Application of Deep Learning to Forecast the South African Unemployment Rate: A Multivariate Approach

Rudzani Mulaudzi
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
0601737r@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—Univariate models have been used successfully to forecast unemployment rates across the world. However, these models are inherently limited as they only use past values of a variable to forecast its future movements: ignoring the influence of external factors. Hence, multivariate models were introduced that generalize univariate models. The most commonly used version of multivariate models for unemployment rate forecasting is the vector autoregression (VAR) model. However, this model requires data to meet particular requirements before use, such as being white noise generated and stationary. This requirement adds a significant amount of overhead in preparing the data for use in the model. The model itself also assumes that features equally impact each other. Therefore, it does not enable the identification of the most impactful features. This model was applied to multivariate data from the South African Reserve Bank, and the model had an error rate twenty times higher than machine learning techniques. LSTM and GRU had the lowest error rate on the same data set with LASSO and elastic net identifying domestic output and government expenditure as important predictors of unemployment.**

***Index Terms*—Forecasting, Machine Learning, Vector Autoregression, Unemployment**

I. Introduction

The South African unemployment rate is currently 30% [1]. Economists speculate that due to the Covid-19 lock-down regulations, this rate is likely to exceed 50% [2]. A high unemployment rate is considered by many to be a sign of an unhealthy economy [3]. Hence, there is a significant interest in forecasting unemployment rates across the world. Corporations often use these forecasts to decide which countries to target for expansion purposes.

Unemployment is typically forecasted using univariate methods, which only look at the past movement of the unemployment rate to predict its future movements. The most common techniques are the Autoregressive Integrated Moving Average (ARIMA) and Holt-Winters [4]–[6]. These are used to predict unemployment rates across the world [5], [7], [8]. However, these methods have limitations because they only rely on the past values of the target variable to predict its future values. Ignoring other variables that might influence the variable of interest.

Multivariate models were introduced to generalize the univariate approaches: allowing for multiple variables to be considered for forecasting purposes. These models allow the influence of variables on each other to be captured without losing the auto-correlation effects. The most common model used to forecast unemployment rates is the vector autoregression (VAR) [4], [7]. The VAR is used as a benchmark model for unemployment rate multivariate forecasts [4], [7].

Although the VAR approach enables the inclusion of variables that impact the target variable into the model, they still have other limitations. According to [5], these models are better suited for linear data. Their forecast accuracy diminishes when the data is nonlinear, as is the case with most economic variables. The models require the input data to be white noise generated, symmetric, and stationary, requiring extensive data transformations to meet these requirements [5]. Furthermore, these models are not suitable for big data settings as they prefer data with few variables that are often selected through expert knowledge.

This research shows that machine learning models, especially deep learning models, overcome the challenges associated with univariate and multivariate models. Concretely, this research i) demonstrates that machine learning models can forecast the South African unemployment rate with greater accuracy than VAR (the benchmark model) and ii) show that machine learning models provide additional benefit as they allow feature selection to be done using data, therefore, not requiring prior expert input or an economic theory as is the case with VAR.

The research approach used by this paper was exploratory, where regression models, kernel-based models, and neural networks were applied to data from the South African Reserve Bank. Deep learning models had the lowest error rates. The data comprised of 147 features with mixed frequencies, the South African unemployment rate, which was sourced from Bureau of Economic Research (BER), being the target variable.

The rest of this paper is structured as follows. Section II discusses the VAR model, a multivariate statistical model, and

deep learning models, a branch of machine learning models used to forecast unemployment rates worldwide. Section III discusses the results that were achieved by applying the VAR and machine learning models to forecast the South African unemployment rate. Section IV is the final section, which discusses the contribution of this research as well as opportunities for future research.

II. Related Work

A. Vector Autoregression Model

Multivariate models are generalizations of univariate models. They allow for the forecasting of multiple variables simultaneously. The Vector Autoregression (VAR) model is discussed in this section. This model is often used as a benchmark model for forecasting unemployment rates across the world [4].

