst a number of machine learning models were.

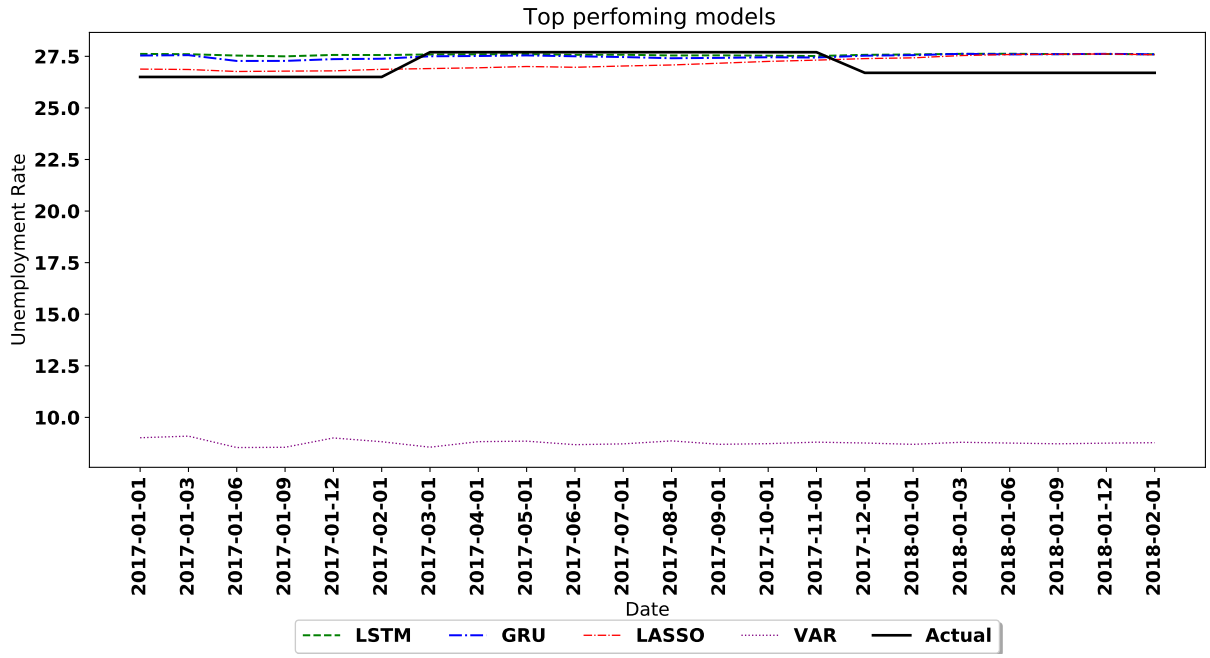


Figure 4.6: The comparison of the VAR model with a few machine learning models to show the difference in performance.

The VAR also has limitations because it assumes all features equally impact each other. Therefore, feature selection is not possible with the VAR. The machine learning models such as ENET and LASSO are able to perform automatic feature importance determination. This is an additional benefit of machine learning models beyond just accuracy improvements.

The machine learning models are also geared towards big data settings, and can cope with a large number of features and observations. The VAR performs better with a limited number of features, making it less suitable for big data contexts.

4.6 The Causal Relationship of the Features

The multivariate feature selection techniques that used ENET as an estimator resulted in the lowest MASE across horizons and models. To determine the causal relationship between the features that ENET identified as most impactful, a Bayesian network was constructed with these features as nodes. The theoretical treatment of Bayesian networks is out of scope for this research but readers are encouraged to consult Pearl [2011] and Koller and Friedman [2009].

A Bayesian network is a directed acyclic graph that shows how nodes (features) influence each other [Ajoodha 2018; Koller and Friedman 2009]. Therefore, in these graphs a node represents a feature and an edge that connects two nodes represents the influence of one node over another [Ajoodha 2018; Koller and Friedman 2009]. These networks provide us an avenue to understand the causal relationship between different features that influence the South African unemployment rate.

A causal network with six nodes was constructed. The nodes being Final consumption expenditure by general government (FGE), domestic output: all groups (GDP), foreign exchange rate: SA rand per USA dollar (FER), total outstanding domestic non-marketable loan levies (TDL), total foreign non-marketable debt (TFD), and the unemployment rate (UR). The Bayesian network structure was learnt from data using the NOTEARS algorithm as implemented by CausalNex [Pamfil *et al.* 2020; Zheng *et al.* 2018; QuantumBlack 2020].

Figure 4.7 shows how features (nodes) that impact the South African unemployment rate influence each other. It is clear to see that the South African unemployment rate is directly influenced by TDL, GDP, and FGE. TFD and FER influence it indirectly. The features themselves also influence each other such as the impact of TFD on TGE: the total foreign non-marketable debt impacts the final consumption by general government. Simply put, the amount of debt that the South African government has impacts how much the government spends.

Figure 4.7: A Bayesian network showing how key features that impact the South African unemployment rate influence each other.

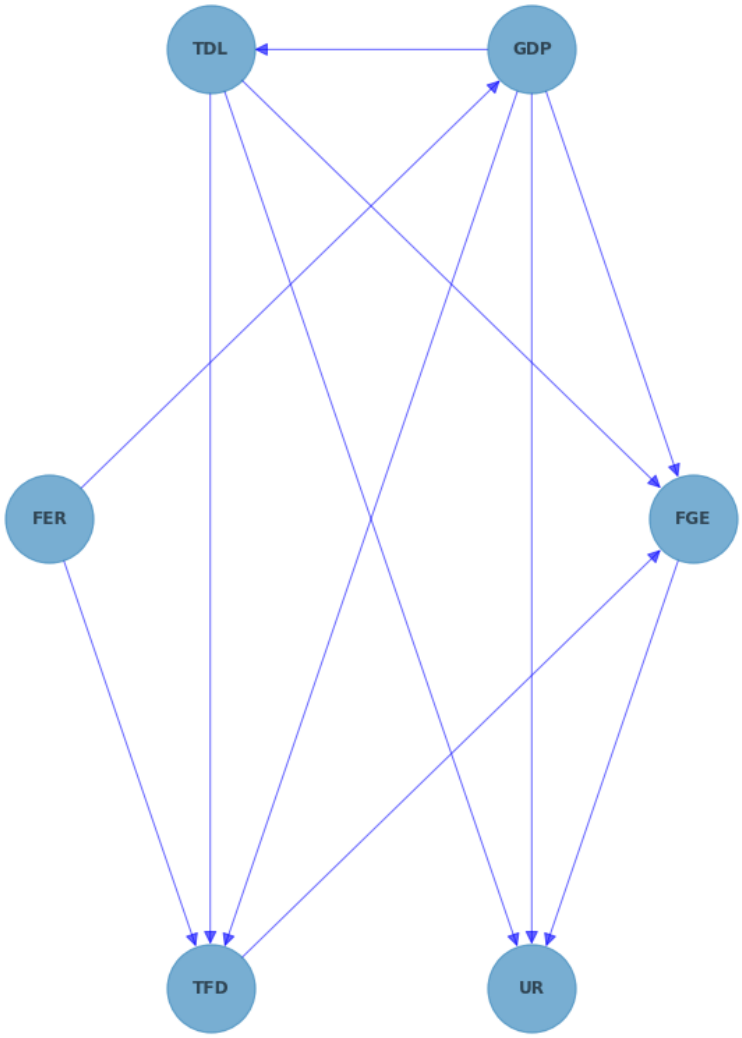


Figure 4.8 shows the dynamic version of the directed graph: a Dynamic Bayesian network (DBN). This shows that not only do the features themselves impact the unemployment rate and each other, the lag values of these have an impact too. The figure only shows one lag order for each variable, however, the second lag order also had an impact as well.

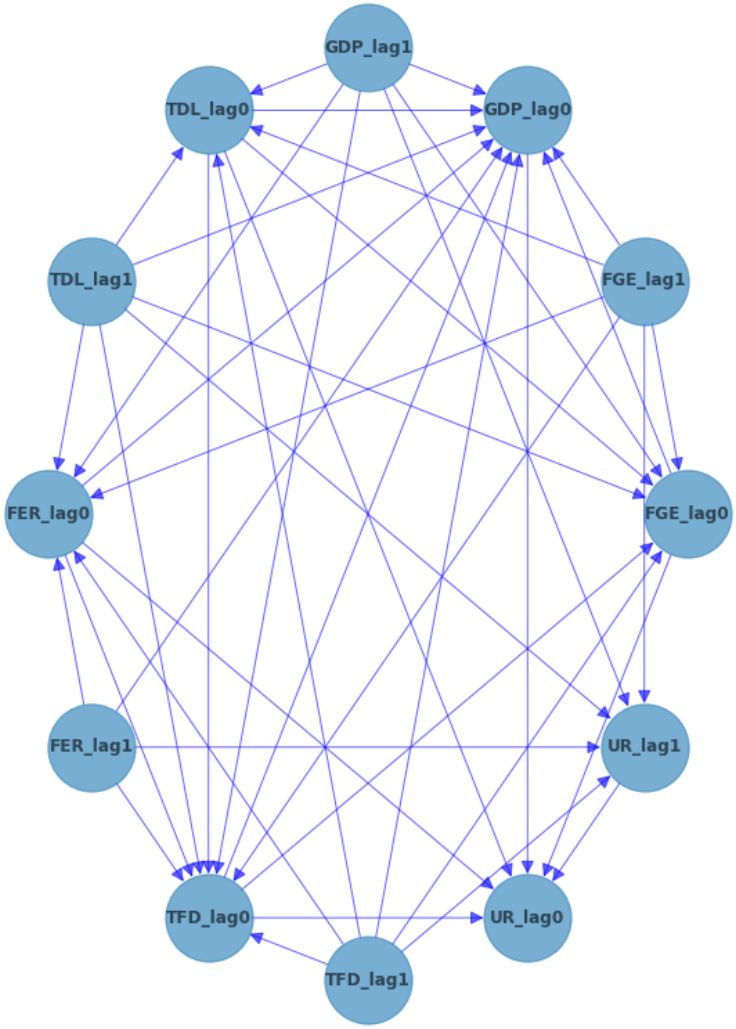
The conditional probability distributions (CPDs) have been omitted from the paper because of the size of the DBN. The source code with detailed implementation information (including CPDs) for the DBN can be found on this link: https://drive.google.com/drive/folders/1XkEJBryoZd3rh-9VwRElw34sh6nJuwT1.

