

Comparison Of 6 Machine Learning Models in Estimating Population Growth

Yitong Huang *

Shenzhen College of International Education, Shenzhen, Guangdong, 518000, China

* Corresponding Author Email: s22035.huang@stu.scie.com.cn

Abstract. With the rapidly increasing population globally, it is essential for policymakers to be able to accurately predict or gain an idea of the forecast of population growth to be able to make effective regulations that can benefit the general public. Therefore, the development of a machine learning model to estimate future population growth is crucial. In this article, various machine learning models such as linear regression, logistic regression, decision trees, random forest, neural networks, and support vector machines are discussed and the benefits and downsides of each are considered. Factors impacting population growth are also discussed to conclude the qualities needed for a model to most suitably perform the task of population prediction. In the end, it is shown that random forest is the best model for this job as it can give a generalized pattern for its results as well as handle complex data types. This paper provides predictions and insights based on machine learning to predict future demographic trends, which can provide useful information for policymakers, researchers, and society in various fields.

Keywords: Machine learning; population growth; population estimation; random forest.

1. Introduction

The world's population is currently experiencing a dynamic shift. The global population is growing at an unprecedented rate, expected to rise from 4.1 billion in 1900 to 10.2 billion in 2100 [1]. This unstable population growth is having a great impact on various parts of our lives. Debates over population policy making continue on, with some seeing the rising standards of living due to rapid globalization and growth, and others arguing for the widespread poverty and environmental stress caused by overpopulation [1]. Therefore, analysis of the factors that influence population growth and the ability to accurately predict these changes has become increasingly important. Firstly, improvements in healthcare and discoveries in medicine have led to a decrease in mortality rates, resulting in longer life expectancy. Economic development and improved living standards in some developed countries have led to a decline in birth rates in many countries. On the other hand, some countries are experiencing rapid population growth. Unstable population growth can generate many issues, one of which is the strain it puts on resources and infrastructure. According to the Malthusian theory, As the population continues to grow, the demand for food, water, and energy increases, which can lead to scarcity and conflict [2]. Furthermore, the aging population in many countries sets up significant obstacles to pensions and social security systems. Thus, it can be seen that it is critical to accurately predict population changes to effectively plan and allocate resources to meet these demands.

Regarding this topic, the advancements in machine learning have made population prediction much easier. Machine learning algorithms are built to analyze vast amounts of data often present in the case of populations, and identify patterns and trends that humans may overlook. This enables more accurate forecasting of population changes and helps policymakers make better decisions. There is a myriad of ways to apply machine learning models to population forecasting and city planning. For instance, it can assist in urban planning as it can recognize patterns, carry out predictions, make decisions, and perform operations with speed and accuracy [3]. It can also aid in healthcare planning by forecasting the demand for medical services and identifying areas that require additional infrastructure. Moreover, machine learning can contribute to policy development in areas such as

education, transportation, and environmental sustainability, by anticipating population shifts and their associated needs.

This article aims to compare and contrast a range of machine learning models to deduce which is the most suitable model for estimating the population. In section 2, the respective advantages and disadvantages of the models are discussed, with the functions and common applications listed. Section 3 explores some factors affecting population growth, and the qualities that a model should have to most accurately and effectively make predictions.

2. Model

2.1. Description of Various Models

Linear regression is made for finding the best-fitting line that reduces the discrepancy between the predicted values and the actual target values. This form of analysis estimates the coefficients of the linear equation [4]. It presupposes that the input features and the target variable have a linear relationship. To provide predictions, the model calculates a best fit line that runs through all the actual data [4].

Logistic regression is a type of model used for binary classification. Using a sigmoid function, it calculates the possibility that an input belongs to a group and translates the result into a range of 0 to 1. In such binary classification tasks, one class could be the normal state, while the abnormal state could be another class [5].

Based on the values of the input features, a decision tree assigns an input to a certain class or value by making a sequence of binary judgments at each node. It is modeled like an inverted tree structure consisting of decision sections composed of points, branches, and leaf nodes. Starting at the root node of the tree, and then moving down the tree branch, a decision tree checks the defined attributes at each node to sort the data [5]. Splitting in decision trees frequently uses standards like Gini impurity or entropy [6].

A random forest is an ensemble model made up of several decision trees. The main purpose of an ensemble method is to integrate the projections of several weaker estimators that are singly trained in order to boost up or enhance generalizability or robustness over a single estimator [7]. A portion of the data is used to train each tree. The final prediction is formed by accumulating the predictions from each individual tree, which reduces overfitting, increases generalization, and increases prediction accuracy.

Neural networks are a mathematical model that simulates biological neurons in the brain for information processing [6]. It is made up of layers of wired-together neurons that process information. Each neuron applies a weighted sum of its inputs, and then transmits the output to the following layer after passing the result through an activation function. Deep neural networks, which have numerous hidden layers between input and output layers, may learn intricate hierarchical features from the data via a method known as backpropagation to adjust the weight values internally while building the model [5].

