Genetic algorithm based features selection for predicting the unemployment rate of India

Sana Mohammed¹, Fakruddin Mohammed²
¹ St. Francis College, Letchworth, Hertfordshire, United Kingdom
² ZAS Consultancy Services Ltd., Hitchin, Hertfordshire, United Kingdom

SUMMARY
The forecasting of employment is critical for policymakers to prevent large impacts on modern life. Many studies have been conducted to predict the unemployment rate using time series models and structural models using data from advanced economy countries, but only a few studies have used data from developing countries. In addition, the two economic theories which focus on short-term inverse relationships between “unemployment and inflation” or “unemployment and gross domestic product” as described by the Phillips curve and Okun’s law, fail to predict the unemployment in low-middle-income countries and advanced economies undergoing stress. In this study, we have chosen India’s labor market, to test the hypothesis described by the Philips curve and Okun’s law because India’s economy exhibits the characteristics of low/middle-income countries with the majority of the labor force employed in agricultural and low-paying seasonal jobs. Given that the unemployment rate depends on a wide number of social and political factors we employed genetic algorithms (GA) for feature selection. Based on our limited literature survey, there were no research studies employing genetic algorithms to explain India’s labor market data so we have chosen India’s macroeconomic data to test the effectiveness of the GA model for features selection. Our results show that, for the chosen period, the Philips curve relationship doesn’t hold well for the Indian labor market. In addition, feature selection using the GA suggested that economic growth, household consumption, government spending and coal imports are the best macroeconomic variables to explain variation in the Indian labor market.

INTRODUCTION
The unemployment rate is the percentage of people who do not currently have a job as a fraction of working-age people who are able and willing to work. This rate is dependent on factors such as macro (economic growth, interest rates, inflation, government spending and currency exchange), social (quality of education, health and cultural bias towards gender and ethnicity), political (government stability, net migration and foreign policies), and natural disasters (floods, earthquakes, loss of housing and skilled labor etc.) (1-8). An increase in unemployment leads to slow economic growth as the unemployed people have less disposable income, which leads to a fall in aggregate demand which may in turn lead to recession (9). If high unemployment persists for a long duration, then it subsequently leads to a fall in the standard of living (10). Therefore, the forecast of the unemployment rate is very important for policymakers and political leaders, as one of the macroeconomic objectives is to keep unemployment low (11). For example, the Federal Reserve, the most powerful economic institution in the United States, which sets interest rates as part of its monetary policy, decided in 2012 that it would keep the interest rate at 0% as long as the unemployment rate is above 7% and inflation does not rise above 3% (12). Similarly, forecasting unemployment data is essential to insurance companies to plan their cash flow (unemployment payouts) and to the traders who trade exchange rates for investments/financial gains (13-14).

Two economic theories try to explain the unemployment rates using macroeconomic variables. Firstly, the Philips curve attempts to explain the short-term inverse relationship between the unemployment rate and the inflation rate. Secondly, Okun’s law describes the inverse relationship between the unemployment rate and gross domestic product (GDP). Many studies have verified the relationship between the unemployment rate against inflation and GDP; however, most of these studies use data from advanced economy countries, and these two economic theories fail when applied to developing nations (low/middle-income countries) and in advanced economies under severe stress (15-18). Therefore, in this study, we take India’s labor market to verify the hypothesis suggested by the Philips curve and Okun’s law models. The reason for choosing India’s labor market is because the majority of its labor force works in low-paying agricultural, seasonal, and cyclical jobs and hence per capita income is a mere 2,283 USD as compared to 47,203 USD in the UK (19-20).

To test the short-term empirical relationship as described by the Philips curve and Okun’s law we used inflation and GDP monthly data of India’s economy from March 2016 to June 2021 (a total of 67 observations). The data was downloaded from The Global Economy website, which provides well over 500 macroeconomic data indicators for 200 countries over different time horizons of monthly, quarterly, and yearly (20).