The VAR model was used by [4] to model unemployment trends in the United States of America (USA). They found that the model was able to achieve higher accuracy rates than univariate approaches. The mathematical representation is shown in Equation 1 [5].

$$\begin{aligned} y_{1,t} &= c_1 + \theta_{11,1} y_{1,t-1} + \theta_{12,1} y_{2,t-1} + e_{1,t} \\ y_{2,t} &= c_2 + \theta_{21,1} y_{1,t-1} + \theta_{22,1} y_{2,t-1} + e_{2,t} \end{aligned} \tag{1}$$

where, $y_{i,t}$ is one of the predictors of the target variable at time t. Fundamental to how VAR works, is that it assumes that features influence each other equally. The coefficients, $\theta_{ii,l}$ and $\theta_{ij,l}$, captures the influence of the l-th lag variable of y_i on y_i and of y_i on y_j, respectively. The error term is $e_{i,t}$ which is a white noise process.

The VAR model can only model stationary data. Therefore, if the data is not stationary, it must be differenced d-times before being used in the VAR model [5]. This requires each feature to be differenced until the entire data set is stationary. The VAR model iteratively forecasts each feature. Another limitation of VAR is that its performance deteriorates with longer forecasting horizons and lag orders.

Furthermore, the VAR requires extensive manipulation of the input data in order to accommodate nonlinear data. This requires the researcher to have extensive domain knowledge and data analysis skills to engineer features suitable for the nonlinear requirements [7]. The Seasonal Additive Nonlinear VAR (SANVAR) was introduced by [9] to address the challenges mentioned earlier with VAR. However, the model still requires extensive skills to set-up and deploy. A 'cousin' of VAR is the threshold autoregressive (TAR) model that provides the ability to model nonlinear data to the autoregressive family of models. TAR was successfully used to forecast unemployment rates in regions where asymmetries in the data were visible [4]. However, these models impose similar data requirements to the VAR on the input data [4].

Due to these challenges and limitations with the VAR model (and its variations), several researchers are experimenting with machine learning approaches as alternative approaches for multivariate unemployment rate forecasting. These approaches include regression, kernel, tree-based techniques, and neural networks.

B. Deep Learning

Although there are multiple branches of machine learning techniques, this sub-section focuses on deep learning techniques. The experimental results show that these models are the most successful in modeling the South African unemployment rate: the interest of this paper.

Neural networks are biologically inspired machine learning models that use concepts relating to how the brain works. According to [10], Frank Rosenblatt was the first to introduce an early version of a neural network, called a perceptron in 1957. The perceptron is a single layer neural network with three parts: input values, weights & bias, and an activation function. The output is calculated as $y = f(\sum_{i=1}^{n} w_i x_i + \theta)$, where x_i is the input (a neuron from layer 1), w_i a weight that enables the function f to map the input to the output. The function f is an activation function that ensures that the output is either 0 or 1, with 0 being an accurate classification and 1 an incorrect one. The perceptron is the foundational structure for modern deep 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. They can function with incomplete or imperfect data, unlike autoregressive models [3]. Neural networks have been described as universal function approximators, making them suitable when forecasting nonlinear time series data [10]. In 1996, it was demonstrated for the first time that neural networks could produce significantly better unemployment forecasting estimates than autoregressive forecasting models, i.e. VAR and TAR models [3]. Since [3], several other researchers have investigated various neural network architectures for forecasting unemployment. These are discussed in this subsection.

1) Feedforward Neural Networks: Feedforward neural networks (FFNNs), also referred to as a multilayer perceptron (MLP), are neural networks where the data flows from input to output in a 'forward' direction without moving backward [10]. Their goal is to find a predictor function, f, which maps an input x to some output y i.e. $y = f(x)$ [10]. FFNNs are the foundations of more advanced neural networks. They effectively model nonlinear relationships, which is important for economic data as this data is often nonlinear.

[7] 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. The researchers also demonstrated that FFNNs offer a significant improvement when forecasting unemployment rates compared to autoregressive models.

In some cases, when the FFNN is too simple - single hidden layer, ten or fewer nodes, with a sigmoid activation - it can be outperformed by autoregressive models [8]. [11] also adds that large data sets are often required for neural networks to outperform autoregressive models. Therefore, it is clear that

deep neural networks improve the performance of time series forecasts when they have an appropriate architecture and are trained on sufficient data.