Support vector models (SVMs) are supervised two-classifiers based on the Vapnik–Chervonenkis dimension theory in statistics, and the principle of structural risk minimization [6]. They aim to find a hyperplane that maximizes the margin between two classes in the feature space. The data points closest to the hyperplane are called support vectors. When handling nonlinear boundaries, SVMs use kernel functions to map the sample data to a high-dimensional space and seek optimal categorical hyperclass planes to isolate different categories of sample data for classification [6].

2.2. Advantages and Disadvantages of Each Model

One of the key advantages of linear regression is that it is simple to understand and interpret. It doesn't take a long time to train with the data. Additionally, linear regression is computationally efficient and can handle large datasets. However, its major drawback is that it assumes a linear

relationship, which may not always be the case in real-world scenarios. It works best when representing direct relationships between the feature and the target.

The advantages of logistic regression, on the other hand, lie in its efficiency for binary classification. Logistic regression provides the probability of a certain outcome, allowing for better sorting. However, it is also limited by the way it always assumes a linear relationship between the independent variables and the log odds of the dependent variable. This assumption may lead to inaccurate predictions when dealing with nonlinear relationships. It is best suited to cases where the classes are well defined and linearly separable.

Decision trees can be easily interpreted and visualized, as the model's decision-making process can be visualized in a tree-like structure. They can handle both numerical and categorical data and can capture non-linear relationships. However, decision trees are prone to overfitting, especially when the tree becomes too complex. On top of that, they can be very sensitive to small variations in the data, and lack memory when dealing with large databases [8]. This can lead to poor generalization and decreased performance on unseen data. More recently developed models are excelling in user behavior analytics and cybersecurity analytics [5].

By combining different decision trees, random forests alleviate the overfitting problem with decision trees. The predictions from various trees are combined in this ensemble learning technique to get predictions that are more accurate. Due to the ensemble nature of the model, they can be computationally demanding and may need extra RAM. They take longer to train on data and produce predictions.

Neural networks are incredibly adaptable and are capable of handling intricate patterns and interactions. They may be used for both classification and regression problems and have the ability to learn from massive volumes of data. In shallow neural networks, only simple computational learning is needed, leading to significantly short training time [8]. However, they can lead to many redundancy problems as they have a myriad of descriptors that are all correlated [8]. Getting stuck in local minima is also a major problem for shallow neural networks [8]. Deep learning neural networks are capable of capturing complex features of data values [8]. However, they are very time consuming and need a lot of data and computer power.

SVMs do a great job of managing high-dimensional data and are good at identifying complex correlations. By using kernel functions, they can handle both linear and nonlinear interactions. The theoretical underpinning of SVMs is also solid, offering a clear knowledge of the decision bounds [8]. However, SVMs may require careful hyperparameter adjustment and be challenging to comprehend. They are also very computationally expensive, as they require large matrix operations [8].

3. Population Growth Prediction

3.1. Discussion of Factors Affecting Population Growth

Population growth is affected by a variety of social, economic, environmental, and demographic factors. Some of the statistics include:

The birth rate, which is the number of births per 1,000 people in a year, is calculated as a percentage. A higher birth rate leads to larger increases in population. On the other hand, the death rate is the number of deaths for every 1,000 people in a given year. Population increase may result from lower mortality rates. The average number of children a woman births in her lifetime is known as the total fertility rate [9]. Increased fertility rates can cause population growth to occur more quickly. The life expectancy refers to the typical number of years that a person can expect to live. Increased life expectancy results in a greater percentage of elderly people and a lower death rate, contributing to population growth.

The immigration and emigration policies and numbers can significantly affect population growth. The amount of people moving in and out of a region, and the relative ease for them to do so can either lead to population growth or slow it down.

Levels of economic development in a country can also be very impactful on its population trends. Residents in a country currently under development will have more incentives to have more children, while inhabitants of a well-developed country may not want as many children. A booming economy will also attract migrant workers in search of possibilities, while having no job opportunities reduces the number of people willing to immigrate. Education and Literacy also come into play, as more developed countries tend to have higher levels of education and literacy in their population, particularly for women, and are associated with lower fertility rates and better family planning practices, which can influence population growth. The process of population growth is exogenous to the processes of income generation, accumulation, technical progress, and institutional change [10].

Improved healthcare infrastructure, access to medical services, and advancements in medical technology can lead to lower death rates and increased life expectancy, contributing to population growth. Environmental conditions, such as access to clean water, sanitation, and healthcare, can impact birth and death rates, thus affecting population growth.

Another aspect that may affect attitudes toward family size and contraception is the country's cultural norms and religious beliefs. Fertility rates can also be impacted by traditional gender roles and expectations. Population increase may be impacted by government regulations on family planning, parental leave, immigration, and social assistance.