In Figure 4.8 the current unemployment rate (UR_lag0) is shown as being impacted by the unemployment rate lag (UR_lag1). This confirms the improved MASE and R-squared

that was observed when lag orders of the unemployment rate were included in the data (see Table 4.3).

The lag values of other features indirectly influence the current unemployment rate. They do this by influencing their own current values (lag_0), as well as the lag value of the unemployment rate: which then influences the unemployment rate directly.

Figure 4.8: A Dynamic Bayesian network (DBN) showing how key features that impact the South African unemployment rate influence each other over time: with only a single lag value for each node.



In order to develop the causal network (DBN), the approach proposed by QuantumBlack [2020] was followed. The afore mentioned features were firstly discretized. Meaning that they were converted from continuous to discrete data. This was done using the percentage change. The percentage change was calculated by looking at the difference between the value

of the series at time t and $t-1$ for every observation. These percentage changes were then discretized by assigning each observation a value based on the change: for positive change, 1 was assigned, whilst negative changes were assigned -1, and no change was assigned 0.

The continuous time series data was used to learn the Dynamic Bayesian network (DBN) from the data using the NOTEARS algorithm. The discretized data is used to learn the states that each node can take (-1, 0, or 1, shown in Table 4.12), as well as the conditional probability distributions (CPDs) for each node. To achieve this the data was split into training and testing: 715 observations for training and 79 (10%) observations for testing. The training data was used to learn the CPDs of the DBN using a Bayesian estimator with the K2 algorithm to learn the priors. QuantumBlack [2020] provides background to this approach.

Table 4.12: The percentage distribution of the various states that the different nodes can have.

	UR	**GDP**	**FGE**	**TFD**	**TDL**	**FER**
0	78%	57%	75%	37%	75%	37%
1	13%	36%	18%	36%	15%	34%
-1	9%	6%	6%	27%	9%	30%

The learned structure with states and CPDs was tested using the testing data. Classification metrics were used to evaluate the overall strength of the DBN. The accuracy, which measures the ability of the model to correctly make predictions, was 72%.

Table 4.13 shows the metrics across the unemployment rate (UR) states; which was set as the prediction target. The results are reflective of the data distribution, there were few examples of state -1, which is a decline in the unemployment rate: the South African unemployment is upward trending. Furthermore, state 1 was also not as frequently observed from 1960 to today, the rapid rise in unemployment rate has been an issue over the last twenty years.

Table 4.13: Summary of the classification metrics used to evaluate the DBN.

	UR (State: -1)	**UR (State: 0)**	**UR (State: 1)**	**Weighted Average**
Precision	11%	80%	0%	65%
Recall	17%	88%	0%	72%
F1-Score	13%	84%	0%	68%

There is additional work that is required in order to address the sample imbalance: this is left to future researchers. However, the evaluation results were satisfactory for this research as they were consistent with the observed data. They were further consistent with Table 4.12. Various techniques such as varying the train and test sizes, reducing the size of the data,

increasing and decreasing the number of states, as well as different discretization approaches were attempted. The results shown in Table 4.13 were the best results achieved.

The DBN was then used to make inference and explore different scenarios that could impact the South African unemployment rate. The marginal likelihood of the unemployment rate (UR) was:

$$P(UR = -1) = 0.28, \tag{4.1}$$
$$P(UR = 0) = 0.43, \ and \tag{4.2}$$
$$P(UR = 1) = 0.30. \tag{4.3}$$

It is observed from the marginals that the unemployment rate is likely to remain as is and if it changes it is more likely to increase than decrease.

The conditional probabilities were also calculated. These are shown in Table 4.14. The probabilities show that the South African unemployment rate is impacted more or less equally by the other features. The GDP is one feature that stands out as it has the highest conditional probability: $P(UR = 0|GDP) = 0.3968$. This means that a change in the GDP is mostly like to keep the unemployment rate the same. Similarly, a change in FER is most likely to lead to a reduction in the unemployment rate, and a change FGE is most likely to increase the unemployment rate.

In Table 4.14 each value in the table shows the Probability (P) of the unemployment rate (UR) being in a particular state (1: increase, 0: remain the same, and -1: decrease) given that one of the nodes has been observed. For examples 0.3059, the first entry in the table, is the $P(UR = -1|GDP)$. The probability of the unemployment rate decreasing (UR in state -1) given that we observe a particular change in the GDP.

Table 4.14: The conditional probabilities of the South African unemployment rate given that the other nodes have been observed.

$P(UR\|GDP)$	**UR (State: -1)**	**UR (State: 0)**	**UR (State: 1)**
GDP	0.3059	0.3968	0.2973
FGE	0.3333	0.3333	0.3333
TFD	0.3153	0.3627	0.3219
TDL	0.3153	0.3662	0.3185
FER	0.3242	0.3548	0.3210

The conditional dependence shows that certain features have a higher impact than others in determining the state of the unemployment rate. Do-Calculus, as explained by Pearl [2011], was used to explored how different changes or intervention to the features could impact the distribution of South African unemployment rate. Thus do-Calculus allows us to perform scenario analysis based on the DBN, by exploring the counterfactual i.e. 'what if node x

changed'.

Table 4.15 shows the results of each change in the feature by considering a counterfactual: what if the feature was increased (set to state 1), decreased (state -1), or remained the same (state 0). It is important to the note that the below discussion, refers to changes in distributions not the actual unemployment rate.

Increasing the GDP will mostly likely to lead to a decrease in the unemployment rate. However, this also results in an increase in the number of those unemployed, although smaller: a net decrease of 5%. This is visible by studying the change in the probability distributions. Concretely, decreasing the GDP results in the likelihood of the unemployment rate increasing by 16%. This shows the asymmetric relationship between GDP and the unemployment rate: in absolute terms, decreasing GDP impacts the unemployment rate more than increasing GDP.

The asymmetric relationship between the South African GDP and South African unemployment rate is discussed in chapter 1. The do-Calculus confirms this observation, demonstrating that the DBN is consistent with the properties observed in the data.

Table 4.15 also shows that if the TDL increases, then there is a high probability of a decrease in the unemployment rate: the likelihood decreasing by 6%. Therefore, changes to the TDL is a lever that policymakers can leverage to improve the unemployment rate. Similar outcomes can be achieved by reduced FGE as well.

Table 4.15: Counter factuals based on the Dynamic Bayesian model, showing how the unemployment rate would change if its features are changed.

	UR (State: -1)		UR (State: 0)		UR (State: 1)	
Before intervention	**0.2770**		**0.4260**		**0.2960**	
Intervention	**Distribution**	**% Change**	**Distribution**	**% Change**	**Distribution**	**% Change**
What if GDP decreased?	0.3299	19%	0.3658	-14%	0.3043	3%
What if GDP increased?	0.2517	-9%	0.4645	9%	0.2838	-4%
What if GDP stayed the same?	0.2932	6%	0.3876	-9%	0.3191	8%
What if FGE decreased?	0.3293	19%	0.3414	-20%	0.3293	11%
What if FGE increased?	0.3283	19%	0.3435	-19%	0.3283	11%
What if FGE stayed the same?	0.2611	-6%	0.4532	6%	0.2857	-3%
What if TDL decreased decreased?	0.3196	15%	0.3557	-17%	0.3246	10%
What if TDL increased?	0.3153	14%	0.3662	-14%	0.3185	8%
What if TDL stayed the same?	0.2705	-2%	0.4384	3%	0.2911	-2%
What if FER decreased?	0.2783	0%	0.4287	1%	0.293	-1%
What if FER increased?	0.2726	-2%	0.4326	2%	0.2948	0%
What if FER stayed the same?	0.2988	8%	0.39315	-8%	0.308	4%

From the initial 147 feature list, there is a subset of features that most impact the South African unemployment rate. This feature subset was determined by ENET, and is made up of features that influence each other: the grouping effect of ENET. This influence was modelled as a DBN.

Using the DBN, it was demonstrated in this section, how the South African unemployment rate responds to changes in other features. Therefore, machine learning approaches not only enable higher accuracy (as measured by MASE) to be achieved, but they also allow for features to be selected based on their predictive capacity, as well as enable the automatic determination of the influence of said features on each other.

4.7 Research Contributions

This chapter demonstrated the value of machine learning models to forecast the South African unemployment rate. It was shown that machine learning models have a higher forecasting accuracy when compared to VAR, the benchmark model. Furthermore, offering the benefit of being able to select features that influence the South African unemployment rate. Therefore, enabling policymakers to focus on the right levers to decrease the unemployment rate. These levers are made even more precise through the DBNs which show how they influence each other, as well as how they directly and in-directly influence the South African unemployment rate.

The research questions have, therefore, been addressed by conducting 1522 experiments. Concretely,

1. **Research Question 1** Compared to the vector autoregression model (VAR), the benchmark multivariate forecasting model, how effective are machine learning models (regression, neural networks, and kernel methods) in forecasting the South African unemployment rate?

   Twelve machine learning models were employed to forecast the South African unemployment rate with the vector autoregression (VAR) used as a benchmark model. The machine learning models were 100 times better than the VAR: as measured by MASE. Most of the models achieved a MASE below 1: machine learning models are better than naïve models. MLP and SVR are generally the top performing models, with the lowest average MASE over four horizons.

