International Conference on “Interdisciplinary Research in Technology & Management” (IRTM 2021)

Conference Proceedings

Edited by

Satyajit Chakrabarti
Director, IEM Kolkata, India

Rintu Nath
Scientist – F, Vigyan Prasar, Department of Science and Technology, Govt. of India

Pradipta Kumar Banerji
Dean (Management Studies), Institute of Engineering and Management, Kolkata

Sujit Datta
Institute of Engineering & Management, Kolkata, India

Sanghamitra Poddar
Institute of Engineering & Management, Kolkata, India

Malay Gangopadhyaya
Institute of Engineering & Management, Kolkata, India

First edition published 2022
by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

CRC Press is an imprint of Informa UK Limited

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalogue entry has been requested.

ISBN: 978-1-003-20224-0 (ebk)

DOI: 10.1201/9781003202240

Chapter 66

Improving the Performance of Multivariate Forecasting Models through Feature Engineering: A South African Unemployment Rate Forecasting Case Study

Rudzani Mulaudzi
School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg, South Africa

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg, South Africa

Abstract—The ability of machine learning models to forecast unemployment rates better than traditional statistical methods has been well established in literature. The ambition of researchers, in this field, over the last decade has been to demonstrate that machine learning models are able to forecast unemployment rates as well as or better than traditional statistical methods. Feature engineering has thus far been applied to a limited extent when forecasting unemployment rates. Especially when compared to feature selection and feature extraction. This research leverages feature engineering to demonstrate that such techniques could improve the performance of the models. The application of feature engineering on multivariate data to forecast the South African unemployment rate increased the R-squared by over 100% on average and decreased the mean absolute scaled error by ~1%. Demonstrating that such techniques are of value when forecasting multivariate data.

Index Terms—Forecasting, Machine Learning, Feature Engineering, Unemployment

I. 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 [1]. According to [2], 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.

Unemployment is typically forecasted using statistical methods such as autoregressive integrated moving average (ARIMA) and vector autoregression (VAR) [3], [4]. However, these models are better suited for linear data [4]. Unemployment rates are not typically linear. Hence, machine learning models have been introduced to the forecasting world [5], [6].

The introduction of machine learning models to time series forecasting has resulted in accuracy improvement i.e. lower errors [2], [7]. Thus far, there has been limited use of feature engineering techniques in these machine learning models. This research shows that machine learning models benefit from feature engineering as their accuracy and ability to explain the variance in the data improves.

Concretely, this research i) demonstrates that the use of feature engineering reduces the mean absolute scaled error (MASE) of machine learning models when forecasting the South African unemployment rate, as well as ii) improving the ability of the machine learning models to explain the variation in the data (R-squared).

The research builds on the demonstration by [7] and [8], that machine learning models can forecast the South African unemployment rate more accurately than traditional models. The research used regression models, kernel-based models, decision trees, and neural networks. However, feature engineering techniques were not applied in those research papers. Hence, this research builds on previous work, by applying feature engineering techniques.

The rest of this paper is structured as follows. Section II discusses the foundation work by [7] and [8], i.e., multivariate models to forecast the South African unemployment rate. Section III discusses the results that were achieved by using feature engineering when forecasting 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

[7] demonstrated that the use of machine learning, compared to the vector autoregression (VAR), improves the forecasting accuracy by over 100%. The VAR achieved a mean absolute scaled error (MASE) of 26.3, whilst the lowest MASE from the machine learning models was 0.91. Therefore, the paper demonstrated that the application of machine learning models improved the MASE, making it 28 times better than VAR.

[8] investigated the impact of feature selection techniques on forecasting the South African unemployment rate. They found that the feature selection techniques improve the forecasting accuracy of machine learning models by at least 15%. Furthermore improving the compute requirements of the models.

Feature selection reduces the number of features by selecting a subset of them for use in the model [8]. If the initial features are not clear or known, the features are extracted from the data: this is known as feature extraction [11]. Feature engineering is different from both of these because it focuses on creating new features from the initial feature set i.e. by combining features, transforming features, performing statistical calculations on the features, etc [13].

To the best of our knowledge feature engineering has not been as widely used for unemployment rate forecasting as feature selection or feature extraction. The benefits of feature selection and feature extraction have been well demonstrated through the usage of innovative data. Examples being the use of GPS data [9], Google search data [10], and smart metering data [11] to forecast the unemployment rate. This research attempts to contribute towards using feature engineering to forecast the South African unemployment rate.

This section provided an overview of the foundational research on which this research is built on. The section discusses how unemployment rates have been successfully forecasted using machine learning models. The next section discusses results from the application of feature engineering to forecast the South African unemployment rate.

III. RESULTS AND DISCUSSIONS

The research builds on previous work to forecast the South African unemployment rate but focuses on feature engineering and the benefits that can be derived from it. The following subsection provides the research methodology and discusses the results of the analysis.