2) Recurrent Neural Networks: Recurrent neural networks (RNNs) are a type of neural network architectures that enable effective modeling of sequences of data as they possess an internal memory structure [10]. At their core, they are essentially FFNNs with feedback loops [10]. RNNs were developed for natural language processing requirements such as language translation. [7] demonstrated that these models could forecast unemployment accurately by using a variation of RNNs referred to as long short-term memory (LSTM) neural networks. The LSTM model outperformed the autoregressive models in an experiment to forecast unemployment rates in the USA, demonstrating their long term memory capabilities.

3) Convolutional Neural Networks: Convolutional neural networks (CNNs) are a class of neural networks used primarily for image classification [10]. 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 in the USA. They were found to perform poorly compared to other neural network architectures but with higher accuracy rates than autoregressive models [7]. The results are consistent with the intention of the architecture, which is to extract features from images and not perform regression type activities. It is worth noting that there is a developing literature of 1-dimensional CNNs referred to as temporal CNNs for time series forecasting.

This section provided an overview of the multivariate and machine learning models that have been employed to forecast unemployment rates across the world. The next section discusses results from the application of machine learning models in South Africa.

III. Results and Discussions

This research was a multivariate analysis. The approach was exploratory, with previous research being used as a guide. The performance of the various models was measured using the mean absolute scaled error. The following subsection provides the results of the analysis.

A. Performance Measures

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) [4], [5], [7], [8]. RMSE is used to compare different models run on the same data set. Whereas MAPE is used to compare models run on the same or different data sets: this is possible because MAPE is not a scale-dependent model [5]. Therefore, for this research, MAPE would have been an appropriate performance measure as it allows models run on different data sets to be compared. However, the challenge 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 actual value is 10 and forecast value is 15 then MAPE would be (10-15)/10 * 100 = 50%, as opposed to the case where actual is 15 and forecast is 10, MAPE would be (15-10)/15 * 100 = 33% [5], [12]. Therefore, [5], [12], proposed 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 2:

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

where, y is the actual value whilst you $\hat{y}$ is the forecast value.

B. Data Preparation

The data was accessed from the South African Reserve Bank (SARB). The SARB database contains 147 macroeconomic variables [13]. The database covers all critical sectors of the South African economy. Data wrangling techniques were used to consolidate the 147 features into a single structured table. The data had 794 observations from January 1970 to December 2019. However, due to the data having different frequencies, monthly and quarterly, the merged data set had a significant amount of missing data. All the quarterly observations had missing data for each month that they did not report while the monthly observations did.

Missing data imputation techniques were employed to address the missing data. Six data imputation strategies were explored [15]:

- 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 data imputation strategy resulted in lower error rates than the other imputation strategies. The 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 prepared data was split into train, evaluation, and test sets: 746 observations were used for training, while the evaluation and test set each had 24 observations. The evaluation set was used to optimize the trained models, with the test set reversed for testing the model on data never seen before.

C. Analysis and Results

Nine different machine learning models were run, five regression techniques (linear regression (LR), least absolute shrinkage selector operator(LASSO), ridge, elastic net (ENet) and bayesian ridge), three deep learning models (multilayer perceptron (MLP), long short-term memory (LTSM), and gated recurrent unit (GRU)), and support vector regression (SVR). The vector autoregression (VAR) was used as the baseline model.

The original data set had 147 features. This feature set was reduced using feature selection techniques. According to [14], 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 data set first then evaluates each subset afterward. Therefore, each possible subset is evaluated where else embedded methods progressively eliminate some subsets: hence embedded methods are 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;

- removal of correlated features (referred to in this paper as 'no correlation'),
- analysis of variance (referred to in this paper as 'ANOVA'),
- mutual information gain (referred to in this paper as 'MIG'), and
- removal of variables with low variance (referred to in this paper as 'variance').

Along with the four feature selection methods, the deletion of duplicated features was treated as a feature selection technique (referred to in this paper as 'unique'). The four feature selections were also combined to form a chain of filter methods e.g. 'unique, no correlation' or 'variance, unique, and no correlation'. The two embedded methods were LASSO and elastic net.