2. **Research Question 2** The South African Reserve Bank tracks 147 key macroeconomic variables, these are used as features in this research. What is the impact of feature selection on the unemployment forecasting accuracy?

   Seventeen feature selection techniques were employed to assess the impact of using different feature subsets. It was observed that multivariate feature selection techniques

have the biggest impact on the South African unemployment rate. There are, therefore, group dynamics in these features that must be taken into account when forecasting the unemployment rate. ENET and LASSO are the top performing feature selection techniques: either directly or as black-box estimators for wrapper methods.

ENET (which was consistently top performing) identified five features as important in forecasting the South African unemployment rate, Final consumption expenditure by general government; domestic output: all groups; foreign exchange rate: SA rand per USA dollar; total outstanding domestic non-marketable loan levies; and total foreign non-marketable debt.

3. **Research Question 3** What is the relationship between the various features of the unemployment rate and how do they influence the unemployment rate and each other?

   A Dynamic Bayesian Network (DBN) was constructed from the data. The DBN showed how different features influence each other and the unemployment rate. The rate is only directly influenced by the current values of the features final consumption expenditure by general government; domestic output: all groups; and total outstanding domestic non-marketable loan levies. The lag value of the unemployment rate also influences it.

   The different features also influence each other almost equally as seen in the DBN. Leveraging this causal structure, do-Calculus was performed to consider different counterfactuals. Decreasing either the final consumption expenditure by general government or the total outstanding domestic non-marketable loan levies results in increases in the unemployment rate.

This chapter demonstrated that this research adds to a growing body of literature on forecasting unemployment rates. The research used an end-to-end machine learning approach: a data-driven process. Furthermore, including DBNs to show how various features influence each other. The next chapter will provide concluding remarks as well as restating the contributions of the research.

5 Conclusion & Future Work

Through 1522 experiments, the research demonstrated that machine learning models can be used to forecast the unemployment rate more accurately than the orthodox approach, VAR. Furthermore identifying impactful features and their influence on each other and the unemployment rate. Therefore, successfully achieving the objectives of the research.

This chapter is the concluding chapter of the research. Section 5.1 summarized the research and its contributions. Section 5.2 discusses three opportunities that future researchers should consider. The last section, 5.3, provides three data determined policy recommendations that the South African government should consider.

5.1 Conclusion

Unemployment is one of the most significant challenges that South Africa faces. Almost a third of the employable population and over half of those aged below 35 are without jobs. This is a problem that has persisted in South African since the dawn of democracy in 1994, and there seems to be no end in sight. Over the last two and half decades multiple policy interventions have been attempted by the South African government to address this challenge but none of them have been widely successful.

The search for policy interventions that will change the South African unemployment narrative is ongoing. This is an important search because the right policy intervention can systematically reduce the unemployment rate. The development of such policies is reliant on accurate forecasts, which are typically performed using traditional statistical tools such as VAR.

The challenge with models such as VAR is that they are built for data that is linear and stationary. However, the South African unemployment rate is asymmetric, nonlinear, non-stationary, seasonal, and trendy. Hence, traditional statistical methods are prone to error

when forecasting unemployment. Because of this challenge, researchers have started experimenting with machine learning models to forecast unemployment rates across the world.

This research set out to i) explore the the ability of machine learning models to forecast the South African unemployment rate; ii) identify key features that assist in forecasting the unemployment rate; and iii) identify how these different features influence each other and the unemployment rate. These objectives have all been met and are discussed in section 4.7.

To achieve the objectives of the research 1522 experiements were run. This was done using twelve machine learning models with VAR as the benchmark with seventeen feature selection techniques. MLP and SVR were the top two performing models. The machine learning models were over 100 times more accurate than the benchmark model. Through ENET, five features (from 147) that impact the South African unemployment rate the most were identified. The feature subset was then used to build a Dynamic Bayesian network (DBN).

Three of the five features have a direct impact on the unemployment rate: final consumption expenditure by general government (FGE); domestic output: all groups (GDP); and total outstanding domestic non-marketable loan levies (TDL). Whilst the remaining two indirectly influence it, these being foreign exchange rate: SA rand per USA dollar (FER) and total foreign non-marketable debt (TFD).

The research also used do-Calculus to demonstrate that increasing GDP can result in a reduction in the unemployment rate. However, decreases in GDP increase the unemployment rate by greater magnitude than the increases in GDP. Therefore, the government should make all the effort to ensure that either GDP remains the same or increases, as its declines can cancel a significant amount of the progress made by its positive changes. Reducing FGE and TDL also results in decrease in the unemployment rate: both these are levers that the government has full control over.

This research and its outputs compliments previous research on forecasting unemployment rates across the world. Whilst previous research is concentrated in North America, Europe, and Asia, this research uses machine learning models for an African context. Particularly South Africa, where, the unemployment rate is upward trending, nonstationary, and nonlinear. Expanding the current literature by incorporating filter, wrapper, and embedded feature selection techniques.

The research used an end-to-end machine learning approach: a data-driven process. Furthermore, including DBNs to show how various features influence each other. To the best of our knowledge this research is the first to demonstrate the value of machine learning in forecasting a macroeconomic variable across the 'whole' machine learning value chain: feature selection, forecasting, feature influence analysis, and do-Calculus scenario analysis. Others tend to only focus one or two of these aspects: as can be seen in Table 2.1.

Furthermore, the focus in unemployment rate forecasting tends to be on feature selection and feature extraction. Feature engineering receives little to no focus and has been used to a limited extent. This research demonstrates that feature engineering improves the performance of machine learning-based unemployment rate forecasts: MASE reduction and an increase in R-squared.

The inclusion of do-Calculus in this research and the demonstrated value is extremely important to the forecasting research space because often the consumers of macroeconomic forecasts, policymakers, are interested in knowing how different interventions would impact the unemployment rate. Therefore, this demonstration is significant and important as it paves the way for more forecasters to embrace such a methodology.

5.2 Policy Recommendations

Multiple policies have been attempted to reduce the South African unemployment rate. Through a data-driven machine learning process, the South African government should focus on three areas to reduce the unemployment rate:

- **Avoid, at all cost, GDP decreases** *Increasing the GDP has been a priority of the South African government since the dawn of democracy. Achieving this priority does positively impact the South African unemployment rate. However, a decline in GDP has a greater negative impact than the positive created by the increase. Therefore, policymakers should ensure that the balance of their policies do not negatively impact the GDP as this can wipe out any progress that has been made in addressing the unemployment rate.*

- **Reduce government expenditure** *Looking at all the policy interventions that the government has attempted to reduce the unemployment rate. It can sometimes feel like there is no end in sight. However, our scenario analysis shows that reducing government spend can result in a decrease in the unemployment rate. This seems like an easy lever for the government to use. Instead of spending more, the government should look to more incentive orientated interventions that leverages private capital without increasing the spend of the government significantly.*

- **Manage government debt** *The South African government debt has been rising at an alarming rate. Improved the management of this directly impacts the government expenditure as well as impacting the unemployment rate positively as well. Therefore, the South African government should look at fiscal responsibility as an indirect lever to influence the South African unemployment rate.*

The analysis, whilst experimental and requiring more work to be useful in the real world, shows that using a data-driven process, the South African government can mine opportunities to influence the unemployment rate through mechanisms that are not always obvious.

5.3 Future Work

This research has demonstrated the effectiveness of using machine learning to forecast the unemployment rate. Three areas have been identified as opportunities for further research.

Research opportunity 1: This research explored how probabilistic graphical models can be used for scenario analysis using probabilistic approaches. Further research is required in this space to demonstrate and develop the do-Calculus approach further. The research demonstrated the approach but much work is required to ensure this is effectively applied for unemployment rate forecasting.

Research opportunity 2: Due to the mixed frequencies, data imputation strategies were employed. Whilst these are widely and effective used, research should be conducted in using high dimensional data with mixed frequencies without any imputation. The challenge with any imputation method is that by definition it adds data to the features that did not exit. This matters more in macroeconomic forecasting because the missing data does not exist because it was never collected by design and there is no intention to ever collect it. Therefore, there are opportunities to develop mixed frequency forecasting approaches for high dimensional data.

Bibliography

[Ahmed *et al.* 2010] Nesreen K. Ahmed, Amir F. Atiya, Neamat El Gayar, and Hisham El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5):594–621, 2010.

[Aiken 1996] Milam Aiken. *A neural network to predict civilian unemployment rates*. Technical report, 1996.

[Ajoodha 2018] Ritesh Ajoodha. *Influence modelling an dlearning between dynamic bayesian networks using score based structure learning*. PhD thesis, University of Witswatersrand, 2018.

[Atsalakis *et al.* 2007] George S. Atsalakis, Camelia Ioana, and Christos H. Skiadas. Forecasting Unemployment Rate Using a Neural Network with Fuzzy Inference System. *ICAP*, 2007.

[Ball *et al.* 2019] Laurence Ball, Davide Furceri, Daniel Leigh, and Prakash Loungani. Does One Law Fit All? Cross-Country Evidence on Okun's Law. *Open Economies Review*, nov 2019.

[Barnichon and Nekarda 2013] Regis Barnichon and Christopher J Nekarda. *The Ins and Outs of Forecasting Unemployment: Using Labor Force Flows to Forecast the Labor Market*. Technical report, Board of Governors of the Federal Reserve System, 2013.

[Belling 2020] Ayal Belling. *South Africa's State of Unemployment Disaster*, 2020.

[Bertsimas *et al.* 2018] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research*, 18:1–39, 2018.

[Bloomberg 2020] Bloomberg. *South African Employment Drops to Lowest Level in Nine Years*, 2020.

[Brooks 2014] Chris Brooks. *Introductory Econometrics for Finance*. Cambridge University Press, 3rd edition, 2014.