A. Performance Measures

[7] 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 the asymmetry issues associated with the mean absolute percentage error (MAPE), which is the typical measure using [4]. 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% [4], [7]. Therefore, [4], 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 (1):

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

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

R-squared was also used as a performance measure. It evaluates the ability of the model to explain the variation in the data used to forecast the South African unemployment rate [4]. In other words, it answers the question 'how well does the model fit the data?'. (2) show the equation of R-squared:

$$R^2 = \frac{\Sigma\left(\hat{y}_t - \bar{y}_t\right)^2}{\Sigma\left(y_t - \bar{y}_t\right)^2} \quad (2)$$

where, y is the actual value being estimated, $\hat{y}_t$ is the forecast value, $\bar{y}_t$ is the mean of the actual values. The denominator is also referred to as the total sum of squared (SST). Hence, R^2 can also be written as

$$R^2 = 1 - SSE/SST \quad (3)$$

where, SSE is the user is the sum of square errors i.e. $\Sigma(y_t - \hat{y}_t)^2$. Therefore, although R^2 should always be between 0 and 1. It can be lower than 1 where the (3) is implemented.

B. 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. [7] demonstrated that this strategy was the most effective for the South African unemployment rate forecasting.

The number of features was increased by increased by adding two lag values. The selection of lags was based on work by [12], who demonstrated that two lags resulted in the highest performance improvement when forecasting the South African unemployment rate.

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

C. Experiments

A total of 12 orthodox machine learning models were deployed: elastic net (ENET), 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 boosting (XGB), and multilayer perceptron (MLP).

The original dataset had 147 features. This feature set was reduced using feature selection techniques.

According to [13], 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. 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 also used. However, it is a computationally inefficient method as it uses greedy search. 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.

Therefore, a total of 216 experiments without lags included, and another 216 with lags included were conducted for this research, i.e. 432 experiments in total.

D. Analysis and Results

Using data from the South African Reserve Bank (SARB). The experiments conducted in [7] were reproduced with a focus on the R-squared. Feature selection techniques were also used, similar to [8]. A key difference with the previous researchers is that the data was forecast over four horizons: horizon 1 (4 months), horizon 2 (8 months), horizon 3 (12 months), and horizon 4 (24 months).

A total of 432 experiments were run. The model that achieved the highest R-squared, without the inclusion of lags, was the SVR model. The model had an R-squared of 0.6705 and Bayes ridge with an R-squared of 0.5821. The results are the average over the four horizons.

Table I shows the R-squared for the top-performing models. Only three models achieved an R-squared above 0.5. The Rsquared being below 0.5 was a cause for concern because it signalled that the models could only explain less than 50% of the variation in the data.

R-squared measures how much variation in the data the model is able to explain relative to a na¨ıve model [4]. Hence, feature engineering was considered as a mechanism to improve this.

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 [12].

Table I 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

** NO FS = No feature selection was applied, Unique = removal of duplicated feature, MIG = Mutual Information Gain, ANOVA = Analysis of Variance, PCA = Principal Component Analysis, EM ENET = elastic net embedded method, and Variance = removal of low variance threshold features.*

[12] found that two was the optimal number of lags through trial and error. In [12] lags between 2 and 7 were tried. Adding that the 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 1 shows that the unemployment rate has a high correlation with its lag values [12]. Demonstrating that there is evidence of temporal dependencies in the South African unemployment rate.

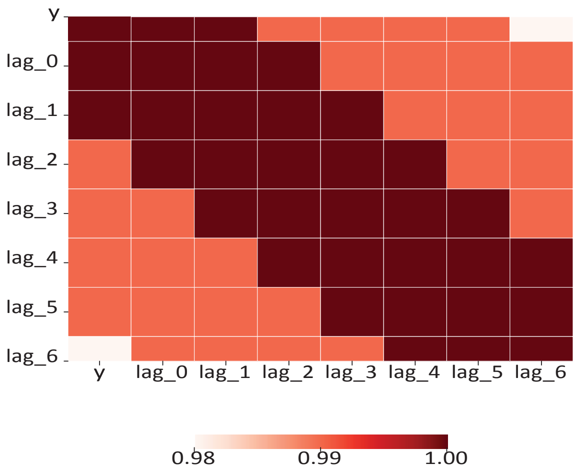

Fig. 1. The Correlation Map for the lag values of unemployment rate showing that the unemployment rate is highly correlated with the lag values.

The addition of the lag variables improved the R-squared by over 100% on average. 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 the addition of lags, which improved to 132 after the lags were added. Furthermore, 73 experiments achieved an R-squared above 0.5, an improvement from 3. Table II shows the top ten R-squared changes.

The inclusion of lag values improved the R-squared of the models significantly. Some models, which had negative R-squared, improved by more than 100%. An example of this is the MLP with the recursive feature elimination feature selection method with elastic net (RFE ENET) (readers should read [8] for a brief overview of the feature selection methods). The model achieved the top MASE score when lags were not included, but the R-squared was –0.2127. After lags were added, the R-squared improved to 0,7273: a 439% improvement.

