The Foreign Exchange Asset Pricing Model Deeply Integrating ARIMA with Decision tree and LSTM

Jiankun Sun [a*], Xin He [b], Weijie Zhang [c], Tianjiao Zhao [d]
* Corresponding author: e-mail: 1183489602@qq.com, [b]e-mail: 714498202@qq.com

[c]e-mail: 1321244032@qq.com, [d]e-mail: 1400249066@qq.com

Harbin Institute of Technology (Weihai), Weihai, China

Abstract. In recent years, with the improvement of economic development level, China's consumer price index has increased year by year. The central bank faces huge inflationary pressure. Many domestic scholars have done a lot of research on the simulation and prediction of inflation behavior. On the basis of previous research results, we selected econometrics, traditional machine learning, deep learning three typical sequence prediction model, to explore their scope of application, using ARIMA, LSTM, decision tree model of China inflation short-term forecast, and according to "sample size preference" and "fluctuation intensity preference" as the characteristics of performance comparison, the results show that the decision tree model prediction effect is better than ARIMA, LSTM inflation prediction model. In view of this, we propose that the decision tree model be used more broadly in the field of inflation rate prediction, thus providing a more valuable reference for formulating macro policies.

Keywords: inflation rate, decision tree model, genetic algorithm, ARIMA model, LSTM model

1. Introduction

Inflation rate refers to the rate of increase in the general level of prices over a certain period of time, and reflects the extent of inflation. Most countries in the world take the consumer Price index (CPI) as the inflation index, and formulate monetary policies and related macroeconomic policies according to the change of inflation rate, so as to achieve the macroeconomic goal of price stability. A stable macroeconomic environment is conducive to investment growth and economic development, while inflation will increase the instability of the economy and seriously damage the long-term development of the macro economy. Macroeconomic policy makers often pay close attention to the changing rules of the inflation rate. So forecasting inflation has important practical implications.

Domestic and foreign scholars have done a lot of research on inflation rate forecast. In terms of inflation forecasting methods, four types of models are mainly studied. First, Phillips curve model. Relevant studies include Stock et al.[1] and Li Dongdong[2]. Second, VAR model. Relevant studies include Fackler[3] et al., Wu Jingshun et al.[4]. Third, ARIMA model and other time series models, related studies include Andrew et al.[5], Zhu Xinrong etal.[6]. Fourth, BP model and other neural network models, related researches include Nakamura[7] and Xue Ye et al.[8].

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.2-12-2022.2328736

In forecasting inflation rate, there are mainly linear models and relatively few studies on non-linear models. Xia rongyao (2009) constructed ARIMA model and concluded that ARIMA (1,1,10) model was more accurate in short-term prediction of CPI in China by comparing SC and AIC criteria[9]. Chen Yanbin (2021) used the LSTM model to forecast China's inflation rate, and concluded that the LSTM model showed good performance in predicting China's inflation rate[10]. At present, there is no literature in China that uses decision tree machine learning model to forecast China's inflation rate. This paper aims to enrich the research results of inflation rate prediction and provide theoretical reference for inflation rate prediction, making macro policies and maintaining stable economic operation.

2. Index selection and data description

This paper analyzes the influencing factors of inflation rate from the theoretical level, and draws on the practice of Chen Yanbin et al. in the relevant literature to construct an index system containing 4 aspects and 7 variables when forecasting CPI: Consumer price index (CPI), which represents the inflation rate, is the explained variable to be predicted in this paper; "Narrow money (M1) and broad money (M2) reflect the impact of money supply on inflation; Industrial value added (IVA) and producer price index (PPI) reflect the impact of production activity level on inflation; Retail sales of consumer goods (MSR), which reflect the impact of household consumption on inflation; Usd/RMB (USDCNY) to reflect the effect of exchange rate changes on inflation.

All indicators in this paper uniformly use monthly data, and relevant data are from Tongdaxin financial terminal and Choice financial terminal. Based on data availability, when predicting CPI, the samples selected in this paper are from May 2012 to July 2022, which contain 101 sample observations.

3. Model setting and solving

3.1 The overview of ARIMA model

Autoregressive moving average (ARIMA) model refers to the model established by transforming a set of non-stationary time series into stationary time series by difference and then regressing the variables on the values of their own lag terms and random error terms. It is the most common model used for time series prediction in statistical models.

