Computer Life
ISSN: 1819-4818 | Vol. 12, No. 2, 2024

KNN model based financial revenue forecast of Guangzhou

Ganzhou Wu

School of science, Guangdong University of Petrochemical Technology, Maoming 525000, China.

Abstract: It is mainly carried on the analysis by the Guangzhou 2013-2020 financial revenue situation, firstly, in Guangzhou The Bureau of statistics downloads the data and preprocesses the data to determine the main factors affecting Guangzhou's fiscal revenue, then the KNN model is established to forecast the fiscal revenue of Guangzhou in 2021. The numerical results show that the method is effective.

Keywords: Fiscal Revenue; Forecast; KNN Model.

1. Introduction

Liu Xiaoying [1] studies the stock price trend with the K-nearest neighbor classification method, and draws the conclusion that it is feasible to use the KNN method based on the data-driven principle to research the stock price trend; The KNN model [2]was used to predict WTI crude oil price, and the results showed that KNN model was better than Arima model and neural network autoregressive model in predicting WTI crude oil price, Tang Zhenyu [3]used multi-modal depth KNN to predict the total survival time of GBM Glioblastoma multiforme, which was the first study to predict the OS time of GBM patients in the framework of DL, and Hong [4] used KNN SUAFA to predict the hourly energy consumption of community buildings, and achieved good prediction results. Yan Feifei [5] used time series analysis method to forecast Guangzhou fiscal revenue data, and obtained the forecast value of local fiscal revenue; Liu Qian's [6] analysis and forecast of the influencing factors of financial revenue in Jiangsu, Zhejiang and Shanghai studies the finance with the elastic network-RBF model, the combination of grey GM (1,1) and neural network is used to forecast the fiscal revenue of Jiangsu, Zhejiang and Shanghai in 2019 Wang Shouying's research on fiscal revenue forecasting based on data mining -- taking Jinan as an example [7], studies Jinan's fiscal revenue with two forecasting methods, the first method uses GM -LRB-1,1) model BPd BP neural network model, and the second method uses the combination model of grey neural network, both of which have good prediction results Jiang Feng [8] and others used Lasso-GRNN neural network model to forecast local fiscal revenue, the results show that Lasso-GRNN model is better than Lasso-BP and Lasso-RBF neural network in both convergence speed and prediction precision Hong-bin W [9] studies the key influencing factors of local fiscal revenue in China, and obtains the main influencing factors of tax revenue; Yiting Wang [10] studies the impact of real estate tax reform on local fiscal revenue based on Shanghai; Yifu Sheng [11] and others used the method of model fusion to forecast Hunan province's fiscal revenue. The gray neural network model with Lasso regression has higher precision than the traditional model Hui Liu [12] provides two improved grey models to accurately forecast the total revenue of Guizhou government, and draws a conclusion that it is feasible and effective to use the grey combination model to forecast the local fiscal revenue. Botri Valerijal [13] used the main time series model to study the local fiscal revenue, and the results show that the econometrics method is a better way to predict the local fiscal revenue.

First, I downloaded the Excel Data from the official website of the Guangzhou Municipal Bureau of Statistics and merged them. Then, I preprocessed the Guangzhou municipal fiscal revenue data from 2013 to 2020. Secondly, normalize the index, split the training set and test set, build KNN model, fit and forecast the fiscal revenue of Guangzhou in 2021.

2. Data pre-processing

This article is based on the data of Guangzhou's financial revenue from 2013 to 2020 on the official website of the Guangzhou Municipal Bureau of Statistics. As the annual data of the Guangzhou Municipal Bureau of Statistics is updated around mid-may, the official data for 2022 are not available at present, the author uses the data of 8 years from 2013 to 2020 to train the model and forecast the financial revenue of Guangzhou in 2021.

This article is based on direct data from the official website of the Guangzhou Municipal Bureau of Statistics. I collected the annual fiscal revenue data from 2013 to 2020 by downloading the Excel table directly. The row labels of the data tables were consistent each year, the column labels are for different years, and you can merge the annual Excel into a single Excel table by using Excel's “Merge cells” and “Copy and paste” functions. At the same time, looking at the consolidated data, the business tax disappeared after 2016. That is because the business tax was abolished in 2016, and the business tax was replaced by value-added tax. That is, the business tax was abolished after 2016, since vat is the only consumption tax, only VAT data should be considered when looking at fiscal revenues after 2016. The author studies the fiscal revenue data from 2013 to 2020, and combines the business tax and the value-added tax before 2015, because before 2015, the business tax and the value-added tax are two independent taxes, so it makes sense to add up the data for the two types of taxes to get the total amount of tax revenue.