Table II The inclusion of lag values of the south african unemployment rate improves the R-squared by over 100% on average.

Model	Feature Selection	Without Lags	With Lags
		R-squared	R-squared
Ridge	MIG	0.4118	0.7875
ENET	MIG	0.0171	0.7875
ENET	PCA	0.4411	0.7872
LASSO	XGBoost	0.3097	0.7872
LASSO	PCA	0.4310	0.7849

LASSO	MIG	0.0511	0.7849
ENET	XGBoost	0.2911	0.7842
LASSO	Random Forest	0.2244	0.7842
RFR	RFE LASSO	0.0092	0.7750
ENET	Unique	0.1513	0.7731
Bayes Ridge	NO FS	0.3438	0.7649

** NO FS = No feature selection was applied, Unique = removal of duplicated feature, MIG = Mutual Information Gain, XGBoost = Extreme Gradient Boost, PCA = Principal Component Analysis, Unique = removed duplicated features, and RFE LASSO = recursive feature elimination with LASSO estimator.*

The accuracy (reduction in MASE) also improved slightly. 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 improved 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 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 their memory capabilities.

The results are shown in Table III, which shows the order change in the top models.

The overall top model after the inclusion of the lags was the SVR. Showing that the inclusion of lags also improves the linear separability of the dataset selected through analysis of variance (ANOVA) feature selection technique: 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 when on the MASE of 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 is well represented in the top models, with and without lags.

Table III 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 group dynamics should be considered or retained when selecting features regardless of the inclusion or exclusion of lags.

There is an increase in the number of univariate feature selection techniques (in the top models) when lags are included. This shows that when the lag values are added, the variation within a feature had a big impact on the ability of the model to forecast the unemployment rate. Removal of low variance features improving the accuracy. This can be noted by observing that four out of the six univariate features in the top models are concerned with variance: variance thresholds removals, the ratio between variance within features and variances between features, or removal of low variance features.

IV. CONCLUSIONS

This paper demonstrates that using feature engineering reduces the MASE of machine learning models by ~1%. Furthermore, improving the R-squared by over 100%.

The work builds on previous similar work that has successfully demonstrated that the South African unemployment rate can be forecasted using machine learning models. This research builds on that work by using feature engineering techniques. This kind of techniques have been used to a limited extent in unemployment forecasting models.

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.

This research only used one feature engineering technique: inclusion of lags. Future research should look to including more features and seeing the outcome. There are over one thousand possible features that can be engineered for time series data.

ACKNOWLEDGMENT

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

Table III The comparison of how the inclusion of lag values change the performing (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

** NO FS = No feature selection was applied, Unique = removal of duplicated feature, MIG = Mutual Information Gain, EM ENET = Embedded method using ENET, Variance = removal of low variance threshold features, and Variance Unique = low variance features are removed along with duplicated features.*

REFERENCES

[1] R. Maluleke, "Quarterly Labour Force Survey - Quarter 1 2020," Statistic South Africa, 2020.

[2] M. Aiken. "A neural network to predict civilian unemployment rates," Journal of International Information Management, Volume 5: Issue 1, Article 3, 1996.

[3] 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.

[4] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice. OTexts," Melbourne, Australia, 2nd Edition, 2018.

[5] 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.

[6] C. Katris. "Prediction of unemployment rates with time-series and machine learning techniques," Computational Economics, 2019.

[7] R. Mulaudzi and R. Ajoodha, "Application of Deep Learning to Forecast the SouthAfrican Unemployment Rate: A Multivariate Approach," Seventh IEEE ComputerScience and Data Engineering Conference, Melbourne, 2020.

[8] R. Mulaudzi and R. Ajoodha, "Demonstration that the Use of Feature Selection onHigh Dimensional South African Macroeconomic Data Results in Improved Performance," Sixth International Congress on Information and Communication Technology Conference, ISBN: 21945357, London, 2021.

[9] D. Moriwaki, "Nowcasting Unemployment Rates with Smartphone GPS Data," International Workshop on Multiple-Aspect Analysis of Semantic Trajectories, Wurzburg, 2019.¨

[10] F. D'Amuri and J. Marcucci, "'Google It!' Forecasting the US Unemployment Rate with A Google Job Search Index," SSRN Electronic Journal, Elsevier, Italy, 2012.

[11] C. Montaklez and W. Hurst, "A machine learning approach for detecting unemployment using the smart metering infrastructure," IEEE Access, 8, 22525–22536, 2020.

[12] R. Mulaudzi and R. Ajoodha, "An exploration of machine learning models to forecast the unemployment rate of South Africa: a univariate approach," The InternationalMultidisciplinary Information Technology and Engineering Conference, ISBN: 978-1-7281-9519-9, East London (South Africa), 2021.

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

[14] A. Bouchard-Cotê and F. Sahin, "Feature Engineering and Selection," CS 294: Practical Machine Learning, 2009.