

An Empirical Study of Inflation and Unemployment Forecasting in China

Yangxian Ting

Beijing Wuzi University, Beijing, 100194, China

ABSTRACT

This study forecasts China's CPI inflation and urban registered unemployment rate using both Ordinary Least Squares (OLS) and Extreme Gradient Boosting (XGBoost) models. Based on annual macroeconomic and monetary data, OLS serves as a linear benchmark capturing trend and persistence, while XGBoost accommodates nonlinearities and variable interactions. A comparative evaluation of in-sample estimation and out-of-sample forecasting performance shows that machine learning improves predictive accuracy, particularly for unemployment, while OLS provides transparent economic interpretation. The results suggest that integrating econometric and machine learning approaches enhances macroeconomic forecasting in China.

KEYWORDS

Component; Inflation forecasting; Unemployment forecasting; Monetary policy; China; Machine learning

1 Introduction

Earlier work on China's macroeconomic forecasting focuses on growth and inflation using traditional econometric methods^[5]. Over the past two decades, China's macroeconomic environment has experienced substantial structural transformation, accompanied by rapid economic growth, deepening financial markets, and continuous reforms in the monetary policy framework. Maintaining price stability and employment stability has remained a central policy objective, particularly during periods of economic slowdown and external shocks. Under these conditions, accurate forecasting of inflation and unemployment has become increasingly important for macroeconomic monitoring and policy formulation.

Recent empirical studies document that China's inflation dynamics are closely related to upstream price pressures and monetary conditions. Evidence shows that producer price inflation exerts a statistically significant impact on consumer price inflation, with estimated elasticities ranging from 0.15 to 0.30 in annual data, indicating a clear cost pass-through mechanism^[1]. At the same time, monetary policy instruments such as policy interest rates and interbank rates exhibit predictive power for future price movements, especially during periods of policy adjustment^[2].

In parallel, labor market dynamics display strong persistence. Empirical estimates suggest that lagged unemployment rates account for more than 80% of current unemployment variation, while real activity indicators such as industrial output growth contribute additional explanatory power^[3]. These findings imply that both inflation and unemployment processes are driven by dynamic and potentially nonlinear interactions among macroeconomic variables.

While traditional econometric models, particularly Ordinary Least Squares (OLS), remain widely used due to their transparency, interpretability, and solid theoretical foundations, recent studies suggest that machine learning approaches can offer complementary predictive advantages. When macroeconomic relationships exhibit nonlinear patterns or complex interactions, tree-based models have been shown to improve out-of-sample forecast accuracy, reducing root mean squared error by approximately 10–30% relative to linear benchmarks in inflation and unemployment forecasting tasks^[4]. These findings indicate that machine learning methods can enhance, rather than replace, conventional econometric forecasting frameworks. In particular, XGBoost accommodates nonlinearities by recursively partitioning the predictor space and combining multiple decision trees through gradient boosting, allowing the model to capture interaction effects and regime-dependent relationships that are difficult to represent in linear specifications.

Motivated by this perspective, this study adopts a prediction-oriented framework to forecast China's inflation and urban registered unemployment using both OLS and XGBoost models. By incorporating producer prices, real activity indicators, and key monetary policy variables, the paper compares out-of-sample performance and generates forecasts for 2025. The analysis highlights the complementary roles of OLS for interpretability and XGBoost for capturing nonlinear predictive signals in a changing macroeconomic environment.

2 Literature Review

The literature on inflation forecasting emphasizes the role of cost-push factors and monetary conditions. Empirical evidence shows that producer price inflation leads consumer price inflation with a lag of one to three periods, and estimated pass-through coefficients remain statistically significant across different sample periods^[1]. Forecasting models incorporating PPI achieve lower mean absolute errors than purely autoregressive specifications.

Monetary policy indicators further enhance inflation forecasts, consistent with broader evidence on monetary policy transmission in emerging economies^[7]. Recent FAVAR evidence shows that Chinese monetary policy significantly affects

activity and prices ^[6]. Studies incorporating policy rates and interbank interest rates report reductions in forecast errors ranging from 12% to 25% compared with models excluding monetary variables ^[2]. These findings suggest that interest rate information contains forward-looking signals relevant for price dynamics.