The recursive feature selection wrapper method was considered. However, it was computationally inefficient as it 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.

Table III-C shows that the removal of correlated features resulted in the lowest error rates of all the models that were run. It is important to note that the table shows results where the missing values were addressed through the last known value imputation strategy. This strategy produced the lowest error rates across all data sets when compared to the five other strategies considered.

Table III-C shows that the deep learning techniques resulted in the lowest error rates as measured by MASE. Specifically, the LSTM and GRU models. These models out performed all the other models across all data sets, some of models and results are shown in the table. The table shows the top-performing models for each machine learning technique that was used for this research.

Both the LSTM and GRU 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. The LSTM had, on average 54 691 parameters with GRU having 42 014 parameters. The other machine learning techniques had an average of 83 parameters. This observation shows that the South African unemployment rate is relatively complex and cannot be easily forecasted by simple models. Furthermore, the comparing performance of the MLP with the LSTM and GRU shows that the data set had long temporal dependencies that cannot be ignored when modeling the South African unemployment rate. This was also confirmed by analyzing the auto-correlation and partial-correlation functions of several features. In addition to this temporal nature, the data was nonlinear, confirmed by the unreliability of the R-squared measure, which had extremely low values and values outside of the acceptable range.

LASSO was the top-performing model outside of the deep learning models. The model is sparse. Its application of the $L1$-regularisation drove most of the features to zero except for four features: 'Final consumption expenditure by general government', 'domestic output: all groups', 'consolidated general government: liabilities', and 'total outstanding domestic non-marketable loan'. ENet also selected the 'domestic output: all groups' and 'Final consumption expenditure by general government' as key features. In both models, these two features had the highest coefficients. ENet also selected 'foreign exchange rate: SA rand per USA dollar', 'Domestic: currency and deposits', and 'total outstanding domestic non-marketable loan levies'. The relatively high performance of LASSO and ENet compared to linear regression (LR), and ridge regression (RR) shows that there is a distinct subset of features that influence the South African unemployment rate more than others.

Therefore, the economic performance (domestic output: all groups) and the consumption patterns of the government (Final consumption expenditure by general government) are the two of the most important features that determine the South African unemployment rate. Mutual information gain and ANOVA analysis also confirmed these observations.

The VAR was used as a benchmark model, table III-C shows that the VAR performed extremely poorly compared to the machine learning models. 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.
The 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 non-linearity of the data: the unemployment rate itself is also nonlinear. The VAR model was only implemented for one data set, the 'variance MIG' data set. The data set had features with low variance removed, and mutual information gain applied to select features. The data set had 15 features. This data set is one of the 89 (with last known value missing data imputation) that the machine learning models were applied to. The other data sets did not work because these resulted in matrices that were singular or not positive definite, which is required when implementing the VAR model.

Figure 1 and figure 2 shows the forecasts over the evaluation period. The LSTM and GRU are very close to each other, while LASSO is more dynamic in capturing the unemployment rate. VAR undershoots its forecasts by a significant margin.

It is worth noting that the errors on LSTM, GRU and LASSO models were all bi-modal, confirming the nonlinearity and non-stationarity of the South African unemployment rate.

TABLE I
PERFORMANCE RESULTS OF MACHINE LEARNING MODELS TO FORECAST THE SOUTH AFRICAN UNEMPLOYMENT RATE. THE VAR IS USED AS A BENCHMARK MODEL AND DEEP LEARNING MODELS DEMONSTRABLY OUTPERFORM ALL OTHER MODELS.

Performance evaluation of machine learning models			
Rank	***Model***	***MASE***	***Feature Selection Technique***
1	LSTM	0.914	Unique, no correlation
2	GRU	0.915	No correlation
20	LASSO	1.211	No correlation
23	ENet	1.217	Unique, no correlation
38	SVR	1.345	Unique, no correlation
43	Ridge	1.375	ANOVA
45	Bayes Ridge	1.410	Unique, no correlation
58	MLP	1.582	Variance, unique, no correlation
58	LR	1.582	Variance, unique, no correlation
-	VAR	26.260	Unique, no correlation, MIG*
-	VARMA	34.775	Unique, no correlation, MIG*

