Advancing the comprehension of consumer price index and influencing factors: insight into the mechanism based on prediction machine learning models

Qian Cui[1,a], Shuai Rong[2,b], and Boyang Zhang[2,c]

[1]School of Public Administration and Law, Liaoning Technical University, 123000 Fuxin, Liaoning, China

[2]School of Public Administration and Law, Liaoning Technical University, 123000 Fuxin, Liaoning, China

[a]Qiancui0128@163.com, [b]348901208@qq.com, [c]200103200221@163.com

Abstract. The consumer price index (CPI) is an important indicator to measuring inflation or deflation. Its changes are closely related to residents' lives, and also affect the direction of national macroeconomic policy formulation. In recent years, the combination of economics with mathematical models such as machine learning and macro-statistics has gradually become a hot topic. In this paper, the impact of different types of CPI on China's overall CPI was discussed. Machine learning prediction and correlation analysis of various types of influencing factors and CPI. The machine learning model of the regression decision tree process predicted CPI more accurately. Spearman correlation analysis showed that CPI was mainly positively related to goods and services, living, medical care, food, tobacco and alcohol, clothing and so on, while CPI was mainly negatively related to the transport and communication of residents.

Keywords: consumer price index; machine learning; Spearman.

1. Introduction

The consumer price index (CPI) is the fluctuation trend of the price level of goods and services purchased by residents [1]. The purpose of compiling the CPI is to analyze and study the impact of price changes on socioeconomic development and residents' living standards [2]. In recent years, with the increasing economic pressure, how to analyze the influencing factors of CPI scientifically and intuitively has become a hot issue.

Machine learning techniques are based on computer algorithms and models that can be automatically developed and empirically predicted in the right context [3]. Machine learning has high prediction accuracy in complex nonlinear problems, reducing unnecessary labor and resource consumption. Machine learning also has applications in public administration [4, 5].

In this paper, the regression decision tree model (RDT) and Artificial neural network model (ANN) in the supervised model were used for comparison prediction. The total CPI and different categories of CPI from 2010 to 2022 were obtained through China's national database, and eight inputs were set, namely food tobacco and alcohol, clothing, living, goods and services, transportation and communication, education, culture and entertainment, medical care, and other goods and services, and the value of CPI was taken as output. In the dataset, 70% were randomly set as training data and 30% as test data. The main formulas of the above three models are shown in eqs. (1) and (2).

$$\mathrm{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{y_{R_j}})^2 \tag{1}$$

$$\mathrm{net}^k = \sum_{j=1}^{m_0} W_{i,j}^k Y_j^{k-1} + b_i^k$$

$$Y^k = f^k(\mathrm{net}^k) \tag{2}$$

In addition, the Spearman correlation analysis matrix was used to study the correlation between various influencing factors. Based on the above methods, this paper explored the potential mechanism of CPI control by various influencing factors.

2. Results and discussion

2.1 Machine learning prediction process

2.1.1 Machine learning prediction analysis

ANN, as a traditional model, is often used in the field of economics, while RDT model is rarely applied. In this paper, Fig. 1a to Fig. 1d compared the predicted and measured values of the RDT and ANN models. The prediction performance of RDT is significantly better than that of ANN model.