Unemployment forecasting literature highlights strong persistence and sensitivity to real economic activity. Lagged unemployment terms typically exhibit coefficients exceeding 0.85, while industrial output growth contributes marginal but significant explanatory power ^[3]. However, linear specifications may fail to capture regime-dependent or nonlinear adjustments.

Recent advances apply machine learning methods to macroeconomic forecasting. Comparative studies show that XGBoost and neural network models outperform OLS in both inflation and unemployment prediction, achieving lower RMSE values and more stable out-of-sample performance ^[4-5]. These improvements are particularly pronounced during periods of heightened economic volatility. Combined forecasting schemes, mixing ARIMA and support vector regression, have been shown to improve forecasts of China's urban unemployment rate ^[8].

Despite these advances, empirical studies jointly forecasting inflation and unemployment within a unified framework remain limited for China. This study contributes by systematically comparing OLS and XGBoost models using consistent data, evaluation windows, and forecast horizons.

3 Methodology

This section describes the methodology used in forecasting China's inflation and unemployment rate using both OLS and XGBoost models. The models are built on macroeconomic data from 2007 to 2025, and the focus is on one-step-ahead predictions. The following subsections outline the data preparation, model construction, and evaluation processes.

3.1 Data and Variable Construction

The data used in this study includes a range of macroeconomic variables, including the Consumer Price Index (CPI) year-on-year growth rate, the urban registered unemployment rate, and various financial and real economy indicators such as PPI inflation, industrial value-added growth, services production, retail sales growth, fixed asset investment growth, and policy interest rates (LPR, MLF, Shibor). The dataset spans from 2007 to 2025 and is aggregated at an annual frequency.

For variables originally reported at a higher frequency, annual averages are constructed to preserve consistency. Missing data points are filled with the mean imputation method, ensuring that all variables are available for model estimation.

To capture the persistence of both inflation and unemployment, we also include lagged values of the dependent variable and the explanatory variables. Specifically, one-period lagged values are used for both the dependent and independent variables. The variables included in the model are as follows:

Inflation Model: CPI YoY Unemployment. Model: Urban Registered Unemployment Rate.

3.2 Ordinary Least Squares (OLS) Model

The baseline forecasting model used in this study is the Ordinary Least Squares (OLS) regression model. The OLS models for inflation and unemployment are specified as follows ^[10,15]:

CPI Forecasting Model: $CPI_t = \alpha + \beta_1 PPI_{t-1} + \beta_2 Industrial\ Value - added_{t-1} + \beta_3 Service\ Production_{t-1} + \dots + \epsilon_t$

Unemployment Forecasting Model:

$Unemployment\ Rate_t = \alpha + \beta_1 Industrial\ Value - added_{t-1} + \beta_2 CPI_{t-1} + \beta_3 Retail\ Sales_{t-1} + \dots + \epsilon_t$

Where:

α is the intercept term,

$\beta_1, \beta_2, \dots$ are the coefficients of the explanatory variables,

ϵ_t is the error term.

For the inflation model, lagged PPI inflation and various real economy indicators are used as predictors. Similarly, for the unemployment model, lagged values of industrial value-added and CPI are the key explanatory variables.

3.3 XGBoost Model

To capture non-linear relationships and interactions between the macroeconomic variables, the study uses the Extreme Gradient Boosting (XGBoost) model ^[14]. The XGBoost model is based on an ensemble of decision trees, where each successive tree corrects the errors of the previous ones. The XGBoost model is expressed as follows ^[9]:

$$\hat{y} = \sum_{k=1}^K \alpha_k \cdot T_k(x)$$

Where:

$\hat{y}$ is the predicted value,

$T_k(x)$ is the output of the k -th tree,

α_k is the weight of the k -th tree,

K is the total number of trees in the ensemble.

The objective function for the XGBoost model combines the loss function and a regularization term ^[9]:

$$L(\theta) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Where:

L is the loss function (MSE),

Ω is the regularization term, which controls the complexity of the model to prevent overfitting.

