Prediction and Influencing Factors of Residents' Ideal Childbirth: Feature Selection Based on Random Forest

Liu Yu

School of Sociology and Political Science, Shanghai University, shanghai, China
liuy505@163.com

Abstract

As the number of new births in China continued to decline from 2017 to 2021, the focus on fertility is particularly important. This paper summarizes four categories of predictor variables from the literature, namely basic personal characteristics, residents' original family characteristics, occupational characteristics and couple relationship characteristics, a total of 16 variables, and screened out the more important 12 variables based on random forest feature selection. The study found that: (1) The variables among the occupational characteristics, original family characteristics and personal basic characteristics of residents have a great influence on the ideal number of children. (2) In the prediction analysis, the support vector machine with linear kernel function has the best prediction effect, and has obvious advantages over logistic regression and random forest. The results of the study are of significance for understanding China's fertility level and alleviating the decline of the newborn population.

Keywords: *Ideal Fertility, Data Mining, Random Rorest, Feature Selection*

1.INTRODUCTION

In 2000, China's elderly population aged 65 and above accounted for about 7% of the total population. According to the standard of The United Nations, China has entered an aging society. According to Statistical Communiqué of China on the 2018 National Economic and Social Development, as of the end of 2017, the elderly population over the age of 60 accounted for 17.3% of the total population, of which those over the age of 65 accounted for 11.4% of the total population. In order to alleviate the trend of aging, it is necessary to improve the fertility rate of Chinese population. However, the number of new births in China has decreased for three consecutive years, and Chinese new birth population may experience a "cliff-like" decline in the future. Therefore, the research on residents' willingness to bear children is particularly important. In the future, the level of urbanization in my country will be further improved, so the fertility of urban residents determines the number of new born population in the country. This study takes Shanghai as an example to study the factors that affect residents' willingness to bear children. In the fields of demography and sociology, there have been many studies on fertility, but most of these studies have used traditional methods to conclude from theories that a certain factor has an impact on fertility willingness, and then use data to verify it. The conclusions and models used in this method have a greater relationship with the sample, so they are highly subjective. In this paper, we will use the random forest method of predicting the importance of variables to examine the influencing factors of fertility.

There is a direct relationship between residents' willingness to have children and their ideal number of children. Therefore, this paper will use the method of data mining to study the influence of varieties of factors on the number of residents' ideal number of children.

2.LITERATURE REVIEW

2.1. Research on the Influencing Factors of Fertility Willingness

There are many literatures on the factors affecting fertility rate, which can be summarized into the following categories: family status theory, economic influence theory, and career development theory.

First, with the increase of women's employment opportunities and the improvement of independent decision-making ability, women's willingness to do business will remain at a low level [1]. There are also studies that believe that women, as the main body of

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 704–709, 2023.
https://doi.org/10.2991/978-94-6463-042-8_101

fertility, are more eager for the right to choose their own reproductive behaviors when their own economic strength is rising [2]. Second, economic impact theory. This kind of research believes that economic factors can affect whether couples choose to have children, and the income level of Chinese married women affects their willingness to have children [3][4]. Third, career development theory. This type of research mainly examines fertility intentions from the perspective of couples' career development. Having a second child makes professional women's career and family life conflict with each other, and employment discrimination in the workplace exacerbates the conflict between the two. The higher the occupational status, the lower the willingness to bear children [5]. The above studies on the factors affecting fertility are insufficient. This paper takes into account the above factors that may affect residents' ideal childbearing, and uses machine learning methods to examine the factors that affect residents' ideal fertility willingness.

2.2. The Use of Data Mining Methods in Fertility Research

The methods of data mining are generally used in classification, while the methods used in population research are generally traditional statistical analysis methods. When it comes to classification, most of them use the method of logistic regression. There are also a small number of researchers who have used other data mining methods in demography. The researchers found that the random forest model performed better than the Logit regression model in multiple classification evaluation criteria [6]. Compared with the original model, the results of using the support vector machine model in China's population prediction have significantly improved the prediction rate and accuracy rate [7]; Random forests have also been used to study the relationship between population characteristics and urban crime rates [8]. Some scholars have used the data mining technology of decision tree to analyze the population information system in e-government affairs [9]; the use of data mining methods can explore the current situation and development trend of urban population aging [10]. To sum up, the data mining method can be used in population research, and it is possible to achieve good analysis results.

