Demonstration that the Use of Feature Selection on High Dimensional South African Macroeconomic Data Results in Improved Performance with Lower Compute Requirements

Rudzani Mulaudzi and Ritesh Ajoodha

The University of the Witwatersrand, Johannesburg, South Africa,
556245@students.wits.ac.za,
WWW home page: https://www.wits.ac.za/

Abstract. As it stands, feature selection in macroeconomic forecasting is arbitrary or based on well-established economic theories. This approach limits the usage of end-to-end data-driven approaches, despite evidence pointing to many potential advantages. This study explores this by using data from the South African Reserve Bank (SARB) with 147 macroeconomic variables: the South African unemployment rate being the target variable. Seventeen feature selection techniques were employed covering filter, wrapper, and embedded methods. The results from this research show that data-driven feature selection offers results that are significantly better than naïve approaches on non-stationary high dimensional macroeconomic data. More specifically, it demonstrates how feature selection techniques can improve macroeconomic machine learning forecasting performance by over 15%, reducing the space and time resource requirements by almost 80% for space and up to 50% for time. Filter methods had the highest accuracy improvement: with almost a 30% accuracy improvement. Wrapper and embedded methods achieved 21% and 15%, respectively. These results show that there are a handful of features that can accurately forecast the South African unemployment rate. In addition, the reduced computational time and resource requirements to deploy machine learning models point to a need for researchers to employ more data-driven methodologies in forecasting macroeconomic variables.

Keywords: Machine Learning, Feature Selection, Forecasting, Unemployment.

1 Introduction

This research attempts to show how feature selection techniques can impact the accuracy and computational resources required to forecast the South African unemployment rate. The research outputs can be used by policymakers to determine how best to address unemployment in the country. Unemployment is

the biggest economic and social challenge in the South Africa. The national unemployment rate is at 30.1% as at June 2020 [1]. The lockdown regulations of COVID-19 are estimated to increase this to above 50% [2].

This research used data from the South African Reserve Bank (SARB), where the South African unemployment rate was used as the label, and 147 macroeconomic variables from the SARB as features [3]. These features cover key sectors of the South African economy: real, fiscal, financial, and external sectors, as well as population data. The data was from January 1960 to December 2019: a total of 790 observations.

Similar research typically has two to fifteen features [4,5]. Hence this research is classified as high dimensional due to the number of features. High dimensional data comes with computational challenges, noise, and accuracy challenges.

This research tests the hypothesis,

- **H1**: The application of feature selection techniques on high dimensional macroeconomic data improves the accuracy when forecasting the South African unemployment rate.
- **H0**: The application of feature selection techniques on high dimensional macroeconomic data *does not* improve the accuracy when forecasting the South African unemployment rate.

The hypothesis is tested through various experiments to see if we can accept the hypothesis (**H1**) or reject it in favor of the null hypothesis (**H0**). The outcome contributes to encouraging end-to-end data-driven approaches in macroeconomic forecasting.

The research paper is organized as follows. Section 2 explains what feature selection techniques are. Section 3 discussed how they have been employed in similar prior research. Section 4 discusses the research methodology with section 5 providing the experimental results, and section 6 concluding with recommendations for future work.

2 Background

High dimensional data present multiple challenges for machine learning models. The most prominent being 'the curse of dimensionality'. This curse refers to how increases in features increase the complexity of the machine learning model, furthermore making the problem intractable [6].

Another key issue with high dimensional data is that it is often full of noise. Using the full feature set to train a model can result in a model of limited use because of the inability to distinguish the 'signal from the noise.' This noise further causes the models to overfit, and therefore, these models become of limited use on new data.

Feature selection techniques assist with these problems because they allow us to use only a subset of the feature set in the models [6]. This operation is performed as part of the data preprocessing step in the machine learning pipeline. Besides addressing the problems mentioned, it comes with additional benefits.

The most common benefits being data reduction, which results in computational space and time savings; feature set reduction, which means the future data collection efforts is easier [6,7]; removing noise resulting in a performance (predictive) improvement [6,7]; and because the importance of each feature is transparent there is also an increase in explainability [6,7].

There are three categories of feature selection algorithms: filter, wrapper, and embedded methods [6]. These methods are further grouped into methods that are either univariate or multivariate [7]. Univariate being those that only rely on the relevance of a specific feature to predict the label without consideration of any other features and their interactions with the feature of interest: statistical scoring is often used [7]. Multivariate try to capture the interrelation between the different features and choose a feature subset considering the interrelated benefits [6,7]. Filter methods are generally univariate, whilst wrapper and embedded methods are multivariate.