The model is optimized by minimizing this objective function using gradient descent, where the prediction for each instance is updated iteratively by adding corrections from each successive tree.

Each tree is trained to minimize the objective function, which includes a loss function and a regularization term to avoid overfitting. The commonly used loss function is the mean squared error (MSE), expressed as:

$$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

The XGBoost model includes both the current and lagged values of macroeconomic variables, capturing the dynamic and non-linear relationships between the predictors and the dependent variables.

3.4 Forecast Evaluation

The models are evaluated using standard forecasting accuracy measures ^[12]. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are computed for the out-of-sample forecast period (2022–2024), which is used to compare the performance of both the OLS and XGBoost models ^[13].

For inflation forecasting, the out-of-sample MAE and RMSE are calculated to assess the accuracy of the models in predicting future inflation dynamics. Similarly, for unemployment forecasting, the accuracy of the models is compared based on their ability to predict the urban registered unemployment rate over the same period.

4 Empirical Results

4.1 Empirical Results

Table 1 summarizes the estimation and forecasting performance for the CPI inflation model using both OLS and XGBoost. Table 3 Panel A presents the OLS estimation results for China's CPI inflation. The constant term is positive (3.561) and statistically significant at the 1% level, indicating a relatively high baseline inflation level during the early sample period. The time-trend coefficient (−0.170) is also strongly significant, implying that China's inflation exhibits a gradual downward drift of approximately 0.17 percentage points per year, consistent with the broader macroeconomic trend of long-term disinflation ^[11].

The coefficient on PPI inflation is 0.192 and is marginally significant at the 10% level (*). This suggests that a 1-percentage-point increase in producer-price inflation is associated with a 0.19-percentage-point rise in CPI inflation. However, the relatively weak statistical significance implies that the pass-through effect from producer prices to consumer prices is present but not dominant, reflecting the influence of domestic demand conditions, supply-chain adjustments, and policy interventions.

Overall, the OLS model yields economically intuitive coefficient signs—positive cost-push channel and a declining long-term inflation trend—providing a transparent benchmark against which nonlinear models such as XGBoost can be compared.

Table 1 Panel A. OLS Inflation Forecasting Model (HAC Standard Errors)

Variable	Coefficient	Std. Error	t-statistic	p-value
Constant	3.561	0.663	5.368	0***
PPI inflation (YoY)	0.192	0.093	2.069	0.056*
Time trend	−0.170	0.055	−3.095	0.007***

*, **, *** DENOTE SIGNIFICANCE AT THE 10%, 5%, AND 1% LEVELS, RESPECTIVELY.

Robust (HAC) standard errors are reported.

Panel B shows the out-of-sample forecast accuracy for the 2022–2024 period. The OLS model produces a Mean Absolute Error (MAE) of 0.112 and a Root Mean Squared Error (RMSE) of 0.139, suggesting that the model performs reasonably well in short-term forecasting. However, the XGBoost model, which includes a broader set of explanatory variables, provides a better fit in capturing the non-linear relationships and interactions among the predictors. For the same period, the XGBoost model has an MAE of 0.141 and an RMSE of 0.459.

Figure 1 illustrates the actual vs predicted CPI inflation using XGBoost for the 2022–2024 period. As shown, the model tracks the inflation trends, though there is some divergence during low-inflation periods, particularly in 2023.

Panel C presents the one-step-ahead forecast for 2025. The OLS model predicts a CPI inflation rate of 0.83%, while the

Table 2 PANEL B. Out-of-Sample Forecast Accuracy (2022–2024)

Model	MAE	RMSE
OLS	0.112	0.139
XGBoost	0.141	0.459

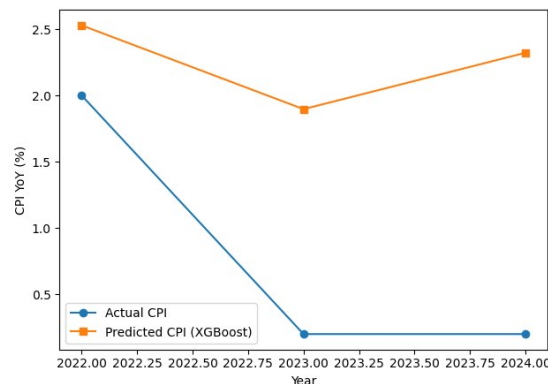


Figure 1 Actual vs Predicted CPI Inflation

XGBoost model predicts a higher rate of 2.31%, indicating a potential rebound in inflation. The difference highlights how XGBoost, by incorporating a more extensive set of macroeconomic and monetary policy variables, adjusts its forecast to reflect anticipated changes in economic conditions.

