

Population Prediction Based on Logistic Regression and Random Forest

Shuzhi Wang^{1, *}

¹ The department of International Education, Guangdong University of Technology, Guangzhou, 511495, China

* Corresponding Author Email: 3119009442@mail2.gdut.edu.cn

Abstract. Due to the continuous improvement of quality of life and more stable social environment, the population is growing rapidly. Among them, population forecasting plays an important role in solving the problems brought by population growth. Based on the data indicators such as population, fertility rate and median age in China's mainland from 1955 to 2022, the logistic regression model is usually used to predict the future population. This model is trained by Hyperopt to obtain the optimal parameters, it can improve the accuracy of the model. In addition, in order to show the differences between different models, this experiment also used random forest model to predict. The experimental results show that the population growth rate will slow down gradually and reach the maximum population in around 2030, after which it will show negative growth. By calculating the mean square error, it shows that the machine learning model can provide learners with accurate population prediction results.

Keywords: Population prediction, Logistic regression, Random forest, Data preprocessing, Optimization parameters.

1. Introduction

Population prediction is based on the existing population size, and considering the social economy, resource, environmental conditions as well as other factors, then make prediction for future population size and structure. Its prediction method is to optimize the parameters of the relevant model through the existing population data, and it is a scientific method for measuring the number and development of the population in a certain period of time [1]. In order to face the increasing number of people, population analysis and prediction has gradually become a main concern to promote social development. The population policy has close relationship to the population that is a kind of policy to control population reproduction. And population policy is established on the basis of a country or region 's current social economy development and the affordability of its resources and environment. It plays an important role in population quantity, structure and quality by affecting the process of population reproduction [2]. Therefore, in our daily life, forecasting populations through data science deserves more attention.

In the last few years, because of the generation of big data and the development of the computer, machine learning is in a rapid development state. In order to find the optimal parameters of the objective function. Most machine learning does not use specific models to assume and limit the relationship between different data, but use the algorithm to directly learns and explores the statistical relationship between the system features and variables of the data [3]. As early as 1920, Raymond et al. had used logistic regression model to analyse the population data of the United States. They found that an S-shaped curve could well fit the population data of the United States from 1790 to 1910. In 1995, Tin Kam Ho proposed the idea of random forest [4]. Until 2001, Breiman systematically expounded random forest and formally proposed the random forest algorithm [5]. Since then, random forest algorithm has become a representative algorithm in machine learning, and widely used in finance, medical and data analysis. In previous research studies, most of them only used logistic regression models to predict population. However, this model cannot take into account the situation of multi-type data, and its versatility is poor. In addition, the decision tree model is a common prediction method, although it can consider more factors and take up less computing resources.

However, this method has obvious errors in the case of unbalanced data indicators, and the error rate on the test data is relatively high. In order to explore more about the correlation between various characteristics and consider the imbalance of some indicators of population data, the advanced algorithm e.g. random forest model that can balance errors and works better with high-dimensional data should be considered.

In this study, the number of Chinese mainland population, fertility rate, median age and other data indicators from 1955 to 2022 that are used to predict the future population. Compared with the traditional prediction method which based on mathematical functions, the parameter modeling through machine learning is more flexible and has higher accuracy. This is the reason why this study choose the logistic regression model and random forest model for population prediction. After using hyper-parameters and obtaining the optimal parameters to optimize model and comparing the model predicted values with actual values in the database. By comparing scatter plots, it can be found that the prediction of random forest is more accurate. The experimental results show that the population growth rate in mainland China is slowing down and will reach a maximum of 1.467 billion around 2030, and the next step will be a period of negative population growth.

2. Method

2.1. Data set description and preprocessing

This study used the population data of China's mainland from Kaggle [6]. The data set provides detailed population data e.g. population size, population density, birth rate. It contains in total of 1,679 population data from 1950 to 2022. This study used a data processing tool called pandas to read the Chinese population data.

Data preprocessing is an important step before data analysis. It can improve the efficiency of program operation and the accuracy of prediction results. Since the data set has invalid data and the format of the data need to be changed, all null values were deleted, and then this study converts string type data into floating point class data.

After preprocessing these data, the data set was divided into training set and test set, 80 % of the data was used as training set and 20 % of the data was used as test set in order to verify the performance of the trained model.

2.2. logistic regression model

Logical regression is one of the most popular models in machine learning. In the experiment, in order to predict the future population accurately, that is very important to consider the influence of natural resources, environmental conditions and other factors on the population. These factors play an important role in blocking the growth of the population. And with the population increases, the blocking effect becomes more and more obvious [7]. The schematic diagram of the logistic regression can be found in Figure 1.

