# Determinants of Childbearing Conceptions in Married Women: A Machine Learning Comparative Study

Li Zhu[1], Qian Long[2*]

[1]Associate Professor, School of Big Data Statistics, Guizhou University of Finance and Economics
[*2]Master's student in the School of Big Data Statistics, Guizhou University of Finance and Economics
*Corresponding Author

Received: 05 March 2025/ Revised: 14 March 2025/ Accepted: 20 March 2025/ Published: 31-03-2025


***Abstract—*** *In recent years, with the continuous progress of society and the continuous change of people's way of thinking, the inclusiveness of society towards women, especially married women, is increasing, and married women have gained more rights and choices, but the fertility rate in China is declining. As a result, the study of demographics and family structure has been gaining traction, especially on the decline in the natural growth rate of the population shown in the national census data, which is also related to the overall characteristics of women, such as whether they are employed or not. This paper uses multiple regression model, decision tree, random forest, gradient boosted regression tree model, support vector machine, XGboost and other machine learning models to train and compare the data, and draws the following conclusions: (1) The comprehensive characteristics of married women such as employment, age, and mother's education level are negatively correlated with their willingness to have children; (2) The influence of married women's husband's husbands and family characteristics on the intention to have children was positively significant. (3) Based on the results of multiple regression models, it warns us that we should consider multiple aspects when doing research or model selection, and that the selected model is good but not necessarily the most consistent with the characteristics of the data; (4) Based on the comparison of various models, it is found that the gradient boosted regression tree model, K-nearest neighbor model, XGBoost model and random forest model are better than the linear regression model.*

***Keywords— Comprehensive Characteristics of Married Women; Fertility Concept; Machine Learning.***

## I. INTRODUCTION

In recent years, with the continuous progress of society and the profound transformation of people's thinking patterns, the tolerance and respect for women, especially married women, have been increasing. This change has given married women more choices and rights, however, China's fertility rate has shown a downward trend in this context, and changes in demographics and family structure have attracted more and more attention from all walks of life. In particular, the declining natural population growth rate revealed by the national census data is closely related to individual characteristics such as women's employment status, which is reflected in the fact that married women begin to consider whether to have children. Among them, the employment status, age, education level, and the relationship between various personal and family characteristics of married women are of great significance for understanding the dynamics of social fertility, promoting women's career development, and optimizing family policy formulation. In this context, it is crucial to use modern machine learning techniques to dig deeper into this topic. In addition, this study not only helps us to comprehensively examine and improve the social reproductive environment from the unique perspective of married women, but also has great significance for the exploration of women's career development, the change of family structure, the optimization of social labor structure, and the improvement of the overall level of China's economic and social development.

In order to comprehensively and deeply explore this problem, this paper constructs a dataset covering a variety of variables based on abundant social survey data resources, and on this basis, it is proposed to use a variety of machine learning models for simulation and comparison, including but not limited to multiple regression, decision trees, random forests, XGboost and logistic regression, and support vector machines. The optimal prediction model is found by comparing the accuracy metrics of each model. Based on the training and comparison of these machine learning models, it can not only provide a comprehensive

analysis of the influencing factors of married women's reproductive age, but also reveal the performance differences and applicability of different models in dealing with such complex problems, so as to provide a solid scientific basis and reference for the formulation and practical operation of relevant policies.

This paper combs and summarizes the existing literature, and finds that many scholars at home and abroad have a lot of relevant research on the relationship between the characteristics of married women and whether they have the youngest child, and the research scope is also very broad. Among them, Huang Qian (2022) et al. believe that when married women have the youngest child, their employment status is related to the age of the child. The presence of the youngest children of different ages has different effects on women's employment. Based on the comprehensive evaluation and analysis of CGSS survey data, Song et al. (2022) concluded that family characteristics such as parents' education level, spouse's occupation, and family economic status have a significant impact on the employment quality of married women. Yan Yu (2022) argues that the youngest child is not significant when married women are employed, but the youngest age of 0-3 years old has a significant impact on lowering the employment threshold for married women. Luo Yuan (2009) concluded that age, health status, and children's age have significant negative effects on married women's labor supply decisions. The education level factor has a significant positive effect on the labor supply decision of married women. Wang Guanghui (2010) argues that as married women become more educated, their income also increases significantly. Dou Zhang Tianzi (2023) established the Logit model to conduct an empirical analysis of the married female labor force from the aspects of personal factors and family factors, and believed that education level and spouse annual income have a significant impact on the willingness of married female labor force to work in the nearest place. Zhang Kang et al. (2021) said that with the improvement of urbanization, the increase in the number of young children has an increasing negative impact on the employment and self-employment of married women, and the increase in the number of children aged 0~2 has a greater negative impact on the employment and self-employment of married women than the increase in the number of children aged 3~6.