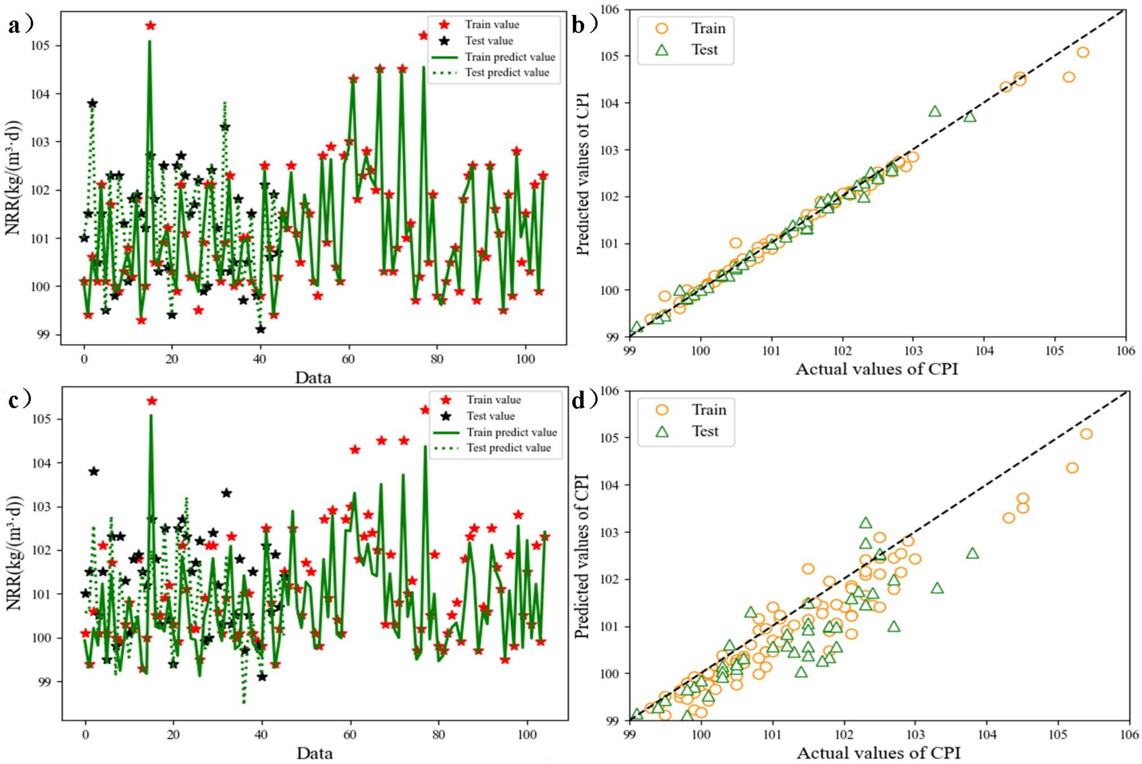

Fig. 1. Performance of two prediction models: RDT (a, b) and ANN (c, d).

Moreover, as can be seen from Table 1, the highest R2 of the RDT model was 0.991, which was much higher than that of the ANN model. In addition, the training RMSE of RDT and ANN models can reach 0.118 and 0.485, respectively, and the testing RMSE was 0.125 and 0.766, respectively. whether training or testing RMSE of the RDT model was smaller than the ANN model. Therefore, combining the above three aspects, this paper finds that the RDT model was the most satisfactory. Based on these results, the calculation principle of the RDT model was more consistent with the change rule of CPI under the influence of different consumption types.

Table 1. Fitting parameters of RDT and ANN**.**

Model	Training R^2	Testing R^2	Training RMSE	Testing RMSE
RDT	0.991	0.987	0.118	0.125
ANN	0.855	0.508	0.485	0.766

2.1.2 Input importance analysis of RDT model

Due to the good predictive performance of the RDT model, the importance of input parameters was further analyzed by SHAP. According to Fig. 2, it can be seen that for the RDT model, the value of tobacco and alcohol was the most important inputs that affect the accuracy of the model. Secondly, the value of education, culture and entertainment also affected the prediction performance. Relatively speaking, the impacts of living and clothing were not significant, which might be because these inputs were daily needs and the changes were not significant.

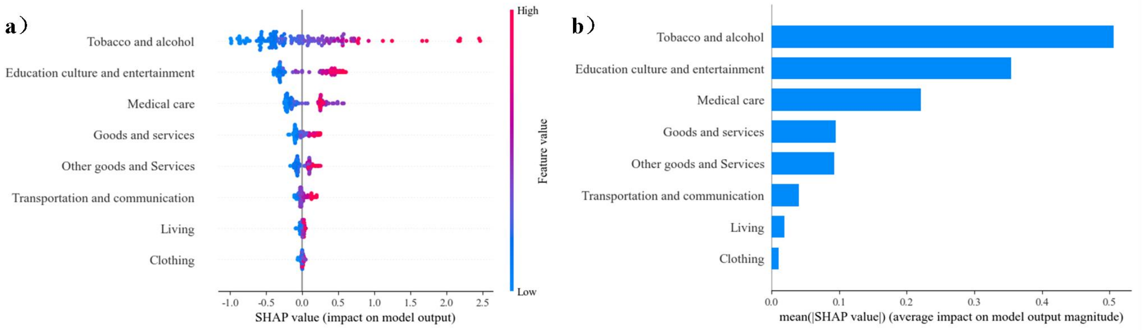

Fig. 2. The SHAP analysis (a) and input importance analysis (b) of the RDT model.