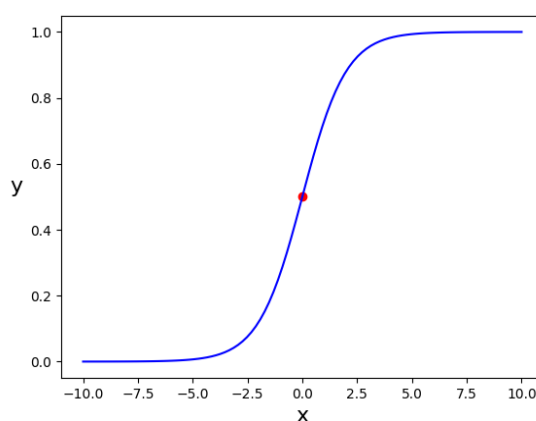


Fig. 1 logistic regression function

2.2.1. Predictive function

From Figure 1, it can be observed that the value range of x in the Sigmoid function from negative infinity to positive infinity, and the range of value is 0 to 1. When x is negative infinity, the function infinitely approaches 0; when it is positive infinity, the function infinitely approaches 1. When $x = 0$, $y = 0.5$.

2.2.2 Establishment and Training of Model

In order to realize training and testing of the model, the experiment is divided into the following steps: 1) Firstly, using StandardScaler function to data normalize and standardize. Then constructed the logistic regression function, and sat parameters to fit the model according to the training data. 2) After training the model, the predicted data is compared with the real data and then calculate the mean square error, and the mean square error is used as an indicator of the optimal parameters for training. 3) Training the defined model, inputting relevant parameters and defining the maximum number of training rounds. After each round of training, the minimum loss and each parameter value were recorded, they were the current optimal parameter of this model. After the training was completed, returned the optimal parameters and outputted the results.

2.3. Random forest model

Random forest is an algorithm based on classification tree, which belongs to ensemble learning in machine learning. It is mainly composed of three parts: the generation of training set, the construction of decision tree and the generation of algorithm [5]. The core algorithm of random forest is to randomly sample from data set and obtain the results through multiple decision trees. And using the average value of the processing results to solve the regression problem (the classification problem usually uses the largest number of classification results as output). In practical, random forest greatly improves the accuracy without significantly increasing the amount of computation. In addition, the random forest model is more accurate than other models when dealing with incomplete data and unbalanced data due to its insensitivity to multi-collinearity [8].

2.4. Training Random Forest Models

For the training of random forest, the fundamental purpose is to find the best parameters of `n_estimators`, `criterion`, `max_depth` and `max_features`, so that the data fitting degree of the model becomes higher. The `n_estimators` is the number of trees in a random forest. Criterion refers to the measure of whether it is the best nodal, which supports three functions: 'mean square error', 'mean absolute error' and 'Poisson deviation'. `max_depth` refers to the maximum depth of the tree, and `max_features` refers to the number of features to consider which can find the optimal division. After multiple turns of training, these optimal parameters were returned and outputted as shown in Table 1.

Table 1. Optimal parameters of the mode used in this study

Parameter	output
criterion	absolute_error
max_features	log2
n_estimators	50
max_depth	991

3. Result and discussion

The following figure shows the population size and fertility rate of China 's mainland as predicted by machine learning models. From Figure 2, it can be seen that since 1990, the percentage yearly change of China 's population has been slowing down and will reach a maximum of 1.44 billion around 2030, and followed by negative population growth. It can also be seen from the population change rate in Figure 3 that the growth rate is 0 around 2030 and then becomes a negative growth.