[Brynard 2011] P A Brynard. *The Implementation of Unemployment Policies in South Africa*. Technical Report 2, 2011.

[Burger and Fourie 2015] Philippe Burger and Frederick Fourie. Macroeconomic policy and South African unemployment: developing a three-segment macroeconomic model. 2015.

[Burger and Fourie 2019] Philippe Burger and Frederick Fourie. The unemployed and the formal and informal sectors in South Africa: A macroeconomic analysis. *South African Journal of Economic and Management Sciences*, 22(1), 2019.

[Cerqueira *et al.* 2019] Vitor Cerqueira, Luis Torgo, and Carlos Soares. Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters. *arXiv*, 2019.

[Chandrashekar and Sahin 2013] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2013.

[Clements *et al.* 2004] Michael P Clements, Philip Hans Franses, and Norman R Swanson. Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, 20(2):169–183, 2004.

[Cook and Hall 2017] Thomas R. Cook and Aaron Smalter Hall. Macroeconomic Indicator Forecasting with Deep Neural Networks. 2017.

[Davies *et al.* 2009] Rob Davies, Dirk Van Seventer, and Miriam Altman. *Leading indicators of employment in South Africa*. Technical report, Human Sciences Research Council, Pretoria, 2009.

[De Gooijer and Hyndman 2006] Jan G De Gooijer and Rob J Hyndman. *25 Years of Time Series Forecasting*. Technical report, 2006.

[De Lannoy *et al.* 2018] Ariane De Lannoy, Lauren Graham, Leila Patel, and Murray Leibbrandt. *What drives youth unemployment and what interventions help? A systematic overview of the evidence and a theory of change high-level overview report*. Technical report, 2018.

[Diana *et al.* 2014] Rita Diana, I. Nyoman Budiantara, Purhadi, and Satwiko Darmesto. Statistical modeling for unemployment rate using smoothing spline in semiparametric multivariable regression model with Bayesian approach. *Model Assisted Statistics and Applications*, 9(2):159–166, 2014.

[Fok *et al.* 2005] Dennis Fok, Dick van Dijk, and Fransesa Hans Philip. Forecasting aggregates using panels of nonlinear time series. *International Journal of Forecasting*, 21, 2005.

[Fourie 2011] Frederick Fourie. The South African unemployment debate: three worlds, three discourses? 2011.

[Fransesa *et al.* 2004] Hans Philip Fransesa, Richard Paapa, and Björn Vroomen. Forecasting unemployment using an autoregression with censored latent effects parameters. *International Journal of Forecasting*, 20(2):255–271, 2004.

[Funke 1992] Michael Funke. Time-series forecasting of the German unemployment rate. *Journal of Forecasting*, 11(2):111–125, feb 1992.

[Goodfellow *et al.* 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[Guyon and Elisseefl 2006] Isabelle Guyon and André Elisseefl. An introduction to feature extraction. *Studies in Fuzziness and Soft Computing*, 207:1–25, 2006.

[Hall 2018] Aaron Smalter Hall. *Machine Learning Approaches to Macroeconomic Forecasting*. Technical report, Economic Review (Kansas City, MO), 2018.

[Hyndman and Athanasopoulos 2018] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2 edition, 2018.

[Hyndman 2006] Rob J Hyndman. *Another look at forecast-accuracy metrics for intermittent demand*, 2006.

[Jelena *et al.* 2017] Mladenovic Jelena, Ilic Ivana, and Kostic Zorana. Modeling The Unemployment Rate At The Eu Level By Using Box-Jenkins Methodology. *KnE Social Sciences*, 1(2):1, mar 2017.

[Jones 2019] Aidan Jones. *Unemployment: our biggest problem*, 2019.

[Katris 2019a] Christos Katris. Forecasting the Unemployment of Med Counties using Time Series and Neural Network models Forecasting the Unemployment of Med Counties using Time Series. *Journal of Statistical and Econometric Methods*, 8(2):37–49, 2019.

[Katris 2019b] Christos Katris. Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Computational Economics*, 2019.

[Koller and Friedman 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2009.

[Kouziokas 2019] Georgios N. Kouziokas. Unemployment Prediction in UK by Using a Feedforward Multilayer Perceptron. pages 65–74. 2019.

[Krauss *et al.* 2016] Christopher ; Krauss, Xuan Anh Do, and Nicolas Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *ECONSTOR*, 3(2016), mar 2016.

[Kreiner and Duca 2019] Aaron Kreiner and John V. Duca. Can machine learning on economic data better forecast the unemployment rate? *Applied Economics Letters*, 2019.

[Kyei and Gyekye 2017] Kwabena A Kyei and Kwame B Gyekye. Determinants of Unemployment in Limpopo Province in South Africa: Exploratory Studies. *Journal of Emerging Trends in Economics and Management Sciences*, 2(1):54, 2017.

[Levinsohn 2007] James Levinsohn. *Two Policies to Alleviate Unemployment in South Africa*. Technical report, 2007.

[Liang 2005] Faming Liang. Bayesian neural networks for nonlinear time series forecasting. *Statistics and Computing*, 15(1):13–29, jan 2005.

[Mahipan *et al.* 2013] Kanlapat Mahipan, Nipaporn Chutiman, and Bungon Kumphon. A forecasting model for thailand's unemployment rate. *Modern Applied Science*, 7(7):10–16, 2013.

[Makridakis *et al.* 2009] Spyros Makridakis, Robin M. Hogarth, and Anil Gaba. Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, 25(4):794–812, oct 2009.

[Makridakis *et al.* 2018] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889, mar 2018.

[Makridakis *et al.* 2020] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, jan 2020.

[Makridakis 1988] Spyros Makridakis. Metaforecasting. Ways of improving forecasting accuracy and usefulness. *International Journal of Forecasting*, 4(3):467–491, jan 1988.

[Maluleke 2020] Risenga Maluleke. *Quarterly Labour Force Survey (QLFS)* . Technical report, Statistic South Africa, Pretoria, jun 2020.

[Marinkov and Geldenhuys 2007] Marina Marinkov and Jean-pierre Geldenhuys. Cyclical unemployment and cyclical output: an estimation of okun's coefficient for South Africa. *The South African Journal of Economics*, 75(3):373–390, sep 2007.

[Mathworks 2020] Mathworks. *Learning Vector Quantization (LVQ) Neural Networks - MATLAB & Simulink*, 2020.

[Meyer 2014] Daniel Francois Meyer. Job creation, a mission impossible? The South African case. *Mediterranean Journal of Social Sciences*, 5(16):65–77, 2014.

[Mitchell 1997] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[Montgomery *et al.* 1998] Alan L. Montgomery, Victor Zarnowitz, Ruey S. Tsay, and George C. Tiao. Forecasting the U.S. Unemployment Rate. *Journal of the American Statistical Association*, 93(442):478, jun 1998.

[Mulaudzi and Ajoodha 2020a] Rudzani Mulaudzi and Ritesh Ajoodha. Application of Deep Learning to Forecast the South African Unemployment Rate: A Multivariate Approach. In *Seventh IEEE Computer Science and Data Engineering Conference*, volume 7, Melbourne, Australia, November 2020. IEEE.

[Mulaudzi and Ajoodha 2020b] Rudzani Mulaudzi and Ritesh Ajoodha. An exploration of machine learning models to forecast the unemployment rate of South Africa: a univariate approach. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference*, volume 2, East London, South Africa, November 2020. IEEE.

[Mulaudzi and Ajoodha 2021a] Rudzani Mulaudzi and Ritesh Ajoodha. Demonstration that the Use of Feature Selection on High Dimensional South African Macroeconomic Data Results in Improved Performance with Lower Compute Requirements. In *Sixth International Congress on Information and Communication Technology Conference*, volume 6, London, UK, February 2021. Springer.

[Mulaudzi and Ajoodha 2021b] Rudzani Mulaudzi and Ritesh Ajoodha. Improving the performance of multivariate forecasting models through feature engineering: A south african unemployment rate forecasting case study. In *2021 International Conference On Interdisciplinary Research in Technology and Management*, volume 6, Kolkata, India, February 2021. Taylor & Francis.

[Olmedo 2014] Elena Olmedo. Forecasting Spanish Unemployment Using Near Neighbour and Neural Net Techniques. *Computational Economics*, 43(2):183–197, jan 2014.

[Pamfil *et al.* 2020] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Paul Beaumont, Konstantinos Georgatzis, and Bryon Aragam. *DYNOTEARS: Structure Learning from Time-Series Data.* Technical report, 2020.

[Patro and Sahu 2015] S.Gopal Krishna Patro and Kishore Kumar Sahu. Normalization: A Preprocessing Stage. *Iarjset*, (April):20–22, 2015.

[Pearl 2011] Judea Pearl. *Causality: Models, reasoning, and inference, second edition.* Cambridge University Press, jan 2011.

[Pelaez 2006] Rolando F. Pelaez. *Using neural nets to forecast the unemployment rate*, jan 2006.

[QuantumBlack 2020] QuantumBlack. *Welcome to CausalNex's API docs and tutorials! — causalnex 0.9.1 documentation*, jan 2020.

[Sermpinis *et al.* 2014] Georgios Sermpinis, Charalampos Stasinakis, Konstantinos Theofilatos, and Andreas Karathanasopoulos. Inflation and Unemployment Forecasting with Genetic Support Vector Regression. *Journal of Forecasting*, 33(6):471–487, sep 2014.

[Shankar *et al.* 2016] Vignesh Shankar, Ansulie Cooper, and Harvey Koh. *Reducing Youth Unemployment in South Africa*. Technical report, 2016.

[Sharma and Singh 2016] Saloni Sharma and Sanjay Singh. Unemployment rates forecasting using supervised neural networks. In *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*, pages 28–33. Institute of Electrical and Electronics Engineers Inc., jul 2016.