2.2 Correlation analysis between the CPI and other factors

Spearman analysis was used in this paper to conduct a correlation of the CPI of various influencing factors (Fig. 3). In Fig. 3, the real and dashed lines were positively and negatively correlated, respectively. The strength of the correlation was indicated by the thickness of the line, that was, the thicker the line was, the stronger the correlation was.

As shown in Fig. 3, CPI was positively correlated with seven labels, including medical care and education. It was indicated that the rapid development of digitalization had expanded the size of consumer groups in the era of big data, making it easier for people to obtain daily necessities, and increasing consumption of various commodities and services. But there was a negative correlation between CPI and transportation and communication. The main fluctuation in transportation came from fuel. The sustainable development of the natural environment and national energy security were under great pressure. The change in oil price had a relatively obvious lag effect on China's economic growth, which restricts the fluctuation of CPI.

In addition, there is a clear correlation between other parameters. For example, living was positively correlated with education, culture and entertainment. Clothing was also positively correlated with goods and services. With the development of the economy and the improvement of people's living standards, the quality of people's lives depended mainly on these parameters.

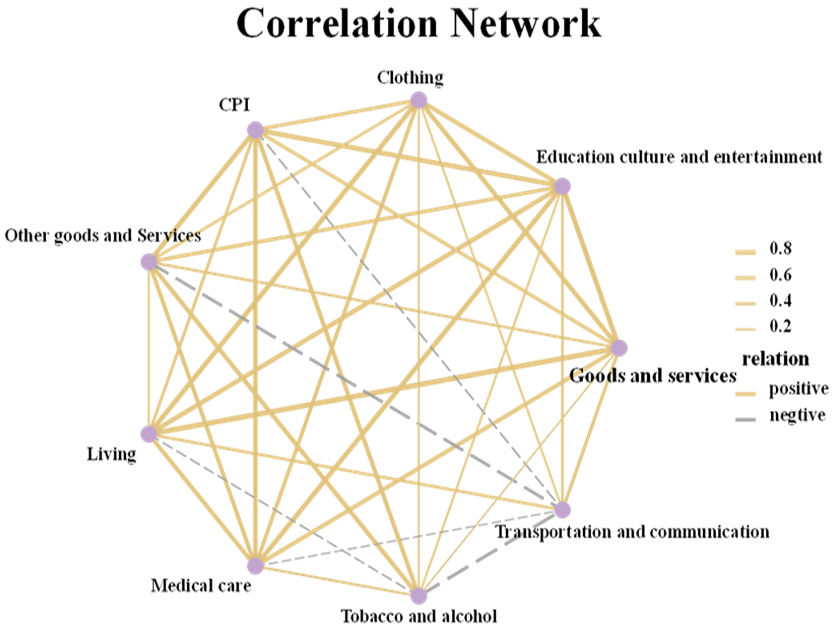

Fig. 3 Spearman Correlation heat map for CPI and other factors.

3. Summary

This paper explored the impact of different factors on China's overall CPI from two aspects: machine learning prediction and correlation analysis. First of all, machine learning results showed that the RDT model can well predict the changing trend of CPI under the influence of various factors. And the value of tobacco and alcohol was the most important inputs that affect the accuracy of the model. Secondly, Spearman correlation analysis found that CPI was mainly positively correlated with commodities, services and education, while CPI was mainly negatively correlated with transportation and communication. This paper provided technical guidance for the study of the relationship between CPI and various influencing factors.

References

[1] White, A. G. Measurement biases in consumer price indexes. international statistical review, 1999, 67(3), 301-325.

[2] Feng, X., Xu, Y. F., Ni, G. Q., & Dai, Y. W. Online leasing problem with price fluctuations under the consumer price index. journal of combinatorial optimization, 2018, 36(2), 493-507.

[3] Zheng, J. H., Cole, T., Zhang, Y. X., Kim, J., & Tang, S. Y. Exploiting machine learning for bestowing intelligence to microfluidics. biosensors & bioelectronics, 2021, 194, 113666

[4] Syah, R., Al-Khowarizmi, A., Elveny, M., & Khan, A. Machine learning based simulation of water treatment using LDH/MOF nanocomposites. environmental technology & innovation, 2021, 23, 01805.

[5] Jorgensen, B. Construction of multivariate dispersion models. Brazilian journal of probability and statistics, 2013, 27(3), 285-309.