The advantage of using ARIMA model for forecasting is that it can reduce the accumulated error of other variables based on the time series forecasting model, and it has high accuracy to forecast inflation rate by using the historical data of CPI. The disadvantage is that the assumption of ARIMA model prediction is strict, the time series data is required to be stable, and the law cannot be captured for unstable data. Stock data, for example, cannot be predicted by ARIMA because it is volatile and often subject to additional factors such as policy and news. ARIMA by nature can only capture linear relationships, not nonlinear relationships, and lacks consideration of other factors in the environment.

Steps of analysis:

1) Unit root stationarity test

In order to reduce the influence of time series autocorrelation and heteroscedasticity without reducing the accuracy of ARIMA model, the monthly CPI data from January 2001 to July 2022 were logized. The first premise of using ARIMA model to forecast CPI is to require CPI time series to remain stable. This paper selects ADF test method for testing.The ADF stationarity test is shown in Table 1.

Table 1 ADF stationarity test

Sequence variable	ADF	1%	5%	10%	P	conclusion
LNCPI	-3.92293	-3.457438	-2.87346	-2.57312	337.3932	stationary
DLNCPI	-7.50321	-3.457215	-2.87336	-2.57307	349.7326	stationary

2) Determination and identification of model parameters

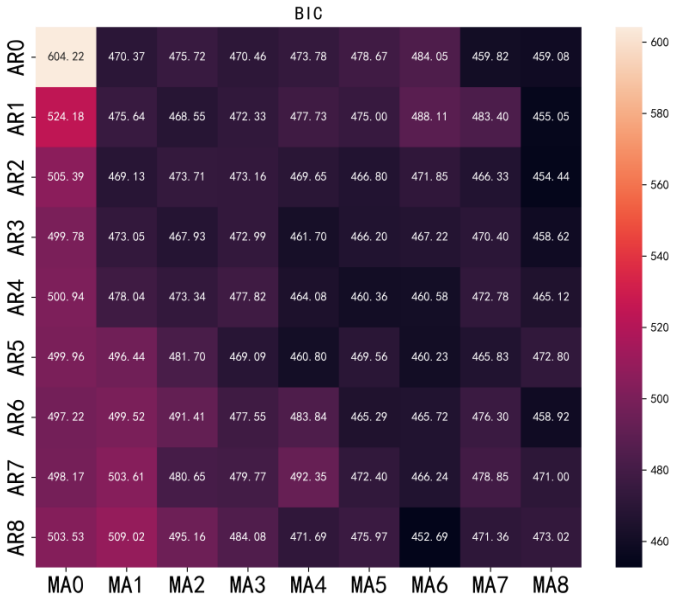

Figure 1 Heat maps based on the BIC criterion

Through ADF stationarity test, it can be concluded that DLNCPI time series is stationary, so it can be determined that d=0 in one of the parameters of ARIMA (p.d.q) model. The BIC criterion is used for analysis to determine the values of p and q parameters in the model. The calculation formula is BIC=kln(n)-2ln(L), where k is the number of model parameters, n is the number of samples, and L is the likelihood function. The hotgraph is shown in Figure 1. As shown in the figure, the best values p=8 and q =6.

3.2 The LSTM model

Long Short term Memory network (LSTM) model is essentially a specific form of recurrent neural network (RNN). LSTM model solves the problem of RNN short-term memory by adding a threshold on the basis of RNN model, so that the recurrent neural network can truly and effectively use the long-distance timing information. When using LSTM model to predict inflation rate, there will be overfitting phenomenon, and the low fitting degree in the face of new samples will affect the forecast results. With the deepening of the layers of neural network, the optimization function is more and more likely to fall into the local optimal solution, and this "trap" is more and more deviated from the true global optimal solution. The performance of deep networks trained on limited data is not as good as that of shallow networks.

Steps of analysis:

1) Selection of training method and optimizer

In this paper, the common practice of deep learning neural networks is adopted to select 80% of the data for model training, which avoids the slow training process and over-fitting problems caused by too many sample points. As far as the optimizer is concerned, in order to make the LSTM model converge quickly, this paper adopts the adaptive learning rate algorithm RMSProp optimizer for optimization training.