In data analysis, the distribution of data needs to be tested to determine whether the data conform to the normal distribution. Normal distribution test usually uses statistical methods, such as chi-square test and so on.

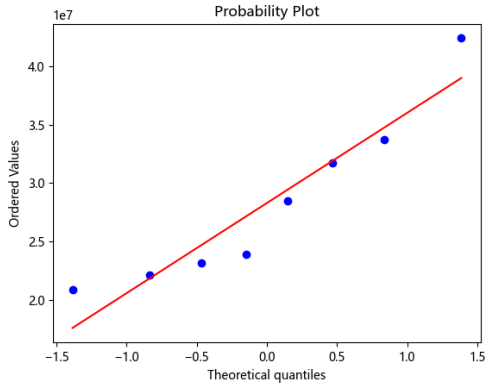

Figure 1: Normal distribution test chart

Figure 1 shows that the total local fiscal revenue conforms to the normal distribution and has a good Linear independence.

3. KNN model

KNN method as machine learning technology consists of three basic elements, k-value selection, distance measurement and classification decision rules. The basic principle is: given a known class training data set to calculate the distance between the sample data and the data to be classified and look for K nearest training examples. Then the majority of the cases to be classified to vote, predict the classification of cases. KNN method does not need explicit learning process, only needs to input case feature vector and output case classification. KNN method is suitable for multi-class classification problems. Although KNN method relies on limit theorem in principle, only a few adjacent samples are considered in the classification decision, thus avoiding the case of unbalanced samples. In addition, KNN method is used to classify the samples whose class domains overlap or overlap greatly.

The basic principle of the KNN method is that for a new input instance, the distance between the instance and each training instance is first calculated in the training data set of the known class tags. Then select the K training instances closest to the instance, count the number of each category in the K instances, and classify the instances into the category with the largest number. Therefore, KNN method is a neighbor-based classification algorithm, which relies on the nearest neighbor instances in the training dataset.

Input: Training Data Set

$$T=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_N,y_N\right)\right\} \tag{1}$$

Among them, $x_i \in X \subseteq R^n$ representing the feature vector, $y_i \in Y=\left\{c_1,c_2,\cdots,c_k\right\}$ representing the class label.

Output: dependent variable Y.

based on a known distance metric T, find the nearest instance of sum k in a given training set, and write the neighborhood that contains k the instance of this point as $N_k(x)$.

In $N_k(x)$, y class labels in which the minority is determined by the principle of majority x.

$$y=\arg\max_{c_j}\sum_{x_i\in N_k(x)} I\left(y_i=c_j\right), i=1,2,\ldots,N;\ j=1,2,\ldots K \tag{2}$$

I is an indicating function, where $y_i=c_j$, $I=1$ at the time, otherwise $I=0$.

In this paper, MAPE, RMSE, R^2 are used as the performance evaluation indexes of KNN model. MAPE denotes mean absolute percentage error, RMSE denotes root mean square error, and R^2denotes goodness of fit.

(1) mean absolute percentage error (MAPE)

The mean absolute percentage error (MAPE) is the absolute value of the difference between the measured value and the true value divided by the average of the true value and multiplied by 100. It is a commonly used indicator to evaluate the accuracy of prediction models, usually used for the prediction of percentage variables.

The formula is:

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-y_i}{y_i}\right| \tag{3}$$

(2) root mean square error (RMSE)

Root-mean-square error (RMSE) is the square root of the square of the difference between the predicted value and the true value. It is a commonly used index to evaluate the precision of prediction model, usually used for continuous variable prediction.

The formula is:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-y_i\right)^2} \tag{4}$$

(3) decision coefficient (R^2)

The decision coefficient (R^2) is the ratio of the sum of squares of the difference between the predicted value and the true value to the sum of squares of the total difference. It is used to evaluate the proportion of variance explained by prediction models, and is usually used to compare the fitting degree of different prediction models. The value of R^2ranges from 0 to 1, and the closer to 1 the better the fit of the model.

The formula is:

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-y_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{5}$$

In formula (3)(4)(5), y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the sample number.

4. Results

This is selected the data of Guangzhou's fiscal revenue from 2013 to 2020 to analyze and model, carries on the correlation analysis to the data which has already been cleaned, determines the index which affects the fiscal revenue, and carries on the normalization processing to the index, it is divided into training set and test set, then training the model with training set, finally using the test set to forecast the financial revenue of Guangzhou.

(1) data sources: download the 2013-2020 Excel data from the official website of the Guangzhou Municipal Bureau of Statistics and consolidate the data annually. The annual Excel

tables are combined into a single Excel table by "Merging cells.". Table 1 and Table 2 are the dependent and independent variables for this article, respectively.