3.2. Requirements for A Prediction Model

A successful and useful model for predicting population growth must be able to accommodate and consider the various factors affecting population growth as much as possible, which is a challenging task that requires a combination of data sources, feature engineering, and model architecture. The model should be able to integrate data from various sources, such as the prementioned demographic statistics, economic indicators and more. This requires data preprocessing and cleaning to ensure consistency and accuracy. Designing relevant features that capture the essence of each factor is crucial. For instance, creating features for birth rates, death rates, fertility rates, education levels, healthcare access, etc., can help the model understand and analyze the different factors. The model should also be capable of performing multivariate analysis to account for the interactions and correlations between factors. This might involve techniques like regression analysis, principal component analysis (PCA), or more advanced methods. It should be noted that most factors change over time, so the model should be able to handle time-series data to understand trends and patterns in population growth. Population growth can vary significantly across regions. Thus the model needs to have some spatial analysis techniques, such as geographic information systems (GIS), which can help the model consider geographical variations. Given that population growth is influenced by a wide range of fields, such as sociology, economics, and geography, an interdisciplinary approach is needed to ensure the model accounts for all relevant factors. The model should be capable of accepting new and changing information because when new data becomes available, it should be able to continuously update itself to improve predictions and adapt to changing trends.

3.3. The Advantages of Using Random Forest for Population Growth Prediction

It can be seen that random forest would be the best model for predicting population growth. This is because it is well suited to deal with complicated, nonlinear, and large datasets, such as those that are typically present in the case of population prediction. There will be many factors present in the estimation, such as economic indicators like unemployment rate, inflation rate and interest rates; social factors, including access to healthcare, gender equality, as well as culture and traditions; and historical data. Random forests can handle a diverse range of 9 variables. Random forests are also very versatile, making it easy to incorporate and add features that enhance its performance in making a forecast of population growth. They can also work around missing data points and anomalies, which are common obstacles when it comes to predicting population growth. Random forests are able to use the other variables in the dataset as a base to fill in information for the outliers. Furthermore, random forest has built-in mechanisms to deal with overfitting. By combining multiple decision trees, it trains

each tree on a different subset of the data and then aggregates their predictions. This ensemble approach helps to reduce variance and improve generalization, making it suitable for estimating populations on unseen data. Random forests also give a simple interpretation, making it a useful tool to policymakers as it can identify the pivotal factors in its estimation.

4. Conclusion

In this article, the best model for estimating population growth is explored, by first discussing the relative advantages and disadvantages of each model. Linear regressions are simple to interpret and train, but can't represent complicated data, making it unsuitable for population prediction. Logistic regression is applied mostly for binary classification, thus also making it unsuitable. Neural networks have the ability to capture complex features in data and can learn from a very large volume of data. However, it can lead to many redundancy problems, get stuck in local minima, and require large computational costs, making it unsustainable. SVMs can do a great job of working with high dimensional data, and have a solid theoretical basis, but the results obtained can be extremely hard to analyze and making it not much help in making policies. While decision trees can handle complex and diverse data types well, it is too prone to overfitting when faced with too many nodes, or small variations. The random forest model solves the problems of overfitting and sensitiveness of the decision tree model as it is an aggregate of many trees, making its results much more general and reliable, therefore leading it to be the best choice in population estimation. Some important factors affecting the population were also mentioned and analyzed, ranging from birth rates and death rates to the level of economic development and access to healthcare and sanitation. These factors include fertility rate, mortality rate, life expectancy, immigration laws and rates, cultural factors, and environmental conditions. All of the above can have tremendous impacts on population growth, thus it is important the prediction model is able to accommodate the factors and take them into consideration. The model should also be capable of accepting many new and rapidly changing data, because population growth doesn't always remain stable and factors are forever changing.

References

- [1] Bongaarts, John. Population policy options in the developing world. Policy Research Division Working Paper no. 59. New York: Population Council, 1994. Version of record: <https://www.science.org/doi/10.1126/science.8303293>.
- [2] Malthus, Thomas Robert. An Essay on the Principle of Population. History of Economic Thought Books, McMaster University Archive for the History of Economic Thought, 1798.
- [3] J Chaturvedi, V.; de Vries, W.T. Machine Learning Algorithms for Urban Land Use Planning: A Review. Urban Sci. 2021, 5, 68.
- [4] About Linear Regression | IBM, <https://www.ibm.com/cn-zh>.
- [5] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI, 2021, 2, 160.
- [6] YANG Jianfeng, QIAO Peirui, LI Yongmei, et al. Statistics and Decision, 2019, 35(06):36-40.
- [7] Das, Kajaree, and Rabi Narayan Behera. "A survey on machine learning: concept, algorithms and applications." International Journal of Innovative Research in Computer and Communication Engineering 5.2 (2017): 1301-1309.
- [8] Fahimeh Ghasemi, Alireza Mehridehnavi, Alfonso Pérez-Garrido, Horacio Pérez-Sánchez. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks, Drug Discovery Today, 2018, 3(10):1784-1790.
- [9] Lutz, Wolfgang, and Ren Qiang. "Determinants of human population growth." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2002, 357(1425): 1197-1210.
- [10] Srinivasan, T. N., "Population Growth and Economic Development" (1987). Discussion Papers. 547. <https://elischolar.library.yale.edu/egcenter-discussion-paper-series/547>