In addition, by comparing various models in machine learning, this paper comprehensively selects multiple regression models, decision trees, random forests, XGboost, logistics, support vector machines and other models for training and simulation, which have their own advantages, among which multiple regression models are used to quantify the linear relationship between multiple independent variables (married female characteristics) and dependent variables (the age of the youngest child). Wu et al. (2011) showed that decision trees are a common machine learning algorithm, and their structure is presented as a tree-like structure, in which each internal node represents an attribute judgment, each branch represents a possible value of the property, and each leaf node represents a category or a numeric value. Zhang Guijie et al. (2008) believe that decision tree algorithm is the most effective method to solve classification problems. Zhang Guijie et al. (2023) believe that decision trees have the advantages of discovering nonlinear relationships of latent variables, dealing with multi-classification problems, and selecting features when using decision trees to study the factors affecting the mental health of college students. It is more conducive to our next step of analysis and has a high comparability compared with multiple regression models. Yao et al. (2014) argue that although random forests may produce generalization errors within a certain limit, random forests do not produce overfitting problems with the increase of decision trees. Random forest (RF) Breiman is an ensemble machine learning method that uses the random resampling technology booststrap and node random splitting technology to construct multiple decision trees and obtain the final classification result through voting. XGboost is an efficient gradient boosting algorithm that is able to handle complex nonlinear relationships and provide ranking of feature importance. Although the logistic model is mainly used for binary classification problems, in this study, it can be applied to the prediction of whether or not to have the youngest child by setting thresholds, for example.

Based on this, there are significant differences between the research methods in this paper and those in the previously mentioned literature. This article is not limited to the application of a single machine learning model, but also compares the advantages and disadvantages of multiple machine learning models and their characteristics. Specifically, this paper carefully selects multiple regression models, decision tree models, random forest models, XGboost models, logistic regression models, and support vector machine models for systematic research and comparison. Rather than only pursuing a major conclusion, this paper focuses more on the simulation effect of various models when processing the data selected in the article. Through detailed comparison and analysis, this paper aims to reveal the differences between different models in terms of prediction performance, data adaptability, and interpretability, so as to find the best prediction model. In this process, this study not only focuses on the prediction accuracy of the model, but also pays attention to the feasibility and reliability of the model in practical application. By comparing the simulation effects of various models, a series of conclusions can be drawn, which will help to deeply understand the influencing factors of married women's reproductive decisions, provide a scientific basis for the formulation of relevant policies, and then promote the optimization of population structure and the sustainable development of society.

## II. DATA DESCRIPTION AND ANALYSIS

### 2.1 Data description:

This study uses the data table in the SATA learning package specifically for the employment of married women as the core data source, and selects the excel format for data storage after downloading, the dataset contains 23 different attributes, covering multiple aspects related to the employment of married women, and contains a total of 753 records, where the meaning of the attributes is shown in Table 1.

**TABLE 1**
**DATA SOURCES AND THEIR MEANINGS**

| Attribute | Meaning |
|---|---|
| inlf | Whether a married woman is employed (1 means employed, 0 means unemployed) |
| hours | Number of hours worked |
| kidslt6 | Whether you have a child younger than 6 years old (1 means yes, 0 means none) |
| kidsge6 | The age of the child under 6 years old (the value here is from 0 to 6) |
| age | Married female age |
| edu | Educational attainment of married women |
| wage | Salary level for married women |
| repwage | Average wage |
| hushrs | Working hours of a married woman's husband |
| husage | The age of the married woman's husband |
| husedu | Educational attainment of the married woman's husband |
| huswage | The level of salary of a married woman's husband |
| faminc | Household income |
| mtr | Average productivity |
| matheduc | Married female mother's level of education |
| fatheduc | Married female father's level of education |
| unem | Number of years of unemployment |
| city | Whether the married woman is a town resident (1 means yes, 0 means no) |
| nwifeinc | Wife's income |
| lwage | Minimum wage level |
| exper | Work experience |
| expersq | Work experience squared |
| college | University education (1 means yes, 0 means no) |

Based on the research in this paper, kidslt6 was selected as the dependent variable of the study, and the rest was used as the independent variable of model training to construct the training set and test set of the research data on the influence relationship between the characteristics of married women and the age of whether they have the youngest child. At the beginning of this paper, multiple regression models were selected to construct regression prediction models. It is compared with decision tree, random forest, xgboost, support vector machine, logistic and other models to observe whether the multiple regression model is the best prediction model for this dataset.