3.DATA AND VARIABLE INTERPRETATION

The data used in this paper is The Shanghai Urban Neighborhood Survey(SUNS) data, which is a large-scale sample survey completed in 2017. The data has undergone strict sampling design, and the data quality has certain reliability. Of the large sample surveys completed, we only selected the adult questionnaires (aged 16+). The data was filtered, and because the study focused on couples, residents who were already married were selected. This article wants to study the influence of occupational factors on the ideal number of children, so the questionnaires of residents who do not have jobs are excluded. In order to make the research meaningful, the age range of the respondents should be fertile. The questionnaire says that the age is limited to under 46 years old. Therefore, this paper will delete the questionnaires for respondents who are older than or equal to 46 years old, and finally get 822 valid questionnaires.

1. The response variable of this paper is the ideal number of children of residents. This variable is processed in this paper. In the SUNS questionnaire, there is an item "How many children do you plan to have in the future". The answer to this question is "less than 2" and assigned a value of 0, and an answer greater than or equal to 2 is assigned a value of 1. The variable becomes a binary response variable.

2. Predictor variables. According to the above literature review part, it can be known that the factors that can affect the ideal fertility willingness are divided into four categories: basic personal characteristics, characteristics of residents' original family, occupational characteristics and characteristics of couple relationship. The 16 explanatory variables and codes of these four types of characteristics are shown in the table below.

Table 1: Related explanatory variables table

variable category	variable number	variable name
basic personal characteristics	1	gender
	2	age
	3	education
	4	nation
	5	register
	6	political status
family of origin characteristics	7	edu-father
	8	edu-mother
	9	father occupation
	10	number of siblings
occupational characteristics	11	category of job
	12	time of job
	13	income
	14	job satisfaction
	15	wife decision-making

marital relationship characteristics	16	couple relationship

4.FEATURE SELECTION BASED ON RANDOM FOREST

Random forest is a kind of combinatorial classification method, which is composed of a large number of decision trees, so it is called forest. The training data for generating a single tree is generated by independent sampling, and the candidate split attributes of each internal node in the single tree are randomly selected from all the candidate attributes. The final classification result of the random forest is determined by voting of each decision tree.

This paper uses the 16 explanatory variables in Table 1 to conduct random forest modeling for the response variable residents' ideal fertility children, and the default random forest is set to 500 trees. After the model is established, the variable importance of the model is extracted, and the variable importance obtained based on the accuracy reduction index and the Gini index reduction value is shown in Figure 1 and Figure 2. From Figure 1, we can see that the household registration, father's occupation, whether he is a party member, the number of siblings and the number of hours worked per week play an important role in accurate classification, while the mother's education level, wife's decision-making power, marital status, and ethnicity did not contribute to the correct classification. From the decreasing value of Gini index in Figure 2, we can see that age, occupational income, number of siblings, years of education, working hours per week, and father's occupation play an important role in correctly classifying more or less ideal children as relative dependent variables. while variables such as ethnicity, wife's decision-making power, and gender have little effect on the correct classification.

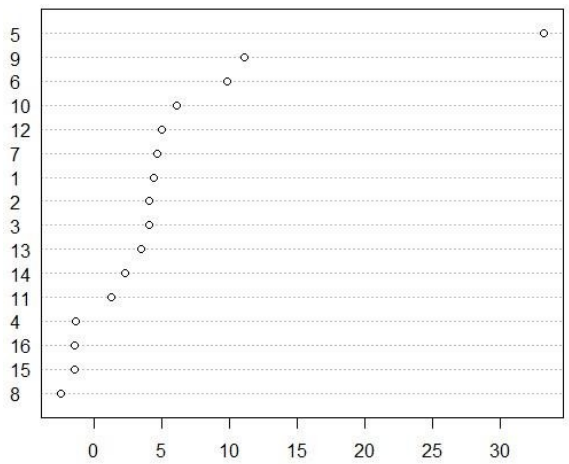

Figure 1: Variable importance plot according to reduction in accuracy.