[South African Reserve Bank 2020] South African Reserve Bank. *Welcome to the South African Reserve Bank - South African Reserve Bank*, 2020.

[Statistics South Africa 2019] Statistics South Africa. *Statistical release P0211*. Technical report, Statistics South Africa, Pretoria, 2019.

[Stock and Watson 2001] James H. Stock and Mark W. Watson. Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115, 2001.

[TradingEconomics 2019] TradingEconomics. *Unemployment rate - Countries list*, 2019.

[Vermeulen 2017] J.C. Vermeulen. Inflation and unemployment in South Africa: Is the Phillips curve still dead? *Southern African Business Review*, 21, 2017.

[Yang 2007] Lijian Yang. *Nonparametric Modelling of Quarterly Unemployment Rates*. Technical report, 2007.

[Yolcu and Bas 2016] Ufuk Yolcu and Eren Bas. The forecasting of labour force participation and the unemployment rate in Poland and Turkey using fuzzy time series methods. *Comparative Economic Research*, 19(2):5–25, jun 2016.

[Zheng *et al.* 2018] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. Montréal, 2018. Neural Information Processing Systems.

Appendices

A

Feature List

Table A.1 shows the list of features that were used for this research. The data was accessed from the South African Reserve Bank (SARB). The SARB mostly collects and reports its own data, however, a few features are sourced from Statistics South Africa. Therefore, this data is reliable and the data generating process is transparent and accessible on the SARB website: this refered to as the Special Data Dissemination Standard (SDDS). The data has coverage of a number of key economic sectors in South Africa: real, fiscal, financial, and external sector as well as population data.

It is also worth noting that economic data is typically available in constant prices or current prices. Constant prices are prices as at a given date, therefore, the value today is not affected by economic changes that would not make a non-financial difference to it. For example inflation causes prices to change, not because anything has changed about the goods or services but because time has passed. Current prices are those that incorporate these financial changes such as inflation. For this research constant prices were used to avoid inflation being a confounding variable across across our data. Inflation (CPI) is itself a feature that was used for this research.

Table A.1: List of features that sourced from the South African Reserve Bank

	Data Frequency			Date Range	Literature Use
Feature	**Monthly**	**Quarterly**	**Annual**		
Gross value added at basic prices of agriculture, forestry and fishing		X		1960 to 2020	
Value added at basic prices of mining and quarrying		X		1960 to 2020	
Gross value added at basic prices of manufacturing		X		1960 to 2020	
Gross value added at basic prices of electricity, gas and water		X		1960 to 2020	
Gross value added at basic prices of construction (contractors)		X		1960 to 2020	
Gross value added at basic prices of wholesale and retail trade, catering and accommodation		X		1960 to 2020	
Gross value added at basic prices of transport, storage and communication		X		1960 to 2020	
Gross value added at basic prices of finance, insurance, real estate and business services		X		1960 to 2020	
Gross value added at basic prices of other community, social and personal services		X		1960 to 2020	
Gross value added at basic prices of general government services		X		1960 to 2020	
Gross value added at basic prices of all industries		X		1960 to 2020	
Gross domestic product at market prices		X		1960 to 2020	
Final consumption expenditure by households: Total		X		1960 to 2020	
Final consumption expenditure by general government		X		1960 to 2020	

Table A.1 continued from previous page

Gross fixed capital formation		X		1960 to 2020	
Change in inventories		X		1960 to 2020	
Residual item		X		1960 to 2020	
Gross domestic expenditure		X		1960 to 2020	
Exports of goods & services		X		1960 to 2020	
Imports of goods & services		X		1960 to 2020	
Final consumption expenditure by households: Total		X		1960 to 2020	
Final consumption expenditure by general government		X		1960 to 2020	
Gross fixed capital formation		X		1960 to 2020	
Change in inventories		X		1960 to 2020	
Residual item		X		1960 to 2020	
Gross domestic expenditure		X		1960 to 2020	
Exports of goods & services		X		1960 to 2020	
Imports of goods & services		X		1960 to 2020	
Physical volume of manufacturing production: Total	X			1963 to 2020	
The leading business cycle indicator is a composite index comprising of time series, which tend to shift direction in advance of the business cycle	X			1960 to 2020	Used by the SARB for the unemployment leading indicator [Davies *et al.* 2009].
Non-agricultural employment: Total		X		1970 to 2020	
Remuneration per worker in non-agricultural: Total		X		1970 to 2020	
Remuneration per worker in non-agricultural: Total		X		1970 to 2020	

Table A.1 continued from previous page

CPI Headline	X			1922 to 2020	Commonly used in economics, the inverse relationship between CPI (inflation) and unemployment is refereed to as Philips Curve [Marinkov and Geldenhuys 2007; Vermeulen 2017])
Domestic output: All groups	X			1970 to 2020	Commonly used in economics, the relationship between GDP (Domestic output) and unemployment is refereed to as Okun's Law [Marinkov and Geldenhuys 2007; Vermeulen 2017])
Consolidated general government: Revenue		X		1991 to 2020	
Consolidated general government: Expense		X		1991 to 2020	
Consolidated general government: Expense: Interest		X		1991 to 2020	
Net cash-flow from operating activities		X		1991 to 2020	
Consolidated general government: Non-financial assets - Net		X		1960 to 2020	
Consolidated general government: Non-financial assets - Net		X		1991 to 2020	
Consolidated general government: Cash surplus/deficit		X		1991 to 2020	

Table A.1 continued from previous page

Consolidated general government: Net cash-flow from financing activities		X		1991 to 2020	
Consolidated general government: Net acquisition of financial assets other than cash		X		1991 to 2020	
Domestic Loans		X		2002 to 2020	
Other domestic financial assets		X		2002 to 2020	
Foreign		X		2002 to 2020	
Consolidated general government: Liabilities		X		1991 to 2020	
Consolidated general government: Liabilities: Domestic		X		1991 to 2020	
Consolidated general government: Liabilities: Domestic: Currency and deposits		X		1991 to 2020	
Consolidated general government: Liabilities: Domestic: Securities other than shares		X		1991 to 2020	
Consolidated general government: Liabilities: Domestic: Loans		X		1991 to 2020	
Agriculture, forestry and fishing		X		1960 to 2020	
Consolidated general government: Liabilities: Domestic: Other domestic liabilities		X		2005 to 2020	
Consolidated general government: Liabilities: Foreign		X		1991 to 2020	
Consolidated general government: Liabilities: Foreign: Currency and deposits		X		1991 to 2020	
Consolidated general government: Liabilities: Foreign: Securities other than shares		X		1991 to 2020	
Consolidated general government: Liabilities: Foreign: Loans		X		1991 to 2020	
Consolidated general government: Net change in the stock of cash		X		1991 to 2020	

Table A.1 continued from previous page

National government revenue adjusted for differences in timing between the recording of transactions and bank clearances.	X			1960 to 2020	
National government expenditure adjusted for differences between the recording of transactions and bank clearances, and late departmental requests/surrenders of funds.	X			1960 to 2020	
The difference between cash-flow revenue and cash-flow expenditure.	X			1960 to 2020	
Cost/profit of revaluation of foreign loans	X			1960 to 2020	
Plus: Accrual adjustment	X			2000 to 2020	
The net borrowing requirement of national government equals the net result of the cash-flow deficit/surplus and the cost of revaluation of foreign loans at redemption.	X			1960 to 2020	
Cash received from the issuance of Treasury bills less cash payments made to redeem Treasury bills.	X			1960 to 2020	
Cash received from the issuance of domestic government bonds less cash payments made to redeem domestic government bonds	X			1984 to 2020	
Cash received from the issuance of other debt instruments less cash payments made to redeem other debt instruments, including non-marketable bills and bonds, and debt of the former TBVC-countries, self-governing territories and the Republic of Namibia that were assumed by the national government	X			1960 to 2020	
Financing by non-marketable bonds	X			1960 to 2020	
Total cash-flow revenue	X			1960 to 2020	

Table A.1 continued from previous page

Including changes in the Exchequer and Paymaster-General Account balances with the South African Reserve Bank and other banks.	X			1960 to 2020	
The net result of all financing activities of national government.	X			1984 to 2020	
Total gross loan debt (nsa)	X			1960 to 2020	
Total outstanding domestic marketable debt	X			1960 to 2020	
Total outstanding domestic marketable bills	X			1960 to 2020	
Total domestic marketable bonds	X			1960 to 2020	
Total outstanding domestic marketable bonds: Not exceeding 1 year (incl SA housing trust loans i.e marketable SP05 bonds)	X			1855 to 2020	
Total outstanding domestic marketable bonds: Exceeding 1 but not 3 years	X			1986 to 2020	
Total outstanding domestic marketable bonds: Medium term	X			1986 to 2020	
Total outstanding domestic marketable bonds: Total exceeding 10 years	X			1986 to 2020	
Total outstanding domestic marketable loan levies	X			1989 to 2020	
Total outstanding domestic non-marketable debt	X			1960 to 2020	
Total outstanding domestic non-marketable bills	X			1960 to 2020	
Total outstanding domestic non-marketable bonds	X			1960 to 2020	
Domestic non-marketable bonds by instrument type: Floating/variable rate bonds		X		1987 to 2020	

Table A.1 continued from previous page

Total outstanding domestic non-marketable loan levies	X			1960 to 2020	
Total foreign debt	X			1960 to 2020	
Total foreign marketable debt	X			1960 to 2020	
Total outstanding foreign marketable bonds by unexpired maturity: Not exceeding 1 year	X			1960 to 2020	
Total outstanding foreign marketable bonds by unexpired maturity: Exceeding 1 but not 3 years	X			1960 to 2020	
Total outstanding foreign marketable bonds by unexpired maturity: Exceeding 3 years	X			1960 to 2020	
Total foreign non-marketable debt	X			1960 to 2020	
Total other debt	X			1993 to 2020	
Gold and Foreign Exchange Contingency Reserve Account	X			2018 to 2020	
Financial guarantees by government		X		2018 to 2020	
Broad monetary credit to the government sector. Net Claims on Government are shown as a counterpart item to the change in M3. It is the total of "gross claims on the government sector" minus "Government deposits".	X			1965 to 2020	