### 2.2 Descriptive statistical analysis of data:

#### 2.2.1 Matrix diagram:

In the Python environment, in order to deeply explore the correlation between the variables in the dataset, especially the potential influence of the independent variable on the dependent variable, the correlation drawing package in Python was used to draw a matrix plot of the correlation scatter plot between the variables. In the matrix chart, the horizontal and vertical coordinates are composed of individual features, and in each cell of the matrix chart is a scatter plot between the corresponding abscissa variable and the corresponding ordinate variable, and the icon reflected by the main diagonal of the matrix chart is a column chart display of the changes of each variable.

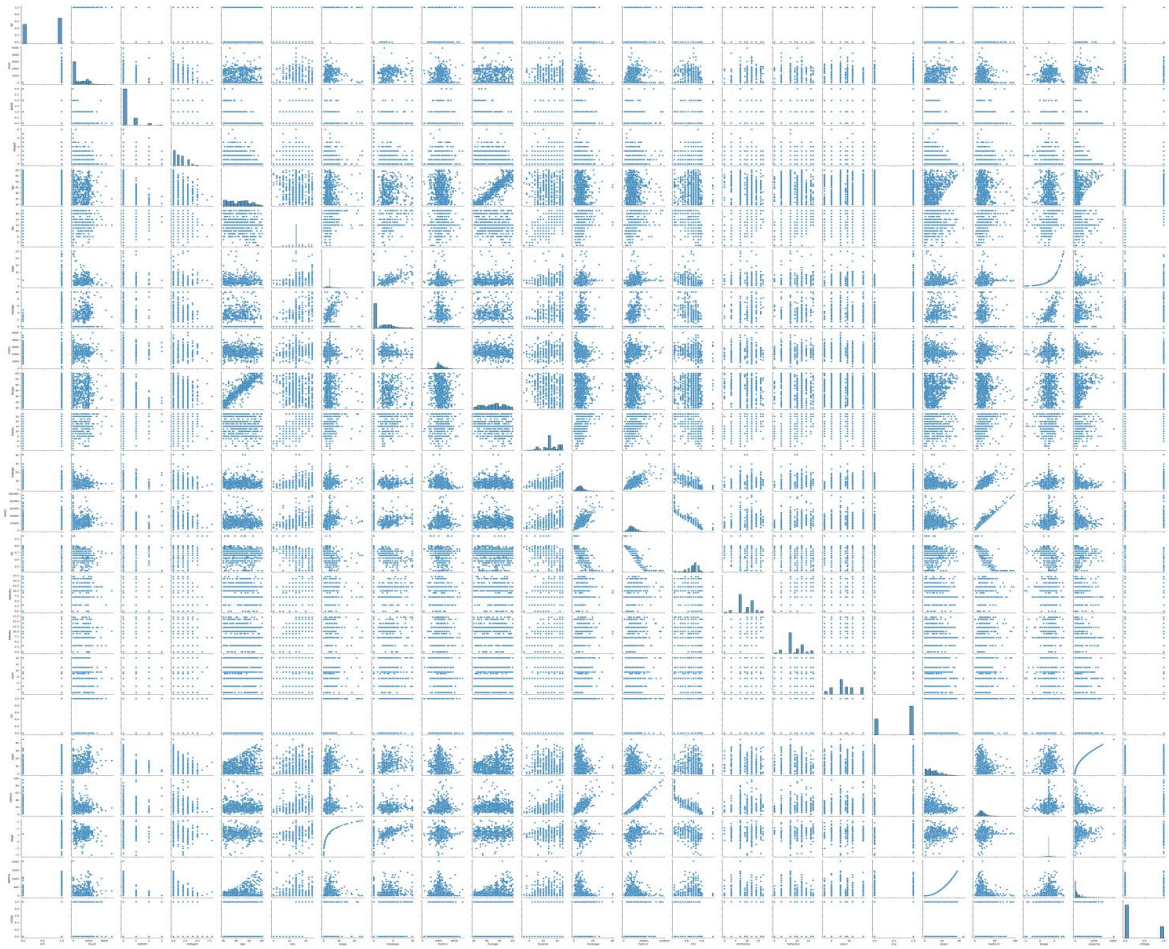

**FIGURE 1: Correlation between variables, scatter plot matrix plot**

You can zoom in on the graph, first look at the information displayed by each column chart on the main diagonal of the matrix chart, the change form of each variable is different, such as inlf variable, because it is classified data, so the column chart shown in the graph is the corresponding 1 and 0 of a data distribution, the same is true for the graphics of kidslt6, kidsge6, city, college and other variables, and then look at age, edu, wage In addition to categorical variables, the column charts of the data show different shapes, from which we can see the distribution of each data, or there is a left-biased distribution, a right-biased distribution, and most of them are normally distributed, which are all information reflected in the data. In addition to observing the graph of the main diagonal, you can also randomly view the relationship between any two variables, some scatter plots are exponentially distributed, some are concentrated distributed, and some show obvious positive correlation trends and so on. Based on the nature of the dependent variable, the relationship between the graph and other independent variables is not obvious, and the correlation between them is very low, but more data information can be viewed through the matrix chart, which is convenient for later prediction and interpretation, and can also reflect certain practical problems and data trends.

#### 2.2.2 Heat map:

The df.corr() code is used to view the correlation coefficients between the variables, and based on this, a heat map of the correlation coefficients between the variables is drawn, and the following figure 2 is obtained.