2) Setting of hyperparameters and selection of network structure

In this paper, with reference to the common partitioning methods of existing studies, a total of 136 monthly data from November 2005 to November 2018 were used as the training set of the model in the initial training, and 34 monthly data from December 2018 to July 2022.07 were used as the validation set of the model. The training set data is used to help the model update the weight and determine the model structure, while the validation set data is mainly used to evaluate the model effect. In addition, since the dynamic rolling prediction method is adopted in this paper, in the subsequent prediction process, the training set expands continuously with the increase of the sample size, which is the input data of equal step size achieved by moving the training set window. The target that the model needs to debug includes the number of neurons in the LSTM layer, the number of training periods, etc. During training, the Mini-Batch Adam algorithm was used to update the model, and then the model accuracy was measured according to the mean square error (MSE) of the model after fitting all variables on the validation set, and the model result with the highest accuracy was selected. The network structure used to predict CPI in this paper is shown in Table 2:

Table 2 Table of LSTM parameters set for each layer

Layer(type)	Shape of the output	Parameter
LSTM_21	(None, 3, 64)	18432
Dropout_21	(None, 3, 64)	0
LSTM_22	(None, 3, 32)	12416
Dropout_22	(None, 3, 32)	0
LSTM_23	(None, 32)	8320
Dropout_23	(None, 32)	0
Dense_7	(None, 1)	33

Total parameter: 39,201
Trainable parameters: 39, 201
Untrainable parameter: 0

3) Dynamic prediction process

$y_{\tau+h}$ are on behalf of forecasts in the τ phase on inflation rate of the $\tau + h$ phase, $y^0_{\tau+h}$ means the observed values of $y_{\tau+h}$. In order to better analyze the prediction effect of time series data, this paper focuses on the dynamic rolling prediction results of the model. h is set to 1 when the adjacent step prediction is performed.

Specific forecasting steps are as follows: First step, take November 2005 to April 2006 as a sample to forecast the value of the next period, that is, the value of the inflation rate in May 2005. The second step is to expand the sample interval from November 2005 to May 2006, and take December 2005 to May 2006 as the sample to forecast the inflation rate in June 2006. The prediction procedure is the same as the first step. The third step is to further expand the sample interval from November 2005 to June 2006, and move the input data window to take December 2005 to June 2006 as the sample data to predict the value of the next period as the sample to predict the value of the inflation rate in July 2006. Repeat the above extended sample rolling prediction process until the prediction of all samples is complete.

3.3 The decision tree

Decision tree is an inductive learning algorithm for single classification regressors, which consists of root nodes, leaf nodes and non-leaf nodes. The decision tree generates the path from the root node to the leaf node by regression analysis on the training set and analyzes the path rules. Categorize or predict new data based on path rules. Decision tree method is capable of handling large sample size data, suitable for high dimensional data, simple easy to understand, there are many kinds of characteristics of coupling and can avoid the influence of the relevant relationship between characteristics of advantages, such as, therefore, to have a certain correlation based on multiple features to predict the rate of inflation has good applicability, and the method has quick computing speed, the advantage of reuse and strong generalization.

Steps of analysis:

1) Correlation analysis

Pearson coefficient was used to divide all explanatory variables. In this paper, USDCNY, M1, M2, VAI, PPI, MSR were taken as explanatory variables, and CPI was taken as explained variable. Pearson correlation was detected by Python language, and the correlation results between explanatory variables and explained variables were obtained. Then, highly correlated independent variables were eliminated according to the size of Pearson's coefficient to reduce variable complexity. The coefficient results are shown in Table 3.

Table 3 The Pearson coefficient of the index

Index name	Abbreviation of index	Pearson coefficient
Producer price index	PPI	0.465
Industrial increase volume	IVA	0.412
Broad money supply	M2	-0.291

Retail sales of consumer goods	MSR	0.281
Narrow sense of money supply	M1	-0.269
USD/RMB	USDCNY	0.114

Pearson coefficient of known characteristic variables greater than 0.8 is considered as highly correlated, between 0.5 and 0.8 is considered as moderately correlated, between 0.3 and 0.5 is considered as weakly correlated, and less than 0.3 is considered as almost irrelevant. Therefore, the 7 indicators selected in this paper are retained as the final explanatory variables of the model.