Table A.1 continued from previous page

Claims on the domestic private sector. Domestic credit extended by monetary financial institutions is disaggregated into: (i) Credit extended to the domestic private sector (Claims on the domestic private sector) which is further subdivided into investments, bills discounted, instalment sale credit, leasing finance, mortgage advances and other loans and advances; and (ii) Net credit extended to the government sector (comprising the central government, provincial governments and social security funds); that is, credit extended less government deposits.	X			1965 to 2020	
Net foreign assets after valuation adjustment. Net foreign assets are shown as a counterpart item to the change in M3. It is the total of foreign assets minus foreign liabilities. Deposits of non-residents are not included in the broad monetary aggregate.	X			1960 to 2020	
M2 plus all long term deposit with monetary banking institutions	X			1965 to 2020	
Total claims on the private sector	X			1993 to 2020	
Total claims on the government	X			1983 to 2020	
Total foreign assets minus total foreign liabilities	X			1978 to 2020	
Deposit of banks,mutual banks with the SARB and notes and coins outside the SARB and SA mint	X			1965 to 2020	
Current account (nsa)		X		1960 to 2020	
Goods: exports (nsa)		X		1960 to 2020	
Goods: imports (nsa)		X		1960 to 2020	

Table A.1 continued from previous page

Services: credit (nsa)		X		2015 to 2020	
Services: debit (nsa)		X		2015 to 2020	
Income: credit (nsa)		X		2015 to 2020	
Income: debit (nsa)		X		2015 to 2020	
Current transfers: credit (nsa)		X		2015 to 2020	
Current transfers: debit (nsa)		X		2015 to 2020	
Capital account (nsa)		X		2015 to 2020	
Capital account: credit (nsa)		X		2015 to 2020	
Capital account: debit (nsa)		X		2015 to 2020	
Financial account		X		2015 to 2020	
Direct investment abroad		X		2015 to 2020	
Direct investment in reporting economy		X		2015 to 2020	
Portfolio investment, assets		X		2015 to 2020	
Portfolio investment, liabilities		X		2015 to 2020	
Financial derivatives, assets		X		2015 to 2020	
Financial derivatives, liabilities		X		2015 to 2020	
Other investment, assets		X		2015 to 2020	
Other investment, liabilities		X		2015 to 2020	
Reserve assets		X		2015 to 2020	
Net errors and omissions		X		2015 to 2020	
Gross reserves of the Reserve Bank - US Dollar	X			2018 to 2020	
IMF reserve position	X			2018 to 2020	
Reserve Bank SDR reserves	X			2018 to 2020	
Reserve Bank gold reserves	X			2018 to 2020	
Reserve Bank other foreign exchange reserves	X			2018 to 2020	
Other reserve assets	X			2018 to 2020	
Gross gold and other foreign reserves : Amount as at end of period	X			2018 to 2020	
Exports of goods, including gold (SARS)	X			2012 to 2020	

Table A.1 continued from previous page

Imports of goods (SARS)	X			2012 to 2020	
Total foreign assets		X		1956 to 2020	
Foreign assets: Total direct investment		X		2015 to 2020	
Foreign assets: Total portfolio investment		X		2015 to 2020	
Foreign assets: Portfolio investment: Equity securities		X		2015 to 2020	
SDDS - Financial derivative assests		X		2015 to 2020	
Foreign assets: Total other investment		X		2015 to 2020	
SDDS - Reserve Assest		X		2015 to 2020	
Total foreign liabilities		X		2015 to 2020	
Foreign liabilities: Total direct investment		X		2015 to 2020	
Foreign liabilities: Total portfolio investment		X		2015 to 2020	
Foreign liabilities: Portfolio investment: Equity securities		X		2015 to 2020	
Foreign liabilities: Portfolio investment: Debt securities		X		2015 to 2020	
SDDS - Financial derivative liabilities		X		2015 to 2020	
Foreign liabilities: Total other investment		X		2015 to 2020	
Net Investment Position		X		2015 to 2020	
Foreign debt of S.A.: Total foreign debt		X		2015 to 2020	
Foreign exchange rate : SA rand per USA dollar		X		2018 to 2020	
The nominal effective exchange rate is a weighted exchange rate of the rand measured against a basket of the currencies of South Africa's twenty most important trading partners. The calculation is based on trade in and consumption of manufactured goods	X			2018 to 2020	
Total South African population			X	2015 to 2020	

Table A.1 continued from previous page

	58	112	1	1922 to 2020	

B

Hyperparameter Settings

Grid search and trial & error were the primary methods for tuning the hyperparameters of the machine learning models. A range was first attempted such as 1e−10, 1e−100, 1e−1000. This would be followed by a deeper grid search such as a range from 1e−10 to 9e−10. The last step was then hand tuned improvements. In cases where a model had multiple hyperparamters, one would be tuned with others being held fixed. Sometimes this would result in loops nested in loops. The hyperparameter settings that had the lowest MASE is show in Table B.1.

Table B.1: The hyperparameter settings for the twelve machine learning models used for this research. Only key parameters are shown.

Model	Hyper Paramers
SVR	Kernel: linear Polynomial degree: 3 Regulization parameter: 1.37 Epsilon: 0.02 Stopping criteria tolerence: 0.001
Bayes Ridge	Paramater for gamma distruction over L1: 0.000006 Paramater for gamma distruction over L2: 0.000006 Compute marginal likelihood at each iteration: True
LR	None
Ridge	L2 regulization strength: 68 Solver: Automatic
LASSO	L1 regulization strength: 0.32
RFR	Criteria: mean squared error Minimum number of samples required to split node: 2 Minimum number of samples required to be a leaf: 1 Number of estimators: 100
ENET	Regulization strength (alpha): 0.03

Table B.1 continued from previous page

Model	Hyper Paramers
LSTM	Number of layers: Six layers with 64, 32, 16, 8, 8, and 1 nodes at each layer Optimizer: adam Loss: mean squared error Metrics: accuracy
GRU	Number of layers: Six layers with 64, 32, 16, 8, 8, and 1 nodes at each layer Optimizer: adam Loss: mean squared error Metrics: accuracy
MLP	Hidden-layer size: 43 Activation function: tanh Solver: lbfgs Alpha: 0.3067 Learning rate: adaptive Learning rate initialization: 0.001 Shuffle: False Early stopping: True Validation fraction: 0.1
BiLSTM	Number of layers: Six layers with 64, 32, 16, 8, 8, and 1 nodes at each layer Optimizer: adam Loss: mean squared error Metrics: accuracy
XGB	Objective: squared error Column sample by tree ratio: 0.3 Learning rate: 0.1 Maximum depth: 6 Regulirization (alpha): 45 Number of estimators: 4

C

Regressions Models Results

Table C.1: The performance (MASE) of all regression machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)					
Model	**Feature Selection**	**H1**	**H2**	**H3**	**AVG**	**H4**	**CV**
Bayes Ridge	EM ENET	0.3108	0.2546	1.0929	0.4422	0.5251	0.3410
LR	EM ENET	0.3097	0.2531	1.1031	0.4434	0.5273	0.4123
Ridge	EM ENET	0.4493	0.2887	1.0046	0.4523	0.5487	0.3410
ENET	MIG	0.3537	0.2747	1.2919	0.5005	0.6052	0.3635
LR	RFE ENET	0.8211	0.5086	0.5899	0.6189	0.6346	0.2744
Bayes Ridge	RFE ENET	0.8246	0.5103	0.5920	0.6192	0.6365	0.7577
LASSO	MIG	0.4415	0.2990	1.2940	0.5153	0.6374	0.4950
Ridge	RFE LASSO	0.4789	0.4983	1.0626	0.5949	0.6587	0.3365
Ridge	RFE ENET	0.8828	0.5383	0.6269	0.6242	0.6680	0.3819
Bayes Ridge	RFE LASSO	0.5176	0.5392	1.0794	0.6014	0.6844	0.5384
ENET	RFE ENET	0.8086	0.6312	0.7500	0.5580	0.6870	0.3856
LR	RFE LASSO	0.5220	0.5438	1.0811	0.6022	0.6873	0.4231