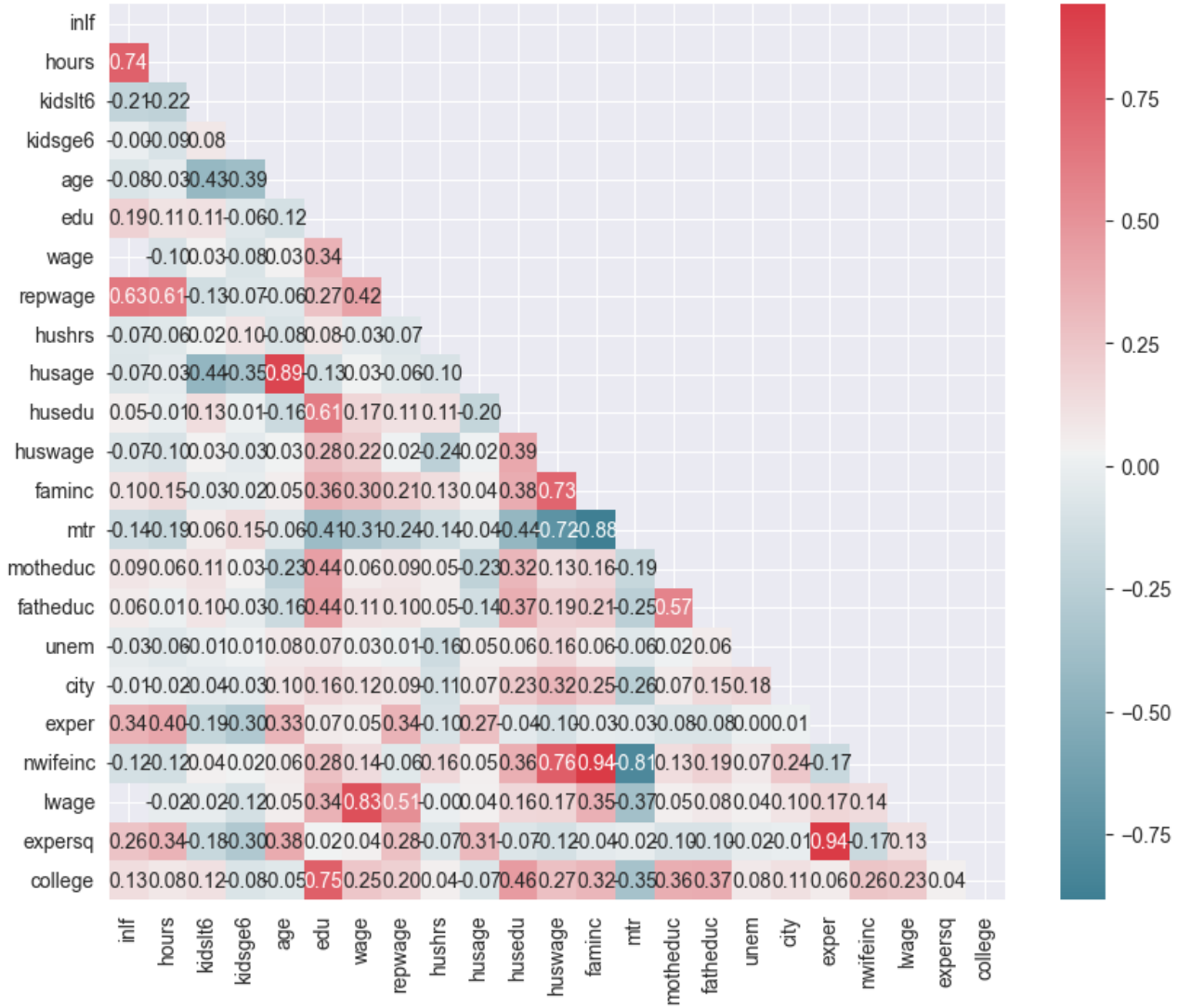


**FIGURE 2: Heat map of the correlation between variables**

Heatmaps are a highly intuitive visualization tool that clearly reveals the strength of correlations between variables in a dataset. In the heat map, it is possible to visualize which variables have significant correlations with each other by varying shades of color. Specifically, cells marked in red and blue represent a high degree of positive and negative correlation, while light blue may indicate a weak correlation (see Figure 2), which allows you to look at the relationship between the dependent and independent variables to set up a model to train and test the data, to test the research hypothesis or find the best predictive model.

## III. THEORETICAL EXPLANATION OF THE MODEL

### 3.1 Theoretical analysis and research hypothesis:

#### 3.1.1 Multiple regression model theory and advantages:

Multiple regression is a statistical analysis method used to explore the relationship between two or more variables, based on simple linear regression analysis, and can be used to predict the correlation between multiple independent and dependent variables. In the multiple regression model, the change of the dependent variable can be explained by the change of the independent variable and the disturbance of random factors, so the results of the multiple regression model analysis can intuitively show the change relationship between the dependent variable and the respective variable and random factors, which is more convenient for empirical analysis. Based on the research question of the article, kidslt6 was set as the dependent variable and the rest of the features as its independent variables.

### 3.1.2 Research hypotheses:

It is well known that the birth of a child is strongly influenced by a combination of family characteristics and a number of factors that make a married woman personal. From a theoretical point of view, with the increase in the proportion of married women in employment and the increase in family income, the number of children they raise may increase correspondingly, because the improvement of economic conditions may provide material security for raising more children, and the improvement of the social status of working women may also prompt them to be more willing and qualified to have more children. Realistically, however, the situation may not be so