Filter methods are used in the data preprocessing, and they allocate a statistical score to each feature based on its predictive ability [6]. These methods use techniques such as Pearson Correlation Coefficient (Corr), Mutual Information Gain (MIG), or Analysis of Variance (ANOVA) to determine the relevance of each feature relative to the label [6,7]. In these methods, the features which have the lowest statistical score would typically be discarded, and only those with high scores are kept as they aid in predicting the label more than others. These models are instrumental in discarding noise from the feature set [7]. However, they can prematurely discard important features because even though a feature appears irrelevant on its own, it might be the case that it is relevant when considered in a group with other features [7].

Wrapper methods are multivariate methods that take into account the group dynamics of the features [6]. Although these wrapper methods are utilized at the data preprocessing stage, they have elements of the modeling step as well. In order for these methods to determine which subset (group) of features are the most relevant. Each subset is selected from the feature set and passed into a machine learning model, which then calculates the error of each subset and recommends a particular subset that achieved the lowest error of all the subsets considered [6,7]. Therefore, the method ranks a group of features, not just an individual feature. Wrapper methods employ greedy search strategies which can result in combinatorial exploration when the data is high dimensional [7]. Another challenge with wrapper methods is that the machine learning pipeline is computationally more expensive because compute resources are deployed to select features and again for modeling using that subset of features.

Embedded methods have the benefits of both multivariate and univariate models. Multivariate being the ability to look at features in the context of other features, thus selecting a group. Univariate being the speed and simplicity. These methods intersect the data preprocessing and modeling steps in the machine learning pipeline. Therefore, reducing the process compute time and resources to an extent. They intersect these steps because they evaluate subsets of features for suitability in predicting the target, whilst simultaneously making a prediction.

They are machine learning models that offer predictive capabilities. Wrapper methods do not offer these capabilities as the calculation of errors is a black box to the user [6,7]. In embedded methods such as least absolute shrinkage and selection operator (LASSO), the method uses the L1-norm to regularize the error optimization: driving features with the least predictive power zero, therefore, returning a subset of features for consideration [7].

Selecting the correct feature selection methods is dependent on the case in question. If group dynamics are important, then wrapper and embedded are a better approach. If there is a limited amount of independence amongst the features, then filter methods are appropriate. This section described different feature selection techniques; the next will discuss how these have been applied in forecasting unemployment rates.

3 Related Work

Unemployment forecasting has been of interest to economists and statisticians for decades. The use of machine learning is more recent, with Aiken [4] being the first to employ neural networks to forecast the United States of America (USA) unemployment rate. However, the use of feature selection techniques is employed inconsistently in the forecasting literature. In most cases, the selection of features is either based on established economic theory or based on the researcher's experience and interests [4,8]. This, therefore, limits the utilization of pure machine learning pipelines in these forecasts. Even in cases where feature selection approaches are used, the initial feature set is arbitrarily chosen [5,12]. This section will discuss how feature selection has been used in similar prior research.

3.1 Elastic Net (ENET)

Hall [5] 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 especially 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 [9]. 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.

Hall [5] demonstrated that the ENET could forecast unemployment over 3, 6, 9, 12, and 24-month horizons more accurately than traditional forecasting models and professional forecasters 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. The most important features identified by the ENET model for the USA were: housing, manufacturing, and interest rates. These were selected from an initial feature set of 138 macroeconomic variables.

3.2 Least Absolute Shrinkage and Selection Operator (LASSO)

Kreiner and Duca [10] used LASSO regression to forecast unemployment rates in the USA. This regression model minimizes prediction error through a regularization process. The model penalizes the reliance on particular variables for prediction, therefore, shrinking the number of features [9].

It was demonstrated that LASSO could improve the forecasting accuracy of unemployment in the USA and identify variables that are most important in forecasting unemployment [10]. 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 rates [10]. The initial number of features is not mentioned. However, these were the 'academic' and 'international' variables in the FRED database: a pre-selection of the category feature categories to be considered.

3.3 Principal Component Analysis (PCA)

PCA is a dimension reduction technique, where the number of features are reduced by transforming them onto a lower dimension [6]. Therefore, if there are p features in $\mathbf{R}^p$ space. PCA will reduce these to $\mathbf{R}^q$ where $p >> q$: a lower dimension. The 'features' in the new spaces are referred to as components [6,7].