Our results show that, for the chosen period of study (2016-2021), the relationship described in Philip’s curve does not hold well for the Indian labor market but Okun’s law. To further enhance the predictability power of the model we employed genetic algorithms for feature selection from a pool of fifteen macroeconomic variables: Economic growth, GDP per capita, Capital investment, Household consumption, Trade openness, Foreign Direct Investment, Current account balance (as GDP fraction), Trade balance, Balance of payments, Government spending (as GDP fraction), Tax revenue, Coal imports,
Electricity consumption kilowatt-hours, Financial system deposits, and Military spending. The reason for only choosing macro data is due to a lack of social (gender bias, health, education etc.) and political data (immigration) at monthly granularity. Our results show that the four features (economic growth, household consumption, government spending, and coal imports) selected by the genetic algorithm outperform the autoregressive (AR) time series models and empirical models suggested by Phillips curve and Okun’s law. We show GA is an alternate model for feature selection as compared to the traditional forward and backward selection methods, Random Forest, or Decision Tree models.

RESULTS

In this study, we conducted four experiments to predict India’s unemployment rate using the monthly unemployment data from March 2016 to June 2021 (a total of sixty-seven monthly observations). Due to the small sample size, the data was split into a 90:10 ratio for training and testing the model. After the split, the training data had 60 observations which were used to build the model and the testing data had 7 observations which were used to test the model predictions.

The first experiment consisted of building a simple benchmark model. The results of this model were used to compare with the rest of the three experiments. For each experiment, a regression model was built, and p-value and Adjusted R²(Adj. R²) of the training data and root mean square error (RMSE) and directional or trend accuracy (no. of times the estimated and actual values moved in the same direction) of the testing data were captured to test the effectiveness of the models. The second and third experiment consists of verifying the inverse relationship described by Philip’s curve and Okun’s law. Finally, in the fourth experiment, a genetic algorithm was built to find the superior variables which explain variation in the unemployment data from a pool of available macroeconomic variables, and then a multivariate regression model was built.

Auto-regressive time series model as a benchmark

Auto-regressive (AR) models are the simplest models which use past observations to predict future values. The autocorrelation function (ACF) and partial autocorrelation function (PACF) are metrics that measure the number of past observations that were most useful in explaining future data. The ACF and PACF plots of India’s monthly unemployment data show that the future month of unemployment can be predicted using the immediate previous two months of unemployment data (Figure 1). Therefore, the AR model for predicting the unemployment rate at time t is given by Equation 1, where \( U_t \) is the unemployment at time t, and \( U_{t-1} \) and \( U_{t-2} \) are the unemployment rate at time periods t-1 and t-2 in the past. \( \phi_1 \) and \( \phi_2 \) are the regression coefficients, and \( \epsilon_t \) is the error term.

\[
U_t = \phi_1 \cdot U_{t-1} + \phi_2 \cdot U_{t-2} + \epsilon_t \quad [\text{Eqn 1}]
\]

The benchmark model results show that the past two observations of unemployment rates play a significant role in predicting the next period’s unemployment role (p<0.05). However, the model can explain only 50% of the variation in the unemployment rate of the training data and struggle to predict the levels of unemployment and directional trend of the testing data (Adj. R²=50%, accuracy=17, RMSE=2.16, Figure 2).

Does the Indian Labor Market Obey Phillips Curve Economic Theory?

In the second experiment, a regression model of the Phillips curve, which describes the inverse relationship between inflation and unemployment is built. In this model Equation 2, \( U_t \) is the unemployment rate at time period t, \( \pi_t \) is the inflation rate, \( \phi_i \) is the regression coefficient, and \( \epsilon_t \) is the error term.

\[
U_t = \phi_1 \cdot \pi_t + \epsilon_t \quad [\text{Eqn 2}]
\]

Surprisingly, for the chosen period (2016-2021), India’s labor market data exhibits a small positive correlation of 0.3 instead of a negative correlation between the unemployment rate and inflation. Despite the correlation, our results show that inflation is a significant variable but can only explain 8% of the variation in the data (p<0.05, Adj. R²=8%, Figure 3A). However, inflation was able to predict changes in the direction of the unemployment trend half the time correctly but with large prediction errors (RMSE=2.13, directional trend prediction accuracy=50%, Figure 3B).

Does the Indian Labor Market Obey Okun’s Law?