straightforward. As society continues to progress and develop, the improvement in the employment situation of married women may mean that they invest more time and energy in the workplace, which may in turn show a negative correlation with having the youngest children. Similarly, the relationship between other independent variables and dependent variables can be analyzed. Here are some assumptions made in this article:

H1: inlf, hours, age, kidsge6, matheduc, city were negatively correlated with kidslt6

H2: hushrs, husage, husedu, mtr, faminc, wage, expersq, college, etc. were positively correlated with kidslt6

### 3.1.3 Model theory:

Based on the selection of data features, the equations of the multiple regression model are as follows:

$$\hat{Y}(kidslt6) = \beta_0 \text{inlf} + \beta_1 \text{hours} + \beta_2 kids\text{ge6} + \beta_3\, exp\, e\, rsq + \beta_4 age + \beta_5 wage + \beta_6 repwage + \beta_7 hushrs + \beta_8 husage + \beta_9 husedu + \beta_{10} fa\, min\, c + \beta_{11} mtr + \beta_{12} mathedu + \beta_{13} \text{fathe}du + \beta_{14} unem + \beta_{15} city + \beta_{16} \text{exper} + \beta_{17} nwifeeinc + \beta_{18} college + \mu \quad (1)$$

The variables are interpreted as follows: inlf indicates whether or not employed, where "inlf=1" indicates employment and "inlf=0" indicates unemployment. Kidsge6 indicates the number of children under the age of 6; hours indicates the number of hours worked; expersq denotes the square of the experience; age denotes age; wage indicates the level of wages; Repwage indicates average salary; hushrs indicate the number of hours worked by married female husbands; husage indicates the age of the married woman's husband; husedu indicates the level of education of the married female husband; faminc indicates the status of household income; MTR stands for Household Productivity; Matheduc indicates the education of a married female mother, and fatheduc indicates the education of a father; unem indicates unemployment status; City means city, 1 means in city, 0 means not in city; Exper means work experience. Nwifeinc indicates the wife's income; College indicates whether you have attended a university, where "college=1" indicates that you have attended a university and "college=0" indicates that you have not attended a university; μ denotes the random perturbation term of the model, which is the random error term of the model.