Kreiner and Duca [10] used PCA on 600 000 features from the FRED database. This resulted in a dimension reduction from 600 000 features to 185 components. These 187 were then modeled using a neural network and LASSO to forecast the USA unemployment rate: resulting in a higher accuracy than traditional statistical methods.

4 Experimental Design

In order to test the stated hypothesis: ***H1****: The application of feature selection techniques on high dimensional macroeconomic data improves the accuracy when forecasting the South African unemployment rate.* Two hundred sixteen (216) experiments were tested with 12 different machine learning models and 17 different feature selection techniques. This section describes how these experiments were conducted.

4.1 Data Preparation

Data was accessed from SARB with 147 features and the South African unemployment rate being the label. The data was from January 1960 to December 2019. The data came with mixed frequencies, and the last known value data imputation strategy was employed. This is a strategy that is commonly employed in macroeconomic forecasting. It also makes logical sense because in macroeconomics, when, for example, GDP is being referenced, it is usually a reference to the last known value of GDP.

The data was then split into training and test, where 24 observations were used for testing and 746 for observations. Previous literature used the similar test sizes.

4.2 Methodology

All three feature selection techniques were used: filter, wrapper, and embedded. These methods cover both the univariate and multivariate approaches. Three univariate techniques were employed, all of which were filter methods: Pearson Correlation Coefficient calculation (Corr), Mutual Information Gain (MIG), and Analysis of Variance (ANOVA). These three are the most common filter methods used for regression purposes [6,7]. Furthermore, duplicate features were also removed from the feature set. This is because features that gave the same information do not improve the ability to model the unemployment rate but consume computational space and time. Low variance features were also removed because this represented features that had too little information to contribute meaningfully to the unemployment rate over the period chosen. This was done using a variance threshold of 5%. Therefore, a total of five univariate techniques were employed.

Eight multivariate techniques were employed, only one of which was a wrapper method. The wrapper method was the recursive feature elimination (RFE). The RFE employs a greedy search strategy to find the optimal subset of features with the low-est performance error [7]. Because of the greedy search strategy's computational demands, the research limited the search to just five features and used regression techniques as the black-box models used in RFE. The embedded methods were LASSO, ENET, Random Forests, and Extreme Gradient Boost (XGBoost). The Random Forest and XGBoost are decision tree approaches that can give a feature relevance score as part of their forecasting activity.

These feature selection techniques were also combined to form feature selection chains, such as a technique with 'Corr applied then MIG' or 'Corr applied and duplicates deleted as well as low variance feature removed.' These combinations resulted in 17 feature selection techniques being employed for this research: listed in Table 1.

Models that have typically been employed in forecasting unemployment rates were used for this research. These models were within the regression, kernel, neural network, and decision tree domains. The aim being to apply these to the SARB feature set to forecast the South African unemployment rate. Then to assess the impact of different feature selection techniques on the outcome. Based on the outcomes, determining if the hypothesis should be accepted or rejected.

A total of 12 models were deployed: elastic net, bayesian ridge regression (Bayes Ridge), LASSO, long-shot term memory (LSTM), gated recurrent unit (GRU), ridge regression (Ridge), support vector regression (SVR), bi-directional LSTM (BiLSTM), random forest regression (RFR), linear regression (OLS), extreme gradient boost (XGB), and multi-layer perceptron (MLP).

Feature Selection	**Number of Features**
No Filter Selection (NO FS)	147
Removal of Duplicates (Unique)	145
Pearson Correlation Coefficient (No Cor)	55
Mutual Information Gain (MIG)	36
Mutual Information Gain and Pearson Correlation Coefficient (MIG No Corr)	14
Variance Threshold (Variance)	145
Variance Threshold and Removal of Duplicates (Variance Unique)	144
Removal of Duplicates and Pearson Correlation Coefficient (Unique No Corr)	54
Variance Threshold, Removal of Duplicates, and Pearson Correlation Coefficient (Variance Unique No Corr)	53
Analysis of Variance (ANOVA)	36
Recursive Feature Elimination with elastic net (ENET) as Black-box estimator (RFE ENET)	5
Recursive Feature Elimination with Ridge as Blackbox estimator (RFE Ridge)	5
Recursive Feature Elimination with LASSO as Blackbox estimator (RFE LASSO)	5
Random Forest	8
Extreme Gradient Boosting (XGBoost)	5
Principal Component Analysis (PCA)	50
Least Absolute Shrinkage and Selection Operator (LASSO)	4
Elastic Net (ENET)	5

Table 1: Feature selection techniques used to test the hypothesis H1.