*Mutual Information Gain

IV. CONCLUSIONS

[16] successfully demonstrated that the South African unemployment rate can be forecast using univariate models. However, these models are limited because only the past of a variable is considered in forecasting future values. Therefore, multivariate models such as VAR are used to capture the influence of other factors. VAR is the most commonly used traditional statistical model for forecasting unemployment rates. However, this model is limited because it only works with stationary data and is not suitable for nonlinear data. Computationally, the model is not suitable when there are too many features, and, therefore, it typically would require prior knowledge to determine which features to include in the model or not. Furthermore, the model assumes that each feature is equally important in predicting the target, making it impossible to determine which features are most important.

Machine learning models offer a way to forecast the South African unemployment using multivariate data overcoming the challenges associated with the VAR model. This paper demonstrated that machine learning models are more accurate than the VAR model when forecasting the South African unemployment rate using multivariate data. As part of the machine learning pipeline, a data-driven approach was used to select features to include in the modeling: removing correlations proved to be the most impactful in improving accuracy rates. Similar feature selection is done using expert knowledge or economic theory in VAR models.

As the machine learning models do not require extensive data transformations to ensure that the data meet particular requirements, it was relatively easy (compared to VAR) to model non-stationary data.

Deep learning models were the most accurate models. These models had four hidden models with dropout and early stopping. LASSO and Elastic Net identified domestic output and government expenditure as key features that influence the South African unemployment rate.

This research has demonstrated the impact of machine learning models for multivariate forecasting unemployment rates, showing that these models can assist in identifying meaningful features. In future research, Bayesian approaches should be explored for both the data imputation and forecasting. These models are suitable because the South African economic data relating to unemployment is very limited. There is also often missing data, a challenge that Bayesian models are very suitable for.

ACKNOWLEDGMENT

This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835).

REFERENCES

[1] Statistics South Africa (StatsSA), "Mbalo briefing May 2020," Statistics South Africa, May 2020.
[2] National Treasury, "Briefing by National Treasury on financial implications of Covid-19 on both the economy and budget: JT standing committee and select committee on finance and appropriations," National Treasury, 2020.
[3] M. Aiken. "A neural network to predict civilian unemployment rates," Journal of International Information Management, Volume 5: Issue 1, Article 3, 1996.
[4] A. L. Montgomery, V. Zarnowitz, R. S. Tsay, and G. C. Tiao, "Forecasting the U.S. unemployment rate," Journal of the American Statistical Association, 93(442):478, June 1998..
[5] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice. OTexts," Melbourne, Australia, 2nd Edition, 2018.
[6] C. Brooks, "Introductory econometrics for finance," Cambridge UniversityPress, 3rd edition, 2014.
[7] R. R. Cook and A. S. Hall. "Macroeconomic indicator forecasting with deep neural networks," Federal Reserve Bank of Kansas City, Research Working Paper 17-11, September 2017.
[8] C. Katris. "Prediction of unemployment rates with time-series and machine learning techniques," Computational Economics, 2019.
[9] L. Yang, "Nonparametric modelling of quarterly unemployment rates," Unpublished Technical Report, 2007.
[10] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," MIT Press, 2016.
[11] V. Cerqueira, L. Torgo, and C. Soares, "Machine learning vs statistical methods for time series forecasting: size matters", arXiv, 2019.
[12] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," Unpublished, November 2005. Available: https://robjhyndman.com/papers/mase.pdf