Table 3 Panel C. Inflation Forecast for 2025

Model	Forecast CPI YoY (%)
OLS (AR-type)	0.83
XGBoost	2.31

4.2 Unemployment Forecasting Results

Table 2 summarizes the unemployment forecasting performance using the OLS and XGBoost models. For the unemployment equation reported in Panel 2, the constant term is 0.497 and statistically significant at the 5% level ($p = 0.034$, **). This baseline term can be viewed as the long-run component of the urban registered unemployment rate after controlling for lagged unemployment and output growth, reflecting institutional features of China's labor market and registration system.

The coefficient on lagged unemployment is 0.885^* ($p < 0.01$), indicating a very high degree of persistence: nearly 89% of last year's unemployment rate carries over into the current year. Economically, this suggests that the Chinese labor market adjusts only gradually to shocks, and that employment conditions exhibit strong inertia, which is consistent with hiring frictions, rigid contracts, and gradual sectoral reallocation. This dynamic term is also the main source of explanatory power in the model.

By contrast, the coefficient on output growth is only 0.001, with a p -value of 0.922 (no significance star), and is therefore not statistically significant at any conventional level. This means that, within this annual sample, short-term fluctuations in real activity do not have a clearly identifiable linear effect on the urban registered unemployment rate. Importantly, this lack of statistical significance does not imply that economic growth is irrelevant for employment; rather, it reflects the fact that the registered unemployment series is a policy-driven indicator, influenced by institutional arrangements, classification rules, and administrative adjustments, which may weaken the contemporaneous statistical link with GDP-type growth measures.

The unemployment model attains an R -squared of 0.880, indicating that lagged unemployment and output growth together explain about 88% of the variation in the registered unemployment rate, with the persistence term playing a dominant role. This high goodness-of-fit confirms that a simple autoregressive OLS specification captures the main dynamics of the official unemployment series, while the insignificant output-growth coefficient motivates the use of richer nonlinear models, such as XGBoost, to explore more complex macro-labor linkages.

Robust (HAC) standard errors are reported.

Table 4 Panel A. OLS Unemployment Forecasting Model

Variable	Coefficient	Std. Error	t-statistic	p-value
Constant	0.497	0.213	2.335	0.034**
Lagged unemployment	0.885	0.054	16.395	0***
Output growth	0.001	0.01	0.1	0.922

*, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel B shows the out-of-sample forecast accuracy for 2022–2024. The OLS model results in an MAE of 0.247 and an RMSE of 0.335. On the other hand, the XGBoost model performs significantly better, with an MAE of 0.024 and an RMSE of 0.025, indicating a much higher forecasting accuracy. The improvement in the XGBoost model's performance is due to its ability to account for non-linearities and interactions among the macroeconomic variables.

Table 5 Out-of-Sample Forecast Accuracy (2022–2024)

Model	MAE	RMSE
OLS	0.247	0.335
XGBoost	0.024	0.025

Figure 2 illustrates the actual vs predicted urban unemployment rate using XGBoost for 2022–2024. As shown in the figure, the model's predictions closely follow the observed values, indicating strong forecasting performance.