2) Genetic algorithm parameter optimization and analysis

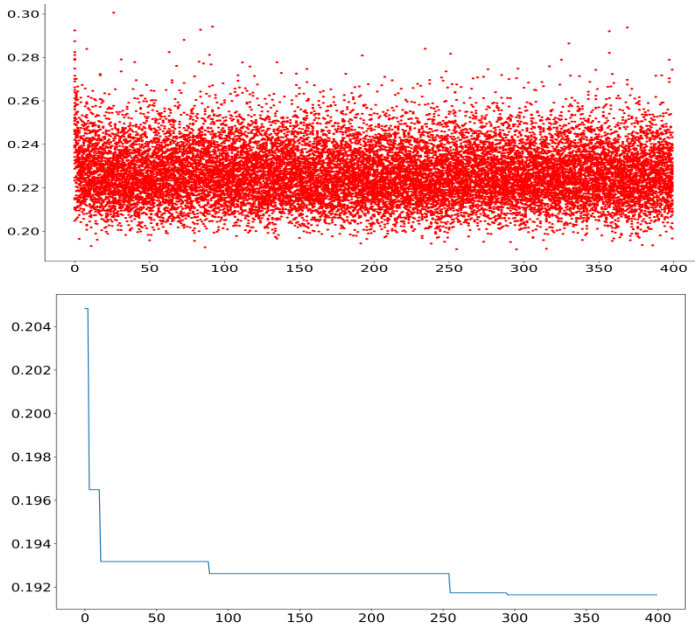

Figure 2 The distribution of loss value of each round of training in random forest which is optimized by genetic algorithm (the figure above) and the minimum loss value of each round of iteration (the figure below)

Enter the filtered explanatory variables. According to the number of explanatory variables, the values of max_depth, min_samples_split and max_leaf_nodes of the decision tree range from 10 to 30,5 to 15, 5 to 15. Firstly, the parameters are optimized by genetic algorithm, and the effect of "natural selection" is achieved by defining the fitness function (the MSE value of the verification set is regarded as loss). The smaller the fitness value, the smaller the individual network error, and the higher the adaptability.

As shown in Figure 2, it can be seen that it has converged to the most value at 130 rounds, when max_depth=16, min_samples_split=6, max_leaf_nodes=15, val_loss_min=0.226.

4. Prediction result analysis

4.1 The result of Arima

From the experimental effect (As shown in Figure 3), the accuracy of ARIMA model still has great room for improvement, especially in some fluctuating time periods, it is difficult for ARIMA itself to play its advantages in linear fitting, and it is difficult to dig out its due rules for non-linear signals. The overall MSE was 0.2773, which was larger than that of the decision tree. Possible reasons were as follows: single data and insufficient explanatory power.

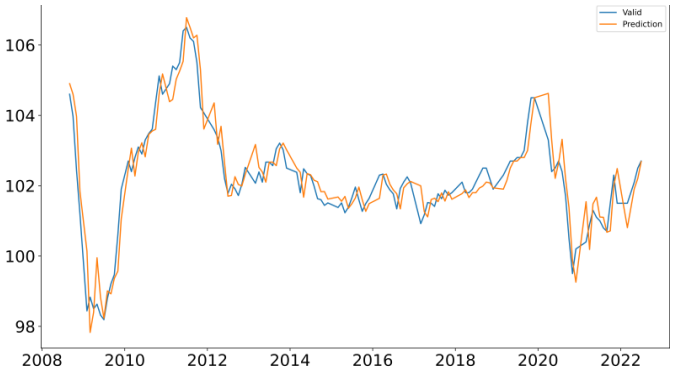

Figure 3 Arima forecast result diagram

4.2 LSTM training results

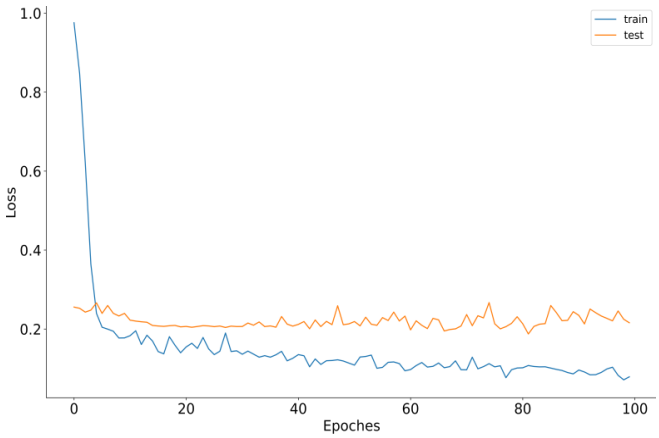

Figure 4 The loss values of both the training set and the validation set during the LSTM training procedure