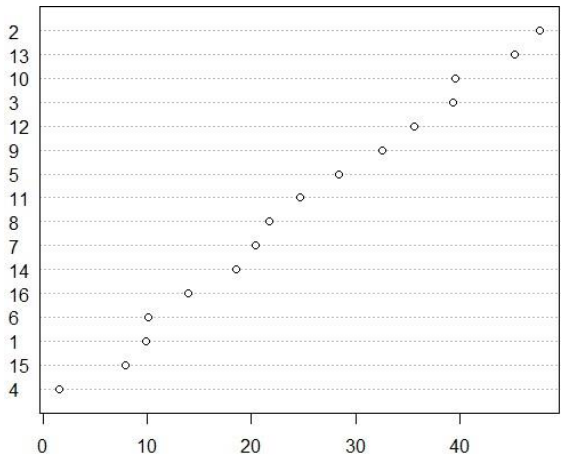

Figure 2: Variable importance plot according to reduction in Gini index

According to the importance map of two variables, the author established the following sets of explanatory variables. Let the complete set of 16 variables be S, S1=S-{8,16,15,4}, S2=S-{1,4,6,15}, S3=S-{1,8,4,15},S4=S-{1,4,6,15,16},S5=S-{1,4,6,15,16,14}, S6=S-{1,4,8,15,16}. That is, after excluding the less important variables, a total of six explanatory variable sets are set, and the error rates of the OOB estimates for these six variable sets are shown in the following table.

Table 2: Out-of-bag error rate table for explanatory variable set

explanatory sets	variables	OOB Error rate(%)
S1		36.01
S2		36.64
S3		35.52
S4		36.13
S5		36.62
S6		34.55

It can be seen from the above table that the out-of-bag estimation of the explanatory variable set S6 is the best, so the explanatory variable set is finally determined to be S6, that is, age, number of siblings, years of personal education, household registration, occupation type, and weekly working hours, job satisfaction, father's education, father's occupation, whether he is a party member, and occupational income are 12 variables to predict the ideal fertility of residents.

5.PREDICTION OF IDEAL NUMBER OF CHILDREN

5.1. Model Evaluation Metrics

The model selection metrics used in this paper are accuracy, precision and recall. These three indicators are based on the positive and negative predictive values. The

meanings of true positives, true negatives, false positives and false negatives are shown in Table 3 below. True positives are predicted to be true and are actually true; true negatives are predicted to be false but are actually false; false positives are predicted to be true but are actually false; false negatives are predicted to be false but are actually false truth value.

Table 3 Schematic representation of model predicting negative and positive

actual value	predictive value 1	0	total
1	TP (True Positive)	FN (False Negative)	P
0	FP (False Positive)	TN (True Negative)	N

1. Accuracy: Accuracy refers to the proportion of correctly classified positive and negative tuples in the total tuples, which measures the overall recognition of the prediction model. The calculation formula is as follows:

$$Accuray = \frac{TP + TN}{p + n} \quad (1)$$

2. Precision: Precision refers to the proportion of tuples that are predicted to be positive tuples that are actually positive tuples. It measures the precision of the prediction model. The calculation formula is as follows:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. Recall: Recall refers to the proportion of tuples that are actually positive tuples that are predicted to be positive tuples. It measures the completeness of the prediction model. The calculation formula is as follows:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

5.2. Prediction

In this section, three models will be used to predict the ideal childbearing situation of residents using the S6 explanatory variable set. The methods to be used are logistic regression, support vector machine and random forest, and then the best model is selected based on the data mining model selection indicators.

1. Logistic regression: Considering that residents' ideal childbearing is a binary response variable, this study first considers the use of a logit model to study the relationship between occupational conditions and willingness to have children again.The mathematical expression of the regression model is:

$$logit(Y) = ln(p_i) = \alpha + \beta X + \varepsilon_i \quad (4)$$

The data set is divided into training set and test set, in which the number of samples in the training set accounts for about 70% of the total, and the test set accounts for about 30% of the total. After using the explanatory variables in S6 to build a logit model for the ideal childbearing children of the residents in the training set, predict on the test set with an accuracy rate of 64%, a precision of 66.67%, and a recall rate of 64.15%.

2. Support vector machine: Support vector machine is generally modeled by a combination of classifiers and kernel functions. Its basic model is a linear classifier with the largest margin defined in the feature space. Support vector machine generally has four kernel functions, namely linear kernel function, polynomial kernel function, radial kernel function and sigmoid kernel function. The four kernel functions and their parameters were tried respectively, and the calculation results are shown in the following table:

Table 4: Prediction results of SVM models with different kernel functions

kernel function	accuracy (%)	precision (%)	recall (%)
linear	77.1	78.6	78.2
polynomial	69.1	71.6	62.2
radial	66.6	66.7	64.2
sigmoid	67.4	61.8	64.3

From the above table, the Support vector machine with linear kernel function has the best performance on the test set, the accuracy is 77.1%, and the precision can reach about 79%, which is very rare. In social sciences, the value of a variable is often determined by a variety of factors, and it is difficult to predict accurately. In this paper, the linear kernel function can reach more than 75%, indicating that the model is well established and the explanatory variables are more appropriate.