The third experiment involves building another simple linear regression model described by equation (3) to verify whether the Indian labor market obeys Okun’s law. In Equation 3, \( U_t \) is the unemployment rate at time t, \( GDP_t \) is the gross domestic product in USD at time period t, \( \epsilon_t \) is the error term, and \( \phi_i \) is the regression coefficient.

\[
U_t = \phi_1 \cdot GDP_t + \epsilon_t \quad [\text{Eqn 3}]
\]

For the chosen period (2016-2021), despite a strong negative correlation of 0.73 between unemployment and GDP, the model was not able to explain the variation in the data (Adj. R² = 54.2%, p=0.000, Figure 4A). However, the GDP data has more power for predicting the unemployment rate directional trend but with large prediction errors (RMSE=2.55, directional trend prediction accuracy=67%, Figure 4B).

Figure 1. India’s employment rate autocorrelation and partial autocorrelation plots. The shaded area represents the coefficients of the terms in the regression equations that are close to zero with a 95% confidence level. A) An autocorrelation plot between the data points of India’s unemployment time series data. B) Is the partial autocorrelation plot that shows the observation at a given time period is dependent on its immediate two previous observations as they lie outside of the blue-shaded areas.
Are there better predictors for forecasting the Indian labor market?

Experiments 1-3 show that past employment observations, inflation, and GDP alone are not enough to explain India’s unemployment data. Therefore, to investigate if there are any superior macroeconomic indicators, we used a GA for feature selection. We inputted a pool of 15 macroeconomic variables to the GA and then used the R statistical programming language along with the GA package to implement the feature selection (Table 1). Equation 4 describes the multivariate regression equation used for GA implementation.

\[ U_t = \sum_{j=1}^{U} \left( \text{Ind}_j \times \phi_j \times V_j \right) + \epsilon_t \quad \text{[Eqn 4]} \]

In equation (4), the upper limit represents the number of macroeconomic variables considered by our algorithm, \( U_t \) is the unemployment rate at time \( t \), \( \text{Ind}_j \) or 1 represents whether a variable exists or not, \( V_j \) represents each macroeconomic variable (Table 1), \( \phi \) represents regression coefficients and \( \epsilon_t \) represent error term in the regression equation.

The algorithm converged after 30 iterations and the results show the superior macroeconomic economic indicators that can explain the variation in unemployment are: economic growth (GDP change), household consumption as a fraction of GDP, government spending as a fraction of GDP, and coal imports in tons (p<0.05). Using these four variables, we built a multivariate regression model as described by Equation 5 to test the model performance.

\[ U_t = \phi_1 \cdot v_1 + \phi_2 \cdot v_2 + \phi_3 \cdot v_3 + \phi_4 \cdot v_4 + \epsilon_t \quad \text{[Eqn 5]} \]

In Equation 5, \( U_t \) is the unemployment rate at time \( t \), \( v_j \) is economic growth (GDP change), \( v_2 \) is household consumption (fraction of GDP), \( v_3 \) is government spending (fraction of GDP), and \( v_4 \) is coal imports (tons), \( \epsilon_t \) is the error residual and \( \phi \) are regression coefficients. The model built on variables identified by the genetic algorithm outperforms all the previous models in terms of the ability to explain the data and ability to predict the change in trend (Adj. \( R^2 \)=78.1%, direction prediction accuracy=83%, RMSE=0.87, Figure 5).

To verify the consistency of the GA model results we used the same dataset as input to the Random Forest (RF) regressor model for feature selection. In order of importance, the top four variables identified by the RF model are Government Spending, Coal Imports, Financial System Deposits and Electricity Consumption. The key points to note here are that: (a) two of these variables (government spending and Coal Imports) are common between the GA and RF model and (b) the RF model shows a significant relationship between the unemployment rate and Financial System Deposits. Financial system deposits is a very interesting variable because they might be capturing the public expectation that the economy is going to shrink and thus increase the savings for consumption in the future. The regression model built using the top four variables identified by RF can only explain 71% of the variation in the data as compared to the GA model which was able to explain 83% of data (Adj. \( R^2 \)=0.71). Also, to test that the variables that were not selected by GA don’t contribute to explaining the data, a regression model was built and the results show the model was able to explain only 31% of the data with a trend prediction accuracy of 17%.
DISCUSSION

To verify whether the unemployment theories suggested by the Phillips curve and Okun’s law satisfy the unemployment data observed in India’s labor market, we built regression models and the results are compared with the simple AR model for baseline results. In addition, in search of superior macroeconomic variables, we employed the GA model for feature selection. The results of those models are discussed in this section below.