As shown in Figure 4, the fitting effect of LSTM is greatly improved than that of ARIMA, and it still shows a good prediction effect in the marked period of high volatility and complex signals. However, in the stationary period near the high fluctuation period, its accuracy will be affected by the parameters of the surrounding non-linear fitting, resulting in a decrease in accuracy.

As shown in Figure 5, there were 100 epochs in total, and the model was overfitted, with MSE =0.26 on the training set and 1.14 on the test set. Possible causes of overfitting: too few data.

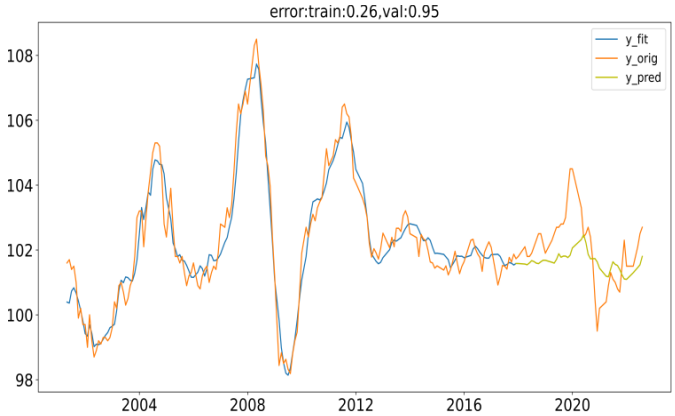

Figure 5 LSTM forecast result diagram

4.3 Decision tree result

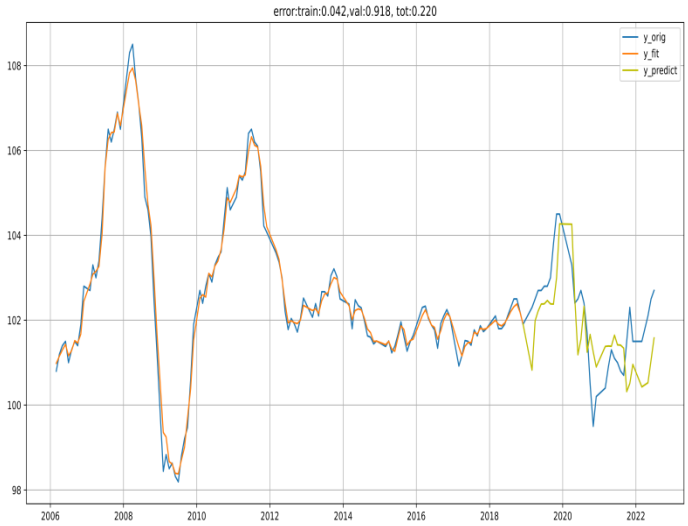

Figure 6 Decision tree forecast result diagram

As shown in Figure 6, the MSE on the training set was 0.042, the MSE on the validation set was 0.918, and the overall MSE was 0.220.

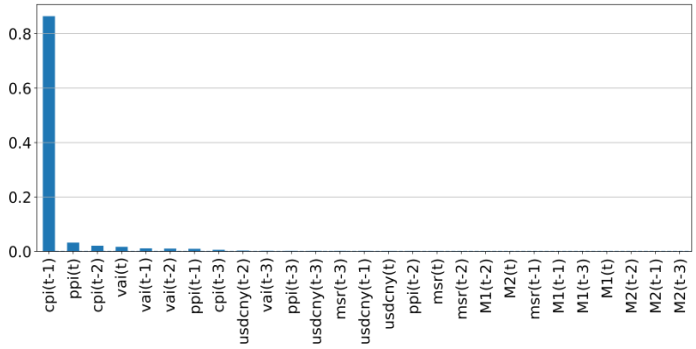

Figure 7 Factor importance

In terms of explanatory power, the first-order lag term of CPI is strongly correlated with factors such as MSR and USDCNY, which provide certain explanatory power.Factor importance is shown in Figure 7.