Table 1: Total revenue of Guangzhou Municipality from 2013 to 2020

Year	Total local revenue (10,000 yuan)
2013	20881374
2014	23188414
2015	23913268
2016	22184824
2017	28448315
2018	31737500
2019	33678876
2020	42374332

Table 2: The Index of financial revenue of Guangzhou

Y	Total local fiscal revenue		
X11	Revenue from the general public budget	X17	Administrative fee income
X12	Value-added tax	X18	Forfeiture of income
X13	Corporate Income Tax	X19	Special income
X14	Personal income tax	X10	Other income
X15	City maintenance and construction tax	X11	Revenue from government funds
X16	Vehicle and vessel tax	X12	Subsidy income from superiors

The data were randomly divided into training set and test set, so that 2015 and 2019 were test set, average absolute percentage error (MAPE) and decision coefficient R^2 were used to evaluate the performance of the model.

Table 3: Comparison of KNN model prediction results with real values

Type	Forecast results	True Value
2015 Year	23188414	23913268
2019 Year	31737500	33678876

Table 4 Error analysis of KNN model

Type	MAPE	R^2
KNN model	4.6%	0.9

As can be seen from Tables 3 and 4, the results of the KNN model show a relatively high accuracy. For the MAPE of 4.6%, it means that the average absolute percentage error of the model is 4.6%, which is relatively small, indicating that the prediction ability of the model is better; for the model with a decision coefficient of 0.91, the average absolute percentage error of the model is 4.6%, it shows that the model can explain about 91% of the dependent variable variation, which is relatively high, indicating that the model fitting degree is good.

5. Conclusion

The results of this paper are as follows: the mean absolute percentage error (MAPE) of KNN model is 4.6% , which is a good fitting effect, it is feasible to use KNN model to forecast the fiscal revenue of Guangzhou.

References

[1] Liu Xiaoying. Research on stock price trend prediction based on KNN method [D] . Northeast Agricultural University. 2015.

[2] Chu dollar, lo oi-chun, Cheung King-shun. Study on crude oil price forecast based on KNN model [J] . Price monthly. 2021(05) ,15-22.

[3] Tang Zhenyu,Cao Hongda,Xu Yuyun,Yang Qing,Wang Jinda,Zhang Han. Overall survival time prediction for glioblastoma using multimodal deep KNN.[J]. Physics in medicine and biology,2022,67(13), 55-62.

[4] Hong Goopyo,Choi GyeongSeok,Eum JiYoung,Lee Han Sol,Kim Daeung Danny. The Hourly Energy Consumption Prediction by KNN for Buildings in Community Buildings[J]. Buildings,2022,12(10), 60-67.

[5] Yan Feifei, Ma Xiaotian, Li Haidi, Wang Tao. Analysis and forecast of Guangzhou fiscal revenue based on data mining technology [J]. Journal of Harbin Normal University Sciences. 2016(01): 31-33.

[6] Liu Qian. Analysis and forecast of the factors affecting financial revenue in Jiangsu, Zhejiang and Shanghai [D]. Hangzhou Dianzi University. 2021.

[7] Wang Shouying. Research on fiscal revenue forecast based on data mining [D] . Shandong Normal University. 2020.

[8] Jiang Feng, Zhang Ting and Zhou Yanling. Forecast of local fiscal revenue based on Lasso-GRNN neural network model [J]. Statistics and decision making. 2018(19): 91-94.

[9] Hong-bin W,Lei D,Yan-fei L. Influencing Factors of Local Fiscal Revenue in China Based on Multiple Regression Model: An Empirical Analysis[C]//Padjadjaran University of Indonesia,University of Electronic Science and Technology of China,American Society for Public Administration ,Chinese Public Administration Journal.Proceedings of 2015 International Conference on Public Administration(11th)(VII) 2015:938-946.

[10] Yiting Wang. Prediction of the Impact of real Estate tax Reform on local Fiscal revenue --- Based on Shanghai[P]. 2022 International Conference on Real Estate, Population and Green Urbanism,2020.

[11] Yifu Sheng,Jianjun Zhang,Wenwu Tan,Jiang Wu,Haijun Lin,Guang Sun,Peng Guo. Application of Grey Model and Neural Network in Financial Revenue Forecast[J]. Computers, Materials & Continua,2021,69(3).

[12] Hui Liu. Prediction of Financial Revenue in Guizhou Province Based on Improved Grey Combination Models[J]. Applied Mechanics and Materials,2014,3183(551-551).

[13]]Botrić Valerija,Vizek Maruška. Forecasting Fiscal Revenues in a Transition Country: The Case of Croatia[J]. Zagreb International Review of Economics and Business;, 2012,15(1),110-120.