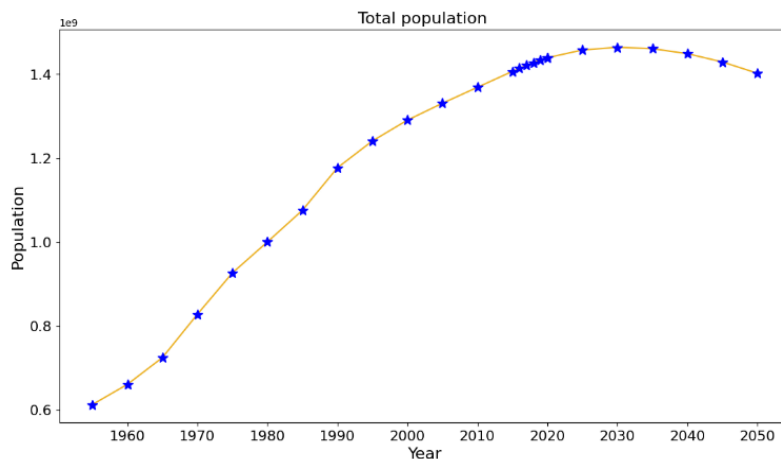


Fig. 2 Total population

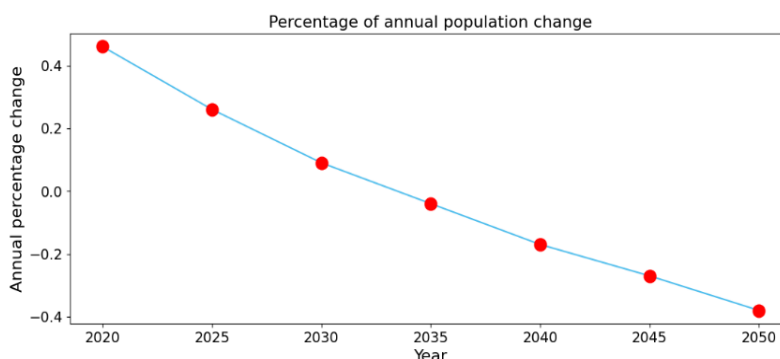


Fig. 3 the rate of change of population

Logistic regression model has important application value for population, resource allocation, environmental capacity and so on. This model is based on the population of China from 1955 to 2022, and predicts the population to 2050. From the perspective of calculation, Logistic regression runs fast and occupies less computing resources. In practical use, the application of Logistic regression is suitable for most of the data with growth retardation. Because of the simple structure, it is relatively simple to use and has strong interpretability, which makes it use in a lot of areas [7].

The population prediction model based on random forest algorithm has higher regression accuracy and fit degree. It has a flexible and stable framework that allows different types of data to interact during the modeling process. Compared with other traditional regression models, the random forest model not only achieve accurate prediction, but also avoid overfitting [9]. In addition, random forest is composed of multiple decision trees. Compared with single classifiers such as decision trees, it has better prediction performance [10]. For the further improvement of the performance, more advanced algorithms e.g. neural network may be considered in the future due to their excellent performance in many tasks [11, 12].

4. Conclusion

Machine learning is a kind of research method to improve the accuracy of the model by analyzing the specified data and optimizing the calculation method continuously through computer technology. In this work, the prediction accuracy was improved by using hyperparameters to optimize logistic regression model and random forest model. By comparing with the existing data, it is found that the trained models are more accurate in predicting population. However, there are many factors affecting population growth, and most of these factors have complex nonlinear relationships, it is necessary to learn about the application of nonlinear regression models in population prediction in the future, such as neural networks, grey prediction and so on. It can be applied to the state that the data is incomplete, or the influencing factors cannot be determined, which is helpful to improve the prediction accuracy.

References

- [1] Wang X et al. 2007 A Review of China 's Population Prediction Research in Recent Years Theory and Reform (06) pp 157-160
- [2] Li J et al. 2018 China's population policy and population structure and its future development trend Economic dynamics (12) p 20
- [3] Hong Y et al. 2021 Big Data, Machine Learning and Statistics: Challenges and Opportunities Journal of Econometrics 1 (1) pp 17-35
- [4] Ho T 1995 Random decision forests in Proceedings of 3rd international conference on document analysis and recognition (Vol. 1 pp 278-282) IEEE
- [5] Hang Q et al. 2018 Application status of machine learning random forest algorithm electronic technology and software engineering (24) p 3
- [6] Kaggle 2022 Population of China <https://www.kaggle.com/datasets/amritharj/population-of-china-19502022>
- [7] Zhang X et al. 2013 Population Forecast Analysis Based on Logistic Regression Model Journal of Chuxiong Normal University (0) pp 9-13
- [8] Li X 2013 Application of random forest model in classification and regression analysis Applied Entomology 50 (4) p 8
- [9] Gao X et al. 2022 Spatial Simulation of Population in Bohai Rim Based on Random Forest Model Journal of Earth Information Science 24 (6) p 13
- [10] Zheng R et al. 2021 Spatial differences and influencing factors of social integration of floating population in China based on random forest model Geographical Sciences 41 (10) p 10
- [11] Yu Q et al. 2020 Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV China Communications 17(3) 46-57
- [12] Han Z et al. 2022 ConvUNeXt: an efficient convolution neural network for medical image segmentation Knowledge-Based Systems 253 109512