4.3 Performance Measure

Mulaudzi and Ajoodha [11] propose that mean absolute scaled error (MASE) be used as a performance measure for South African unemployment rate forecasting. Their proposal is because the MASE overcomes asymmetry issues associated with the the mean absolute percentage error (MAPE), which is the typical measure used [12,13]. MAPE penalizes models with negative errors more harshly than those with positive errors [12,13].

The MASE was proposed by Hyndman [12,13] and is shown in equation (1),

$$MASE = \frac{MAE}{MAE_{naive}} \tag{1}$$

where, MAE is the mean absolute error. Therefore, MASE is the MAE of the test set divided by the MAE of the naïve model. A MASE below 1 indicates that a model is more accurate than the naïve model.

5 Results and Discussions

Eighteen feature subsets were analyzed with twelve machine learning models: a total of 216 experiments. These experiments enable a determination of whether the hypothesis should be accepted or rejected. The results are discussed in this section.

5.1 Filter Methods

The filter methods are, on average, 28% more accurate than not using a feature selection method. Eleven of the twelve models benefited from the feature selection methods: improved accuracy and computational time. Table 2 shows the MASE achieved for different feature subsets: the linear regression (ordinary least squared (OLS)) method is omitted from the table due to extremely high errors as it is unable to capture non-linear data.

Bayes Ridge was the only model that was not impacted by the feature selection techniques. The model was the top-performing model across the experiment. Bayes Ridge is a probabilistic model similar to the ridge regression, which uses L2-norm for the regularization purpose. The L2-parameter is set by 'hand' using trial and error or grid search in ridge regression. Bayes Ridge has a similar regularization concept, but the parameter is tuned to the data, unlike ridge regression.

Background of the Bayes ridge model can be found in most introductory texts on probabilistic machine learning, such as Bishops [14]. The model improves with more data because the tuning process is more precise. Given that feature selection reduces data, it is expected that such a model would decline in performance or remain the same.

The univariate feature selection techniques were also joined together to form chain links such as 'Unique, No Corr' or 'MIG, No Corr', which is the removal

	NO FS	UNIQUE	VARIANCE	NO CORR	MIG	ANOVA
ENET	0,60	0,60	0,60	0,43	0,50	0,60
Bayes Ridge	0,43	0,44	0,43	0,48	0,75	0,70
LASSO	0,62	0,62	0,62	0,44	0,52	0,62
LSTM	0,84	0,85	0,83	0,63	0,81	0,81
GRU	0,85	0,84	0,87	0,64	0,77	0,82
Ridge	0,65	0,65	0,65	0,55	0,47	0,72
SVR	0,88	0,90	0,88	1,58	0,58	0,49
BiLSTM	0,86	0,85	0,85	0,77	0,79	0,80
RFR	0,72	0,70	0,73	0,69	0,67	0,70
XGB	0,75	0,73	0,74	0,68	0,70	0,67
MLP	0,80	1,04	0,88	0,74	0,79	1,09

Table 2: The MASE of the univariate filter feature selection methods compared with not applying any feature selection.

of duplications and correlated features or the using mutual information gain to select top features and removing those correlated. These chains had similar results as above, showing that filter methods improve forecasting accuracy in general.

5.2 Wrapper Methods

The wrapper methods were, in general, 21% more accurate when compared to not employing feature selection. Bayes Ridge and ridge regression both declined in performance. This is due to their employment of the L2-norm. Which essentially drives all features close to zero to avoid the model overfitting. As observed under filter methods, Bayes Ridge declines or stays the same when the data is reduced because of how the L2-norm is employed. Table 3 shows the MASE achieved with the wrapper filter methods applied.

5.3 Embedded Methods

Similar to the wrapper and filter methods, it is generally beneficial to employ embedded feature selection methods. Except for the Bayesian ridge because this method is tuned to data, and therefore when there is more data, this tuning improves. The embedded methods had, on average, 15% improvement. Therefore, these methods offered the lowest accuracy improvements compared to wrapper and filter methods.

Table 4 shows the MASE achieved by the models. It is interesting to note that the embedded method that used LASSO and ENET as estimators identified the

	NO FS	**RFE Ridge**	**RFE ENET**	**RFE LASSO**
ENET	0,60	0,44	0,590	0,590
Bayes Ridge	0,43	1,08	0,783	0,783
LASSO	0,62	0,44	0,585	0,585
LSTM	0,84	0,47	0,450	0,454
GRU	0,85	0,72	0,463	0,484
Ridge	0,65	1,06	0,762	0,762
SVR	0,88	1,06	0,786	0,786
BiLSTM	0,86	0,58	0,507	0,504
RFR	0,72	0,66	0,569	0,566
XGB	0,75	0,77	0,605	0,605
MLP	0,80	0,74	0,803	0,773

Table 3: The MASE of the multivariate wrapper feature selection methods compared with not ap-plying any feature selection.

same features as key to predicting the unemployment rate. Hence their MASE is the same.