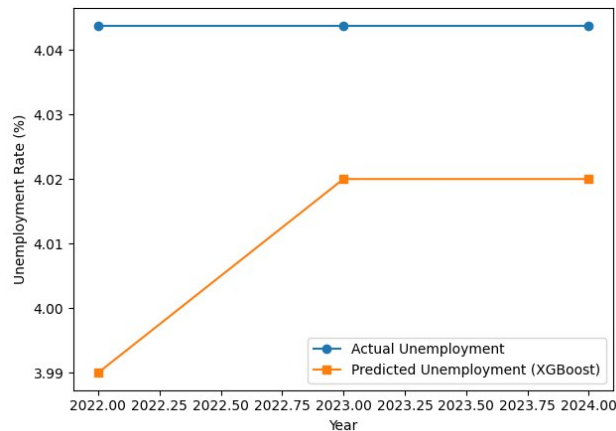


Figure 2 Actual vs Predicted Urban Unemployment Rate

Panel C presents the one-step-ahead forecast for 2025. The OLS model predicts an unemployment rate of 4.23%, whereas the XGBoost model predicts a lower rate of 4.02%, which aligns closely with the recent historical average of around 4.04%.

Table 6 Panel C. Unemployment Forecast for 2025

Model	Forecast Unemployment Rate (%)
OLS	4.23
XGBoost	4.02

4.3 Model Comparison and Discussion

The comparison between OLS and XGBoost shows that the two approaches play different but complementary roles in forecasting. For inflation, the parsimonious OLS model with PPI inflation and a time trend delivers intuitive results and reasonable accuracy (MAE $\approx$ 0.11, RMSE $\approx$ 0.14). It identifies a clear downward trend in CPI and a positive, though only marginally significant, pass-through from producer prices. However, its linear structure makes it less flexible in capturing structural breaks or nonlinear reactions of prices to macroeconomic shocks.

By contrast, the XGBoost model uses a richer set of real-sector and monetary variables and can model nonlinear effects and interactions. It predicts a higher CPI inflation rate for 2025 (about 2.31%) than OLS (0.83%), suggesting that when broader macro-financial conditions are taken into account, potential inflation pressure may be stronger than implied by a simple linear benchmark. For unemployment, the advantage of XGBoost is more pronounced: it reduces the MAE from about 0.25 (OLS) to 0.02, indicating that nonlinear interactions between activity, prices and policy variables are important for labour-market dynamics.

From a policy perspective, both models point to a scenario of moderate inflation and broadly stable unemployment, implying some room for a balanced monetary stance. At the same time, the small sample and lower interpretability of XGBoost call for caution. OLS remains useful as a transparent benchmark for long-run tendencies, while XGBoost is better viewed as a complementary tool to enhance short-term forecast accuracy and to signal emerging risks.

5 Conclusion & Policy Implications

5.1 Conclusion

This study investigates the predictive performance of traditional econometric models and machine learning methods in forecasting China's inflation and urban registered unemployment rate. Using a comprehensive set of macroeconomic and monetary policy variables, the paper compares OLS-based forecasting models with the XGBoost algorithm under a

unified out-of-sample evaluation framework.

The empirical results suggest several key findings. First, for inflation forecasting, the OLS model demonstrates relatively stable short-term predictive performance, particularly during periods of low inflation volatility. However, its linear structure limits its ability to capture complex interactions and potential nonlinear dynamics in inflation formation. In contrast, the XGBoost model, while exhibiting larger short-horizon forecast errors in certain low-inflation years, generates a substantially higher inflation forecast for 2025. This result reflects the model's sensitivity to a broader information set, including upstream prices, real economic activity, and monetary policy indicators, and suggests the presence of medium-term inflationary pressures that linear models may underestimate.

Second, for unemployment forecasting, the XGBoost model significantly outperforms the OLS benchmark. The markedly lower MAE and RMSE indicate that machine learning methods are particularly well suited for predicting highly persistent and low-volatility macroeconomic variables such as the urban registered unemployment rate. By capturing nonlinear relationships and interactions among inflation, output, and policy variables, XGBoost delivers more accurate and stable forecasts than linear autoregressive specifications.

5.2 Policy Implications

The findings of this study carry several important policy implications. First, the superior performance of the XGBoost model in unemployment forecasting suggests that policymakers may benefit from incorporating machine learning-based tools into their macroeconomic monitoring frameworks. More accurate forecasts of labor market conditions can improve the timing and calibration of employment stabilization policies, particularly during periods of economic uncertainty.