Model	Feature Selection	H1	H2	H3	AVG	H4	CV
Ridge	RANDOM FOREST	0.4591	0.4621	1.0995	0.7387	0.6898	0.3450
Bayes Ridge	EM LASSO	0.3201	0.3048	1.4005	0.7787	0.7010	0.5713
LASSO	RFE ENET	0.8264	0.6542	0.7757	0.5594	0.7039	0.4950
LR	EM LASSO	0.3506	0.3078	1.4109	0.7900	0.7148	0.4780
Ridge	EM LASSO	0.5160	0.5357	1.4362	0.6329	0.7802	0.2967
LASSO	RFE LASSO	0.5130	0.4904	1.5839	0.5845	0.7930	0.4950
LR	VARIANCE UNIQUE NO CORR	1.1843	0.6598	0.7350	0.6933	0.8181	7.80E+09
LR	UNIQUE NO CORR	1.1843	0.6598	0.7350	0.6933	0.8181	7.75E+10
LR	NO CORR	1.1843	0.6598	0.7350	0.6933	0.8181	1.38E+09
ENET	RFE LASSO	0.5447	0.5235	1.6258	0.5896	0.8209	0.3541
Bayes Ridge	XGBOOST	0.8611	0.6984	1.1979	0.6285	0.8465	0.4493
ENET	MIG NO CORR	0.9204	0.8261	1.2535	0.4831	0.8708	0.3770
ENET	NO CORR	0.9204	0.8261	1.2535	0.4831	0.8708	0.3831
ENET	UNIQUE NO CORR	0.9204	0.8261	1.2535	0.4831	0.8708	0.3831
ENET	VARIANCE UNIQUE NO CORR	0.9204	0.8261	1.2535	0.4831	0.8708	0.3831
LASSO	MIG NO CORR	0.9553	0.8386	1.2581	0.4694	0.8803	0.4950
LASSO	NO CORR	0.9553	0.8386	1.2581	0.4694	0.8803	0.4950
LASSO	UNIQUE NO CORR	0.9553	0.8386	1.2581	0.4694	0.8803	0.4950
LASSO	VARIANCE UNIQUE NO CORR	0.9553	0.8386	1.2581	0.4694	0.8803	0.4950
Ridge	MIG NO CORR	0.8242	0.6339	1.2670	0.8884	0.9034	0.3291
LR	XGBOOST	1.1499	0.8355	1.3117	0.6261	0.9808	0.9860
Ridge	XGBOOST	0.6515	1.0774	1.4412	0.7773	0.9869	0.3631
Ridge	ANOVA	1.3284	1.1093	1.2257	0.5032	1.0416	0.3136
ENET	RANDOM FOREST	1.2323	1.0042	1.4005	0.5347	1.0429	0.3771
LASSO	RANDOM FOREST	1.2673	1.0113	1.4029	0.5193	1.0502	0.4950
Bayes Ridge	VARIANCE UNIQUE	1.6836	0.9916	1.0082	0.5226	1.0515	0.3863
Bayes Ridge	Unique	1.6836	0.9916	1.0082	0.5226	1.0515	0.3863
Bayes Ridge	MIG NO CORR	1.0239	0.7417	1.4248	1.0693	1.0649	0.3615
Bayes Ridge	VARIANCE	1.7204	1.0173	1.0299	0.5223	1.0725	0.3864
Bayes Ridge	NO FS	1.7204	1.0173	1.0299	0.5223	1.0725	0.3864

Model	Feature Selection	H1	H2	H3	AVG	H4	CV
ENET	EM LASSO	1.3911	0.8083	1.5295	0.5998	1.0822	0.3556
ENET	ANOVA	1.3913	0.8084	1.5295	0.5998	1.0822	0.3635
ENET	NO FS	1.3913	0.8084	1.5295	0.5998	1.0823	0.3745
ENET	Unique	1.3913	0.8084	1.5295	0.5998	1.0823	0.3745
ENET	VARIANCE	1.3913	0.8084	1.5295	0.5998	1.0823	0.3745
ENET	VARIANCE UNIQUE	1.3913	0.8084	1.5295	0.5998	1.0823	0.3745
ENET	EM ENET	1.3914	0.8084	1.5296	0.5998	1.0823	0.3655
Bayes Ridge	MIG	1.6291	1.0824	1.2523	0.3846	1.0871	0.4423
LR	RANDOM FOREST	1.1335	1.0545	1.4052	0.7566	1.0875	0.8127
LR	MIG NO CORR	1.0642	0.7621	1.4530	1.0950	1.0936	0.4522
Bayes Ridge	RANDOM FOREST	1.1909	1.0672	1.4389	0.7840	1.1202	0.4789
LASSO	EM LASSO	1.5068	0.8645	1.5541	0.6196	1.1362	0.4950
LASSO	Unique	1.5069	0.8646	1.5541	0.6196	1.1363	0.4950
LASSO	ANOVA	1.5069	0.8646	1.5541	0.6196	1.1363	0.4950
LASSO	NO FS	1.5069	0.8646	1.5541	0.6196	1.1363	0.4950
LASSO	VARIANCE	1.5069	0.8646	1.5541	0.6196	1.1363	0.4950
LASSO	VARIANCE UNIQUE	1.5069	0.8646	1.5541	0.6196	1.1363	0.4950
LASSO	EM ENET	1.5070	0.8646	1.5542	0.6196	1.1363	0.4950
Ridge	MIG	1.6176	1.1465	1.3494	0.4956	1.1523	0.3072
ENET	XGBOOST	1.3109	1.2828	1.5530	0.5171	1.1659	0.3771
LASSO	XGBOOST	1.3438	1.2832	1.5509	0.5007	1.1696	0.4950
ENET	PCA	1.6737	1.1122	1.4652	0.5025	1.1884	0.5095
LASSO	PCA	1.6847	1.1222	1.4710	0.5101	1.1970	0.4950
Bayes Ridge	VARIANCE UNIQUE NO CORR	1.4783	1.1517	1.5713	0.5899	1.1978	0.4144
Bayes Ridge	UNIQUE NO CORR	1.4783	1.1517	1.5713	0.5899	1.1978	0.4144
Bayes Ridge	NO CORR	1.4783	1.1517	1.5713	0.5899	1.1978	0.4144
Ridge	VARIANCE UNIQUE	2.0608	1.2680	1.6906	0.6919	1.4278	0.3324
Ridge	Unique	2.0608	1.2680	1.6906	0.6919	1.4278	0.3324
Ridge	VARIANCE	2.0669	1.2721	1.6950	0.6941	1.4320	0.3325
Ridge	NO FS	2.0669	1.2721	1.6950	0.6941	1.4320	0.3325

Model	Feature Selection	H1	H2	H3	AVG	H4	CV
ENET	RFE RIDGE	2.2906	1.2958	1.5378	0.6986	1.4557	0.4950
LASSO	RFE RIDGE	2.3084	1.3082	1.5234	0.7106	1.4626	0.4950
Ridge	UNIQUE NO CORR	1.8634	1.4065	1.9017	0.7235	1.4738	0.3455
Ridge	VARIANCE UNIQUE NO CORR	1.8634	1.4065	1.9017	0.7235	1.4738	0.3455
Ridge	NO CORR	1.8634	1.4065	1.9017	0.7235	1.4738	0.3455
Ridge	PCA	2.5643	1.7517	2.1371	0.6407	1.7735	0.4551
Bayes Ridge	PCA	2.6929	1.8715	2.2961	0.6932	1.8884	0.5406
LR	PCA	2.7090	1.8846	2.3110	0.6979	1.9006	8.3460
Ridge	RFE RIDGE	4.5819	2.8943	3.7466	1.5208	3.1859	0.4836
Bayes Ridge	RFE RIDGE	4.6286	2.9269	3.7854	1.5395	3.2201	0.7752
Bayes Ridge	ANOVA	5.2533	3.5875	3.1667	0.8798	3.2218	0.4636
LR	ANOVA	10.3528	7.1340	6.4048	1.5448	6.3591	4.3569
LR	MIG	9.35E+02	5.58E+02	5.51E+02	1.60E+02	5.51E+02	0.9884
LR	RFE RIDGE	4.24E+03	2.96E+03	1.12E+04	1.54E+04	8.45E+03	2.46E+25
LR	Unique	1.45E+10	1.01E+10	1.02E+10	3.12E+09	9.46E+09	1.88E+08
LR	NO FS	1.50E+10	1.05E+10	1.05E+10	3.24E+09	9.82E+09	8.11E+08
LR	VARIANCE UNIQUE	2.53E+10	1.76E+10	1.77E+10	5.45E+09	1.65E+10	7.77E+08
LR	VARIANCE	3.17E+10	2.21E+10	2.23E+10	6.84E+09	2.07E+10	4.78E+08 [t]

D

Neural Networks Models Results

Table D.1: The performance (MASE) of all neural network machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)					
Model	**Feature Selection**	**H1**	**H2**	**H3**	**AVG**	**H4**	**CV**
MLP	RFE ENET	0.3506	0.3357	0.3103	0.5593	0.3890	0.2334
LSTM	RFE ENET	0.4532	0.5024	0.4684	0.6563	0.5201	0.2308
LSTM	RFE LASSO	0.5244	0.4961	1.2502	0.5187	0.6973	0.2308
MLP	RFE LASSO	0.6188	0.5982	1.1594	0.5895	0.7415	0.2422
MLP	MIG	0.4745	0.6219	1.3744	0.5680	0.7597	0.3642
GRU	RFE LASSO	0.7624	0.7359	1.1550	0.6091	0.8156	0.2308
BiLSTM	RFE LASSO	0.7704	0.7528	1.1957	0.6160	0.8337	0.2308
GRU	PCA	1.2858	1.1915	1.3790	0.5269	1.0958	0.2308
BiLSTM	RFE RIDGE	1.8084	0.9594	1.1880	0.6101	1.1415	0.2308
LSTM	RFE RIDGE	1.8254	0.9712	1.1929	0.6137	1.1508	0.2308
LSTM	UNIQUE NO CORR	2.2450	1.2600	1.4842	0.6918	1.4203	0.2308
LSTM	PCA	2.2472	1.2718	1.4856	0.6929	1.4244	0.2308
LSTM	VARIANCE UNIQUE NO CORR	2.2611	1.2681	1.5137	0.6917	1.4337	0.2308