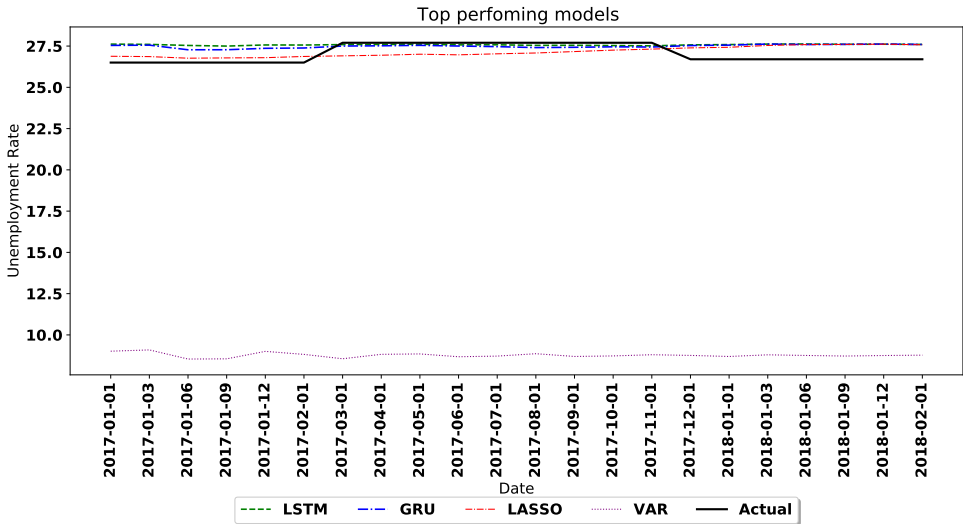

Fig. 1. The performance of LSTM, GRU, LASSO and VAR are the displayed. VAR centered its prediction around 8.5 and therefore undershot in its predictions.

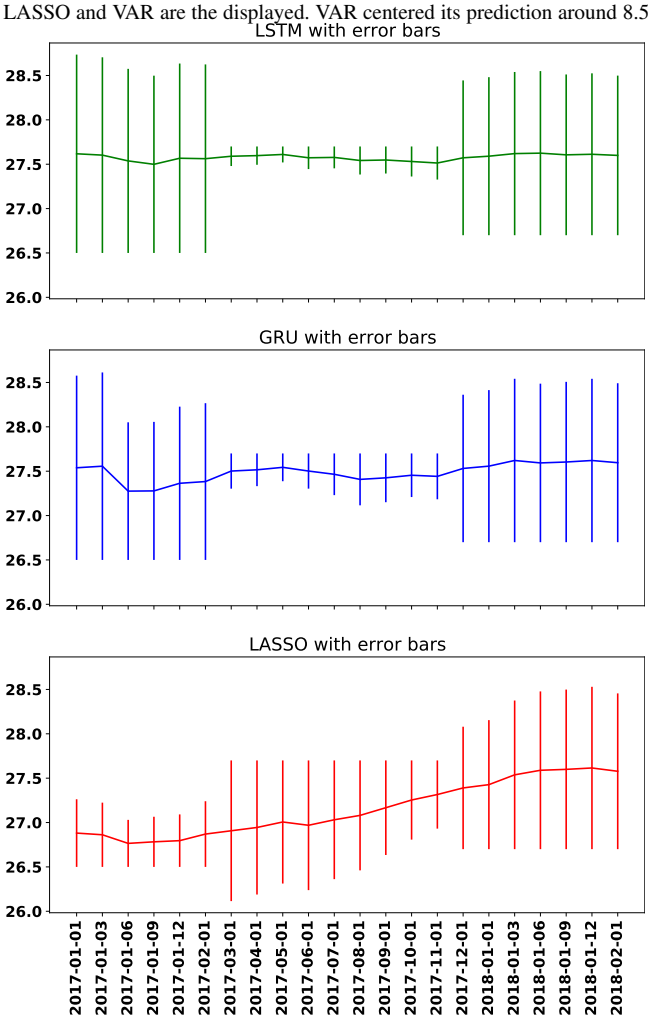

Fig. 2. The errors of LSTM, GRU, and LASSO follow a bi-modal distribution. Which confirms the non-stationary nature of the South African unemployment rate.

[13] South African Reserve Bank (SARB),"Economic and financial data for South Africa," South African Reserve Bank. Available: https://wwwrs.resbank.co.za/webindicators/EconFinDataForSA.aspx.

[14] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers and Electrical Engineering, Volume 40, Issue 1, January 2014.

[15] D. Bertsimas, C. Pawlowski and Y. Zhuo, "From predictive methods to missing data imputation: an optimization approach," Journal of Machine Learning Research, 1- 39, 2018.

[16] R. Mulaudzi and R. Ajoodha, "An exploration of machine learning

models to forecast the unemployment rate of South Africa: a univariate approach," In Press, 2020.