	NO FS	**EM ENET**	**EM LASSO**	**XGBoost**	**Random Forest**
ENET	0,60	0,60	0,60	0,51	0,53
Bayes Ridge	0,43	0,78	0,78	0,71	0,66
LASSO	0,62	0,62	0,62	0,49	0,52
LSTM	0,84	0,86	0,85	0,65	0,60
GRU	0,85	0,85	0,87	0,65	0,66
Ridge	0,65	0,63	0,63	0,71	0,72
SVR	0,88	0,77	0,77	0,85	0,71
BiLSTM	0,86	0,87	0,85	0,83	0,80
RFR	0,72	0,77	0,76	0,78	0,80
XGB	0,75	0,77	0,77	0,70	0,67
MLP	0,80	0,74	0,75	0,82	0,95

Table 4: The MASE of the multivariate embedded feature selection methods compared with not applying any feature selection.

5.4 Dimension Reduction

PCA provided significant computational improvements, more than the filter, wrapper, and embedded methods. Table 5 and Table 6 shows the time (in miliseconds) and space (number of parameters) that the PCA had relative to the average of the other methods. The time savings are shown in Table 5. There was a significant saving in computation time in a number of cases: 13, 12, and 17 milliseconds for the LSTM, GRU, BiLSTM compared to 15, 15, and 44 milliseconds when PCA was not used. The space savings were better than not using a

feature selection technique or using filter techniques. Wrapper and Embedded methods, however, offered more savings.

	NO FS	**Filter**	**Wrapper**	**Embedded**	**PCA**
ENET	0,00868	0,0045	0,0022	0,0040	0,0028
Bayes Ridge	0,02612	0,0132	0,0037	0,0034	0,0064
LASSO	0,00771	0,0042	0,0022	0,0043	0,0026
LSTM	15,75152	15,3353	17,5586	15,4537	12,6019
GRU	15,31274	15,0273	17,7250	19,4560	11,8673
Ridge	0,00675	0,0063	0,0041	0,0047	0,0038
SVR	0,25344	0,2136	0,0595	0,0637	0,1643
BiLSTM	28,08850	44,5586	83,3304	47,6442	16,8032
RFR	2,56104	1,4069	0,2141	0,8185	1,7519
XGB	0,05348	0,0146	0,0034	0,0092	0,0144
MLP	1,04305	0,8435	0,3477	0,3087	0,7577

Table 5: The computation time, in milliseconds, for different feature selection methods, PCA, and, original data without applying any feature selection.

	NO FS	**Filter**	**Wrapper**	**Embedded**	**PCA**
ENET	147	75	5	17	50
Bayes Ridge	147	75	5	6	50
LASSO	147	75	5	6	50
LSTM	93463	72528	51918	35783	65084
GRU	50326	39054	27956	19268	35045
Ridge	147	75	5	6	50
SVR	71177	52859	34825	35337	46345
BiLSTM	163561	126925	90857	62621	113897
RFR	38744	38744	38744	38744	38744
XGB	147	75	5	21	50
MLP	53769	40030	26505	26889	35145

Table 6: A comparison of the number of parameters in different machine learning methods with feature selection methods, PCA, and, the original data without applying any feature selection.

Feature selection techniques have, in general, a significant impact on the space and time required per model. On average, it was observed that the number of parameters reduced by almost 80% and the computational time by almost 50%.

Even though computational time improvement was higher than other methods (in a number of cases). The MASE was inconsistent across different machine learning models. Table 7 shows that in 50% of cases, MASE without using feature selection was higher than using PCA. Therefore, using PCA with South

African macroeconomic data is not a worthwhile effort. Furthermore, PCA has the additional disadvantage of not being as easily explainable: an important requirement for policy makers.

	NO FS	**PCA**
ENET	0,60	0,51
Bayes Ridge	0,43	0,78
LASSO	0,62	0,52
LSTM	0,84	0,65
GRU	0,85	0,65
Ridge	0,65	0,70
SVR	0,88	0,77
BiLSTM	0,86	0,65
RFR	0,72	0,77
XGB	0,75	0,76
MLP	0,80	0,76

Table 7: The MASE of the multivariate dimension reduction techniques, PCA, compared with not applying any feature selection.