Second, the divergence between OLS and XGBoost inflation forecasts for 2025 highlights the importance of considering a broad information set when assessing inflation risks. Relying solely on linear models may lead to an underestimation of medium-term inflationary pressures, especially when monetary policy transmission and real economic conditions interact in nonlinear ways. Machine learning models can serve as a useful complement to traditional tools by providing early warning signals of potential inflation turning points.

Finally, this study suggests that combining econometric models with machine learning approaches can enhance the robustness of macroeconomic forecasting. Rather than replacing traditional methods, machine learning models should be viewed as an additional layer of analytical support, helping policymakers and analysts form more comprehensive and forward-looking assessments of inflation and employment dynamics in China.

References

- [1] Ting-guo Z., Xin-yue, F. Wei, J. (2026). CPI forecast and model comparison in a data-rich environment. [online] Tju.edu.cn. Available at: <https://jmsc.tju.edu.cn/jmsc/article/abstract/20250801> [Accessed 6 Feb. 2026].
- [2] Feng H. (2023). Analysis and Forecast of CPI in China Based on LSTM and VAR Model. World Scientific Book Chapters, [online] Vol. 10, No. 1(1), pp.339–357. Available at: https://ideas.repec.org/h/wsi/wsichap/9789811267505_0025.html [Accessed 6 Feb. 2026].
- [3] Cao D., Zhao, Y. Xu, X. (2026). Interpretable Nonlinear Forecasting of China's CPI with Adaptive Threshold ARMA and Information Criterion Guided Integration. Big Data and Cognitive Computing, [online] 10(1), pp.14–14.
- [4] Liu, X. Li, L. (2022). Prediction of Labor Unemployment Based on Time Series Model and Neural Network Model. Computational Intelligence and Neuroscience, 2022(8), pp.1–8.
- [5] Higgins P., Zha, T. Zhong, W. (2016). Forecasting China's economic growth and inflation. China Economic Review, 41(25), pp.46–61.
- [6] Boniface Yemba, Tang, B. Kitenge, E. (2024). Monetary policy in China: A Factor Augmented VAR approach. International Review of Economics & Finance, 89(89), pp.975–1008.
- [7] Das, S. Song, W. (2022). Monetary Policy Transmission and Policy Coordination in China. IMF Working Papers, 2022(074), p.1.
- [8] Wang Y. (2023). Prediction of China's Urban Population Unemployment Rate based on Combined Forecasting Model. Prediction of China's Urban Population Unemployment Rate based on Combined Forecasting Model, [online] 4(2692-7608), p.2023.
- [9] Chen T. and Guestrin C. (2016). XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 1(1), pp.785–794.
- [10] Stock J. and Watson M. (1999). Forecasting in#ation. Journal of Monetary Economics, [online] 44(293-335), pp.293–335. Available at: https://www.princeton.edu/~mwatson/papers/Stock_Watson_JME_1999.pdf?utm_source=chatgpt.com [Accessed 15 Feb. 2026].
- [11] Stock J. and Watson M. (2007). Why Has U. S. Inflation Become Harder to Forecast? [online] Available at: https://www.princeton.edu/~mwatson/papers/Stock_Watson_JMCB_2007.pdf?utm_source=chatgpt.com [Accessed 15 Feb. 2026].
- [12] Newey W.K. and West K.D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55(3), p.703.
- [13] Diebold F. and Mariano R. (1995). Comparing Predictive Accuracy. Source: Journal of Business & Economic Statistics, [online] 13(3), pp.253–263. Available at: https://www.ntuzov.com/Nik_Site/Niks_files/Research/papers/mut_funds/Diebold_Mariano_1995.pdf?utm_source=chatgpt.com [Accessed 15 Feb. 2026].
- [14] Xu X., Li, S. Liu, W.-H. (2024). Forecasting China's Inflation Rate: Evidence from Machine Learning Methods. SSRN Electronic Journal.
- [15] Liu, X. Li, L. (2022b). Prediction of Labor Unemployment Based on Time Series Model and Neural Network Model. Computational Intelligence and Neuroscience, 2022, pp.1–8.