Table 4 Characteristic comparison of measurement timing model, deep learning, and traditional machine learning model (take ARIMA, LSTM, and decision tree, for example)

	ARIMA	LSTM	decision tree
Sample size preference	Suitable for samples with small sizes	Suitable for large-scale samples	There was no significant preference for the sample size
Signal feature preference	Strong ability to mine the linear features	Strong ability to mine nonlinear and fluctuating signals	Both linearity and nonlinearity can mine well and have strong interpretable power

As shown in Table 4, in terms of sample size preference, ARIMA is suitable for small size samples, LSTM is suitable for large-scale samples, and the decision tree has no obvious preference for sample size; ARIMA is strong for linear feature mining, LSTM for nonlinear and fluctuating signal mining, and decision tree linearity and nonlinear mining.

5. Conclusion

In this paper, we use ARIMA model, LSTM model and decision tree model to forecast and compare the short-term inflation rate in China. When forecasting the trend of CPI, an index system of 7 variables including money supply, production activity level, household consumption and exchange rate is constructed from the theoretical level. In this paper, the decision tree model was pioneeringly used to predict the trend of China's CPI. The importance ranking of random forest was used to screen out the important variables affecting the inflation rate. The genetic algorithm was used to adjust the parameters of the decision tree and improve the tuning

speed, so that the prediction effect was improved. The combination algorithm has a good effect and high accuracy on the short-term prediction of CPI.

Based on the research results and literature analysis, this paper suggests that in the future, in addition to strengthening the research on CPI prediction by decision tree model, the research on core CPI prediction by decision tree model should also be strengthened. Because CPI includes parts of food and energy prices that are easily disturbed by random factors such as climate and season, it has strong volatility, which is not conducive to the decision-making judgment of macro policy makers and may lead to the risk of overreaction to temporary factors of inflation. The core CPI, which strips out food and energy prices, removes or subtracts the volatile components of inflation and better reflects movements in the general price level. Therefore, in the next step, the application of decision tree model in core CPI prediction should be further strengthened, so as to provide more valuable reference for macroeconomic policy making.

Reference

[1] Stock, J. H., Watson, M. W. Phillips Curve Inflation Forecasts [R]. NBER Working Paper No. 14322, 2008. 1-82.

[2] Li Dongdong. (2021). Study on the Long-term Relationship between Inflation, Expected Inflation and Output Gap The - An empirical study of Phillips Curves based on Chinese empirical data. China's Collective Economy (12), 28-30.

[3] James S.Facker W, Douglas MclMliuin. lnflation Forecast Targeting: An Altermative Approach to Estimating the lnflation-Output Vvariability Tradeofif [J].Southern Economic Jounnal, 2011.78(2) :424-451

[4], Wu Kam-shun. (2013). The Chinese core inflation rate was estimated using structural VAR models. The Southern Economy (02), 41-55.doi:10.19592/j.cnki.scje. 2013.02.005.

[5] Andrew Ang, Geert Bekaert, Mlin We.Do Macro Variables, Asset Markets Or Surveys Forecast lnfation Better?[J].Nber Working Papers.2005(3) :21-28.

[6] Xinrong Zhu & and Peng Chao. (2017). Short-term inflation prediction based on the ARIMA model with a structural breakpoint test. Journal of Guangxi University of Finance and Economics (02), 36-43.

[7] Nakamura, E. Inflation Forecasting Using A Neural Network [J]. Economics Letters, 2005, 86(3): 373-378.

[8] Xue Ye, Lin Qizhu, Ren Yao. The predictive model of inflation risk in China, is based on the decision tree-BP neural network [J]. Economic issues, 2016, (1): 82-89.

[9] Yan Zhengxu, Qin Chao & Song Gang. (2021). Stock price model of stock price model based on Pearson feature selection. Computer Engineering and Application (15), 286-296.

[10] Chen Yanbin, Liu Lingjun, Chen Xiaoliang. Study on Inflation Rate Prediction Based on Deep Learning LSTM Model [J / OL]. Research on financial issues: 1-14