Firstly, it is very surprising to see that, the simple AR model which uses two immediate previous months’ unemployment data to predict the next month’s unemployment rate was able to explain the 50% variations in the data. However, this model was not good at predicting the directional trend (actual and predicted values moving in the same direction) as the model accuracy was only 17% (**Figure 3B**). Our results are consistent with previously published attempts to predict the French unemployment rate where they found a time lag of four time periods was required to improve the accuracy (5). The reason for this lagging behavior could be that unemployment is a lagging economic indicator because any actions policymakers take to control unemployment may take time to trickle down the effects of those policies on the economy.

The interesting point to note with the second experiment is that, as per the Phillips curve, unemployment and inflation are inversely related. However, for the chosen period, India’s labor market data exhibits a positive correlation of 0.3 between unemployment and inflation. Despite the correlation, the model was able to explain only 8% of the variation in the unemployment rate. These results are consistent with existing research studies, which highlight that for developing economies (low and lower-middle-income countries) the short-term inverse relationship as described in Philip’s curve doesn’t hold because in developing countries most employment contracts are short-term and seasonal (15-18). This observation is consistent within the Indian labor market because the agriculture sector in India contributes 19% to its GDP and employs 50% of the labor force with employment contracts that are short-term and seasonal (19). Additionally, due to the small sample size of the data, we would like to test this hypothesis using data from other developing nations in Southeast Asia like Bangladesh, Pakistan, Sri Lanka, and Malaysia.

The results of the third experiment, which captures the inverse relationship between GDP and the unemployment rate as described by Okun, show that GDP growth was able to explain 54.2% of the variation in the unemployment rate of the Indian labor market, and it was also able to predict the trend change with 67% accuracy. Our results are consistent with economic theories because when the economy is growing, the aggregate demand for goods and services increases; therefore, firms are willing to increase production to meet the increased demand from consumers. To increase the supply in response to increased consumer demand, the firms are likely to hire labor, leading to a lower unemployment rate. However, the large errors in the training dataset indicate that the model is biased, and GDP alone was not enough to fully explain the variability of India’s unemployment rate. Some other economic variables are likely needed to fully explain the variability. A report from the Federal Reserve Bank highlights that Okun’s law is a good rule of thumb; however, it is not consistent enough to use it as forecasting to explain the unemployment data (18).

To further improve the predictive power of the model suggested by Okun’s law, we employed a GA to find the best
economic indicators that can fully explain the variability of the data. The results of the fourth experiment show that the prominent variables for explaining the Indian labor market data are economic growth, government spending, household consumption, and coal imports. The correlations of these four variables in the aforementioned order are 0.05, 0.51, -0.32, and -0.40. One interesting observation of the correlations is that GA identified Growth as a prominent variable even though its correlation coefficient is 0.05 as compared to other variables whose correlation is much higher. The output of the genetic algorithm logically makes sense as an increase in any or all of these four input variables leads to an increase in aggregate demand. The increase in consumer aggregate demand entices firms to increase the supply by investing in capital goods or by hiring labor, thereby lowering unemployment. The multivariate regression model built using these four economic indicators was able to explain 83% of the variation in India’s labor market and also predicted the directional change in the trend with 83% accuracy.

Even though the GA model outperformed the other models, we strongly feel that there is more work that needs to be done to test the model’s robustness. The sample size of the dataset used in this study is very small (67 observations) with 15 variables. This ratio of sample size to number of observations doesn’t meet the best practices. Therefore, we would like to test the GA model’s robustness by taking the economic data from developing countries in the Southeastern Asian region of varied sample sizes (long and short term). Additionally, GA suffers from the sensitivity of the objective function used to find the optimal solution. Therefore, we would like to study the robustness of the model if we use different objective functions like RMSE, number of correct predictions, etc. In addition, we would like to test the model stability by taking the data from existing studies and testing whether identical conclusions can be reached by using GA.

In summary, the significance of this study is that the Philips curve does not hold for present-day India, but to extrapolate to other developing nations is outside of the scope of this study. As suggested by Okun’s law we found GDP to be a good economic indicator in explaining the unemployment rate in India; however, to fully explain the variability of the unemployment rate, we needed to add some other economic variables. The study shows that GA is a potentially good candidate for feature selection, especially if the number of variables is large since it is fast due to parallel execution compared to backward or forward selection techniques. However, the GA model needs some more thorough testing for its robustness as the sample size used in this study is small.