Model	Feature Selection	H1	H2	H3	AVG	H4	CV
BiLSTM	PCA	2.2751	1.2854	1.5001	0.6950	1.4389	0.2308
BiLSTM	NO CORR	2.2908	1.2867	1.5211	0.6940	1.4482	0.2308
LSTM	NO CORR	2.2905	1.2889	1.5365	0.6905	1.4516	0.2308
MLP	EM ENET	1.7497	1.3541	1.9410	0.8138	1.4646	0.2531
BiLSTM	VARIANCE UNIQUE NO CORR	2.3242	1.3114	1.5455	0.6990	1.4700	0.2308
MLP	XGBOOST	1.3883	1.7957	2.4384	0.8079	1.6076	0.2427
GRU	RANDOM FOREST	1.5507	1.8060	2.4502	0.6716	1.6196	0.2308
LSTM	XGBOOST	1.8325	1.9168	2.4662	0.6861	1.7254	0.2308
MLP	PCA	2.4854	1.7360	2.1971	0.6745	1.7733	0.3325
GRU	XGBOOST	2.0678	2.0708	2.6379	0.7130	1.8724	0.2308
LSTM	RANDOM FOREST	2.2393	2.0769	2.6850	0.7461	1.9368	0.2308
MLP	EM LASSO	2.6963	2.0294	2.3556	0.8429	1.9811	0.2091
BiLSTM	XGBOOST	2.2832	2.1579	2.7418	0.7488	1.9829	0.2308
MLP	VARIANCE UNIQUE	2.9352	1.8948	2.3398	0.8119	1.9954	0.3549
MLP	MIG NO CORR	2.9628	2.0551	2.5030	0.7859	2.0767	0.2326
GRU	RFE ENET	2.9423	2.1202	2.3115	1.0901	2.1160	0.2308
MLP	RANDOM FOREST	1.9689	2.4348	3.1540	0.9697	2.1319	0.2418
MLP	UNIQUE NO CORR	3.0931	2.1988	2.6144	0.7449	2.1628	0.2827
GRU	EM LASSO	3.1402	2.3381	2.6420	0.7074	2.2069	0.2308
MLP	VARIANCE UNIQUE NO CORR	3.1537	2.2789	2.6669	0.7837	2.2208	0.2827
GRU	EM ENET	3.5442	2.2730	2.4402	0.7886	2.2615	0.2308
MLP	NO CORR	3.2667	2.3115	2.7249	0.7710	2.2685	0.2792
GRU	MIG	3.3224	2.3890	2.8715	0.7763	2.3398	0.2308
BiLSTM	RANDOM FOREST	3.1555	2.4801	3.0146	0.8250	2.3688	0.2308
LSTM	EM LASSO	3.5896	2.4893	2.6564	0.7585	2.3734	0.2308
MLP	NO FS	3.5321	2.3897	2.9037	0.9578	2.4458	0.3503
BiLSTM	RFE ENET	3.6525	2.5080	2.5712	1.1411	2.4682	0.2308
GRU	MIG NO CORR	3.8485	2.4136	2.9917	0.7916	2.5114	0.2308
LSTM	EM ENET	3.9699	2.5640	2.7019	0.8228	2.5146	0.2308
MLP	ANOVA	3.7443	2.6076	2.8772	0.9459	2.5437	0.3174

Model	Feature Selection	H1	H2	H3	AVG	H4	CV
MLP	VARIANCE	3.7665	2.4926	2.9800	0.9458	2.5462	0.3548
BiLSTM	EM ENET	4.0196	2.6147	2.7700	0.8189	2.5558	0.2308
GRU	RFE RIDGE	4.0163	2.4998	2.9716	0.7910	2.5697	0.2308
MLP	Unique	3.9511	2.5978	3.0533	0.9487	2.6377	0.3790
LSTM	MIG NO CORR	4.1863	2.5870	3.0899	0.8370	2.6751	0.2308
BiLSTM	MIG	4.1278	2.7205	3.1279	0.8551	2.7078	0.2308
BiLSTM	EM LASSO	4.2438	2.8296	2.9936	0.8136	2.7201	0.2308
BiLSTM	ANOVA	4.2734	2.7012	3.0788	0.8566	2.7275	0.2308
MLP	RFE RIDGE	4.4949	2.8337	2.9893	0.8906	2.8021	0.4399
BiLSTM	MIG NO CORR	4.4823	2.8160	3.2555	0.8792	2.8583	0.2308
LSTM	ANOVA	4.5632	2.8909	3.2471	0.8836	2.8962	0.2308
LSTM	VARIANCE UNIQUE	4.5611	2.8923	3.2671	0.8925	2.9032	0.2308
LSTM	MIG	4.4580	2.9334	3.3437	0.8854	2.9051	0.2308
BiLSTM	VARIANCE UNIQUE	4.6168	2.9243	3.2975	0.8991	2.9344	0.2308
GRU	ANOVA	4.6442	2.9502	3.3164	0.9077	2.9546	0.2308
BiLSTM	UNIQUE NO CORR	4.6573	2.9455	3.3424	0.8989	2.9610	0.2308
GRU	VARIANCE	4.6769	2.9737	3.3626	0.9043	2.9794	0.2308
GRU	VARIANCE UNIQUE NO CORR	4.7067	2.9744	3.3648	0.8993	2.9863	0.2308
GRU	Unique	4.7107	2.9913	3.3695	0.9078	2.9948	0.2308
GRU	VARIANCE UNIQUE	4.7377	3.0134	3.3968	0.9091	3.0142	0.2308
GRU	NO CORR	4.8002	3.0407	3.4419	0.9113	3.0485	0.2308
LSTM	Unique	4.7918	3.0524	3.4422	0.9124	3.0497	0.2308
LSTM	NO FS	4.8146	3.0665	3.4570	0.9146	3.0632	0.2308
BiLSTM	Unique	4.8402	3.0807	3.4696	0.9179	3.0771	0.2308
GRU	NO FS	4.8445	3.0848	3.4743	0.9187	3.0806	0.2308
LSTM	VARIANCE	4.8848	3.1171	3.5127	0.9207	3.1088	0.2308
GRU	UNIQUE NO CORR	4.8942	3.1155	3.5165	0.9221	3.1121	0.2308
BiLSTM	VARIANCE	5.0066	3.1962	3.5953	0.9329	3.1828	0.2308
BiLSTM	NO FS	5.0472	3.2228	3.6222	0.9368	3.2072	0.2308 [t]

E

Decision Trees Models Results

Table E.1: The performance (MASE) of all neural network machine learning models over four horizons: 4-months (H1), 8-months (H2), 12-months (H3), and 24-months (H4).

		Mean Absolute Scaled Error (MASE)					
Model	**Feature Selection**	**H1**	**H2**	**H3**	**AVG**	**H4**	**CV**
RFR	RFE LASSO	0.4243	0.4859	0.3843	0.4828	0.4443	0.1917
RFR	RFE ENET	0.4491	0.6190	0.4831	0.5629	0.5285	0.1889
RFR	EM ENET	0.4683	0.7914	1.2908	0.5654	0.7790	0.1683
XGB	RFE ENET	1.5102	0.8208	0.9413	0.5999	0.9680	0.4662
RFR	EM LASSO	0.5893	1.1111	1.7803	0.6406	1.0303	0.1544
RFR	MIG NO CORR	0.5481	1.2367	1.8036	0.6277	1.0540	0.1831
RFR	ANOVA	1.1570	1.2212	1.5935	0.6393	1.1528	0.2049
XGB	RFE LASSO	1.7982	1.0217	1.1609	0.6309	1.1529	0.4622
XGB	EM ENET	1.8226	1.0635	1.2598	0.6460	1.1980	0.4521
RFR	NO CORR	0.5195	1.4603	2.1599	0.7090	1.2122	0.2006
RFR	UNIQUE NO CORR	0.5391	1.4637	2.1643	0.7099	1.2192	0.2041
RFR	VARIANCE UNIQUE NO CORR	0.5279	1.4650	2.1700	0.7144	1.2193	0.1980
XGB	ANOVA	1.9029	1.2378	1.5373	0.6814	1.3398	0.4524

Model	Feature Selection	H1	H2	H3	AVG	H4	CV
XGB	RANDOM FOREST	1.9640	1.2263	1.5530	0.6677	1.3527	0.4487
XGB	EM LASSO	1.9607	1.2845	1.5534	0.6943	1.3732	0.4534
XGB	UNIQUE NO CORR	2.0054	1.2961	1.5582	0.6923	1.3880	0.4662
XGB	VARIANCE UNIQUE NO CORR	2.0553	1.3267	1.6104	0.6957	1.4220	0.4692
XGB	MIG NO CORR	2.0025	1.3676	1.6621	0.6908	1.4307	0.4605
RFR	MIG	1.5276	1.5262	2.0046	0.7014	1.4399	0.1772
XGB	XGBOOST	2.0469	1.4413	1.7443	0.7107	1.4858	0.4588
RFR	Unique	1.8517	1.5813	1.9543	0.7009	1.5220	0.2264
XGB	VARIANCE UNIQUE	2.1244	1.4620	1.8238	0.7280	1.5345	0.4609
RFR	VARIANCE	1.8207	1.6053	1.9988	0.7136	1.5346	0.2221
XGB	MIG	2.2547	1.4111	1.7803	0.7198	1.5415	0.4543
XGB	VARIANCE	2.1383	1.4982	1.8344	0.7386	1.5523	0.4434
XGB	Unique	2.2056	1.5164	1.8091	0.7276	1.5647	0.4607
RFR	NO FS	1.9403	1.6448	2.0363	0.7206	1.5855	0.2217
RFR	VARIANCE UNIQUE	1.8566	1.7074	2.1249	0.7337	1.6056	0.2259
XGB	NO CORR	2.3682	1.5254	1.8119	0.7179	1.6058	0.4641
XGB	NO FS	2.4559	1.6122	1.9272	0.7555	1.6877	0.4625
RFR	XGBOOST	1.8288	1.9497	2.4872	0.7893	1.7637	0.1603
RFR	RANDOM FOREST	1.9689	1.9628	2.4702	0.7934	1.7988	0.1699
RFR	PCA	1.9806	1.9741	2.5337	0.7838	1.8180	0.2212
XGB	PCA	2.7753	1.7325	2.0077	0.7689	1.8211	0.4868
XGB	RFE RIDGE	3.1299	1.8813	2.1499	0.7989	1.9900	0.5272
RFR	RFE RIDGE	3.9995	2.4880	2.9659	0.8721	2.5814	0.2735 [t]