3. Random Forest: We know that the prediction result of random forest is determined by two important factors, they are: (1) the number of pre-selected variables in each tree node, (2) the number of trees in the forest. We tested the number of trees in the forest with the lowest OOB error rate when the number of preselected variables was 2, 3, 4 and 5. The experimental analysis shows that when the number of pre-selected variables for each tree node is 2, 3, 4, and 5, the error is the smallest when the number of trees in the forest is about 160, 260, 220, and 120, and then the error tends to be stable. We find that the overall prediction effect of random forest is not so good, and the highest prediction accuracy is about 65%.

Comparison of models: based on the prediction results of the variables in the previous part, we calculated the following table6. From the table, we can see that the support vector machine performs the best in the prediction of the test set, and its accuracy, precision and recall are both above 75%. In summary, prediction result of the SVM with linear kernel function is the best.

Table 5: The performance of the three types of models on various evaluation indicators

models	Accuracy (%)	Precision (%)	Recall (%)
logistic regression	64	66.67	64.15
best support vector machine	77.1	78.6	78.2
best random forest	65.21	65.05	65.37

6.CONCLUSIONS

This paper studies the factors that affect the ideal number of children born by residents through the method of data mining. The study found that: (1) Variables such as household registration, number of siblings, and age of residents' occupational characteristics, original family characteristics and basic personal characteristics have a greater impact on the number of children ideal for residents; The effect of the variable on the ideal number of children born by residents is small. (2) In the process of predicting the ideal number of children to be born, the SVM with linear kernel function has better prediction accuracy and precision; the prediction effect of logistic regression and random forest is relatively poor.

Policy Suggestions: In order to alleviate the declining trend of China's new born population, this study proposes the following suggestions. In the four categories of characteristics that affect fertility, some are congenital and some can be changed. Occupational characteristics and the relationship between husband and wife can affect the ideal number of children for residents to have, so the government should also limit the weekly working hours of laborers, raise the minimum wage standard, and increase the income of residents; create a good social atmosphere to make the relationship between couples more harmonious, so that China can have conditions for raising its fertility rate.

This article has shortcomings. Because the data are from a sample survey in Shanghai, China, the results are not representative of cities across the country. However, the results of this study are still meaningful. First, based on the research results, we know that factors such as household registration, characteristics of the original family and occupation play an important role in improving fertility. Therefore, in order to alleviate the situation of declining fertility in China, measures such as appropriately loosening restrictions on household registration and increasing labor remuneration can be adopted. Second, the data mining method can be well applied in the study of the population field, and the support vector machine has a good effect in this study.

REFERENCES

[1] Zheng Zhenzhen (2004). Fertility Desire of Married Women in China. J. Chinese Journal of Population Science. 05, 75-82.

[2] Zhao Menghan. & Ji Yingchun. (2019). Husbands' Housework Share and Women's Hazards of Entering Parenthood. J. Population Research. 01, 64-77.

[3] Liu mina (2010). An Analysis of Affecting Factors of the Desired Childbearing Number of married Women inChina. J. Northwest Population Journal.01.

[4] Qing Shisong. & Ding Jinhong. (2015). Couple's Fertility Intention Difference across Only Child Families: Evidence from Matched-couple Survey in Shanghai. J. Chinese Journal of Population Science. 05,81-93+128.

[5] Yang fang. & Guo xiaomin (2017). Research on the Influence of "Universal Second Child" on Professional Women and Policy Support—Based on the Perspective of Work-Family Balance. J. China Youth Study. 10,31-36+22.

[6] Li Dongling (2017). Prediction to the Second Childbearing Desire of Fertile Woman Based on Data Mining. J. Software .08.

[7] Li Feiya. & jiang ruofan (2012). Population Prediction Application Based on PCA and SVM Model. J. Northwest Population Journal. 01, 29-32.

[8] Wang Yuchen. & Guo Zhongyang. & Wang Yuanyuan (2017). A forecasting model of crime risk based on random forest. J. Journal of East China Normal University (Natural Science).4.

[9] Meng Sheng (2013). The Application of Data Mining in Analysis of the Census Data. D. Xiamen University.

[10] Li Miao (2016). Harm, trend and countermeasures of population aging in Chongqing. J. Rural Economy and Science-Technology. 10.