6 Conclusion

Forecasting the South African unemployment rate using macroeconomic variables from the SARB was used to test the hypothesis: ***H1****: The application of feature selection techniques on high dimensional macroeconomic data improves the accuracy when forecasting the South African unemployment rate.* This was tested by running 216 experiments, with twelve machine learning models, and eighteen feature sets. In 98% of the experiments, feature selection techniques improved the accuracy of models and reduced the computational resources required to run these models.

The hypothesis is, therefore, accepted, and the null hypothesis is rejected. The experiments had coverage of all the major feature selection techniques and used a wide selection of machine learning models. Thus, a robust result that gives confidence in accepting the hypothesis: feature selection does improve forecasting accuracy, on average, by more than 15%.

This work demonstrated that feature selection techniques can be leveraged to uncover which features influence the South African unemployment rate. The highest performance (MASE reduction) improvement was achieved through filter methods. This is likely due to the non-stationary nature of macroeconomic variables in the country. Furthermore, filter feature selection approaches are ideal for noise reduction; thus, even though one can model the unemployment rate of South Africa using hundreds of macroeconomic variables, a number of these are likely just to be noise.

It was also seen that PCA does not offer performance improvement even though it offers significant computational savings. PCA is not ideal for forecasting the South African unemployment rate given the need for the country is to forecast the unemployment rate accurately; the computational speed is a secondary concern. This model is also not as easily explainable, which is a quality that is required in macroeconomic forecasting.

Feature selection techniques offer performance and computation improvements when forecasting the South African unemployment rate using high dimensional data. Further work should explore more complex feature selection techniques such as those that combine the statistical techniques used by filter method with grouping effects of multivariate techniques. It should also be noted that not all possible 'chains' of feature selection methods were employed in this research; future work should consider 'chains' that were not explored, such as 'PCA, MIG, and LASSO.'

7 Acknowledgments

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, http://www.statssa.gov.za/?p=13633. Last access 15 December 2020
2. South African National Treasury.: The COVID-19 shock and the revised economic outlook in brief. In: South African National Budget, pp. 21–28. South African National Treasury, Pretoria (2020)
3. South African Reserve Bank, resbank.co.za/webindicators/EconFinDataForSA.aspx. Last access 10 September 2020
4. Aiken, M.: A neural network to predict civilian unemployment rates. Journal of International Information Management **5**(1), (1996)
5. Hall, A.: Machine Learning Approaches to Macroeconomic Forecasting. In: Economic Review Technical Report. Federal Reserve Bank of Kansas City, Kansas (2018)
6. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2013)
7. Guyon, I., Elisseefl, A.: An introduction to feature extraction. In: Feature Extraction, Guyon, I. Nikravesh, M. Gunn, and S. Zadeh, L. (eds.) Studies in Fuzziness and Soft Computing 2006, vol. 207, pp. 1–25, Springer, Heidelberg (2006).
8. Kyei, K., Gyekye, K.: Determinants of Unemployment in Limpopo Province in South Africa: Exploratory Studies. Journal of Emerging Trends in Economics and Management Sciences **2**(1), 54 (2017)
9. Zou, H., Hastie, T.: Regularization and Variable Selection via the Elastic Net. In: Journal of the Royal Statistical Society. Statistical Methodology 2005, vol. 67, Stanford (2005). 10.1111/j.1467-9868.2005.00503.x

10. Kreiner, A., Duca, J.: Can machine learning on economic data better forecast the un-employment rate?. In: Applied Economics Letters, Guyon, I. Nikravesh, M. Gunn, and S. Zadeh, L. (eds.) Studies in Fuzziness and Soft Computing 2006, vol. 27, pp. 1434–1437, Routledge, Dallas (2020). 10.1080/13504851.2019.1688237
11. Mulaudzi, R., Ajoodha, R.: Application of Deep Learning to Forecast the South African Unemployment Rate: A Multivariate Approach. The 7th Asia-Pacific Conference on Computer Science and Data Engineering, (2020)
12. Montgomery, A., Zarnowitz, V., Tsay, R., Tiao, G.: Forecasting the U.S. unemployment rate. Journal of the American Statistical Association **93**(442), (1998)
13. Hyndman, R., Koehler, A., https://robjhyndman.com/papers/mase.pdf. Last access 10 December 2020
14. Bishop, C.: Pattern Recognition and Machine Learning. Springer, Cambridge (2006)