MATERIALS AND METHODS

Data pre-processing and model assumptions

In this study we used India’s monthly labor and macroeconomic data from March 2016 to June 2021 (a total of 67 observations). The summary statistics like mean, median, and standard deviation of the monthly macroeconomic data used in this study are tabulated in Table 1. To ensure the data used in this study satisfies all the assumptions we have normalized the data around the mean and correlation matrix and QQ plots were used to verify the collinear and normal distribution of error terms.

Experimental procedure

Due to the small sample size (67 observations), the data was split into training and testing datasets in a 90:10 ratio instead of the usual practice of an 80:20 ratio. The training data was used to build the model and the testing was used to test the model performance.

To build the regression models proposed in previous sections we used the statsmodel Python package. The regression models for experiments 1-3 are described by equations 1-3, respectively.

For each experiment we conducted, the regression model output of p-value and Adj. R² of the training data were captured. Similarly, for the test data, the RMSE (square root of the mean of squares of errors between actual and predicted values) and the directional trend accuracy (no of times the actual and estimated values moved in the same direction) were captured for comparison purposes.

Finally, a genetic algorithm was used to select the superior predictors from a pool of macroeconomic variables listed in Table 1 that explain the variations in India’s unemployment data. The algorithm implementation details are described below. Due to lack of monthly data, the study was restricted only to the variables listed in Table 1.

Initial Population for GA

The population consists of \( N \) number of samples, where each sample has a subset of macroeconomic variables \( \{v_1, v_2, v_3, \ldots, v_m\} \). The presence of a macroeconomic variable \( v_i \) in a sample was represented by 0s or 1s where 1 represents a macroeconomic variable was present otherwise not present. For example, a sample represented by a set...
\{0,1,1,0,1,0,0,0,1,1,0,0,0,0,1\} implies that the macroeconomic variables \{0,v_4,v_6,v_7,0,0,0,0,0,0,0,0,0\} were present in the model and all other macroeconomic variables were ignored. At the start of the genetic algorithm implementation, \(N\) number of samples were considered and the macroeconomic variables present in each sample were selected randomly. As a general guideline, the initial size of the population is 1 to 2.5 times the number of macroeconomic variables and then gradually increasing the population to several hundred. Therefore, in this research study we have chosen an initial population of \(N = 50, 100, 150 \text{ and } 200\) to carry out model simulations and the optimum results are found when the population size was \(N=200\).

**Fitness Function for GA**

The genetic algorithm needs a metric or a fitness function to find the top \(n\) samples from an initial population of \(N\). In this specific example, a regression model was built using the predictors that were present in the sample. For example, if a sample is represented by the sequence \{0,1,1,0,1,0,0,0,1,1,0,0,0,1\} then the corresponding regression equation is described below.

\[
U_t = l_1 \cdot \phi_1 \cdot v_{i1} + l_2 \cdot \phi_2 \cdot v_{i2} + \ldots + l_{11} \cdot \phi_{11} \cdot v_{i11} + \epsilon_t \quad \text{[Eqn 6]}
\]

In Equation 6, \(t\) is the time period, \(\phi\)'s are regression coefficients, \(v\)'s are macroeconomic variables, \(l_i\) is 1 or 0 depending on whether a macroeconomic variable exists or not in the multivariate regression equation and \(\epsilon\) is an error term \(U_t - \hat{U}_t\). The above regression equation is fitted to the training data and Adj. \(R^2\) was used as a fitness metric. We used the programming language library called GA for the implementation of the algorithm.

Finally, two regression models were built to test the statistical significance of the GA model. The first model comprises regression on four superior variables identified by the GA model and the second model comprises the remaining eleven variables that the GA model ignored. The regression result shows all four superior variables are statistically significant (p<0.05) and the remaining eleven variables are insignificant (p>0.05).

Received: August 12, 2022  
Accepted: June 19, 2023  
Published: March 16, 2024

**REFERENCES**


Copyright: © 2024 Mohammed and Mohammed. All JEI articles are distributed under the attribution non-commercial, no derivative license (http://creativecommons.org/licenses/by-nc-nd/4.0/). This means that anyone is free to share, copy and distribute an unaltered article for non-commercial purposes provided the original author and source is credited.