Through the construction of the model, more direct results can be obtained in the subsequent training and testing, and the data can be regressed by using the least squares method (originally the data was initially processed, including the test and correction of the multicollinearity of the model, the test and correction of heteroskedasticity, autocorrelation, etc., and the stationarity of the model, but in view of the needs of this paper, these processes are directly ignored, and the original data is directly used for regression testing and the results are used as the basis for comparison between models). The results of each variable and parameter are obtained.

## IV. FEATURE PROCESSING AND MODEL TRAINING

### 4.1 Feature processing:

In the data preprocessing stage, this study uses the code user_info.info() to view the information of the data and observe whether there are potential problems such as outliers and missing values in the data. It is important to note that none of these data with missing values are categorical data, but continuous numerical data. Considering that the lack of categorical data may involve data integrity and classification accuracy, and the absence of continuous data may be compensated by a reasonable filling strategy, this study decided to use the mean filling method for the variables with missing values, and the results in Table 3 were obtained.

**TABLE 2**
**CLASS 'PANDAS.CORE.FRAME.DATAFRAME**

| Number | Column | Non−Null Count | Dtype |
|---|---|---|---|
| 0 | inlf | 753 non−null | int64 |
| 1 | hours | 753 non−null | int64 |
| 2 | kidslt6 | 753 non−null | int64 |
| 3 | kidsge6 | 753 non−null | int64 |
| 4 | age | 753 non−null | int64 |
| 5 | edu | 753 non−null | int64 |
| 6 | wage | 753 non−null | int64 |
| 7 | repwage | 753 non−null | int64 |
| 8 | hushrs | 753 non−null | int64 |
| 9 | husage | 753 non−null | int64 |
| 10 | husedu | 753 non−null | int64 |
| 11 | huswage | 753 non−null | int64 |
| 12 | faminc | 753 non−null | int64 |
| 13 | mtr | 753 non−null | int64 |
| 14 | matheduc | 753 non−null | int64 |
| 15 | fatheduc | 753 non−null | int64 |
| 16 | unem | 753 non−null | int64 |
| 17 | city | 753 non−null | int64 |
| 18 | nwifeinc | 753 non−null | int64 |
| 19 | lwage | 753 non−null | int64 |
| 20 | exper | 753 non−null | int64 |
| 21 | expersq | 753 non−null | int64 |
| 22 | college | 753 non−null | int64 |

### 4.2 Model training:

#### 4.2.1 Multiple regression model:

Based on the data obtained by feature processing of the data, the multiple regression model was constructed, and the least squares (OLS regression) regression analysis was realized by the code, and the following least squares regression result table was obtained.

**TABLE 4**
**OLS REGRESSION RESULTS**

| | | | |
|---|---|---|---|
| **Dep. Variable:** | kidslt6 | **R-squared (uncentered):** | 0.425 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.409 |
| **Method:** | Least Squares | **F-statistic:** | 27.11 |
| **Date:** | Thu, 28 Dec 2023 | **Prob (F-statistic):** | 1.32e-74 |
| **Time:** | 22:12:20 | **Log-Likelihood:** | -443.38 |
| **No. Observations:** | 753 | **AIC:** | 926.8 |
| **Df Residuals:** | 733 | **BIC:** | 1019 |
| **Df Model:** | 20 | | |
| **Covariance Type:** | nonrobust | | |

| Variable | Coefficient | Std. Error | t-value | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| inlf | -0.1820 | 0.053 | -3.413 | 0.001 | -0.287 | -0.077 |
| hours | -6.307e-06 | 4.24e-05 | -0.149 | 0.882 | -8.95e-05 | 7.69e-05 |
| expersq | 0.0003 | 0.000 | 1.418 | 0.157 | -0.000 | 0.001 |
| kidse6 | -0.0538 | 0.014 | -3.787 | 0.000 | -0.082 | -0.026 |
| age | -0.0186 | 0.005 | -4.020 | 0.000 | -0.028 | -0.010 |
| edu | 0.0134 | 0.012 | 1.073 | 0.284 | -0.011 | 0.038 |
| wage | 0.0102 | 0.009 | 1.151 | 0.250 | -0.007 | 0.027 |
| repwage | 0.0068 | 0.011 | 0.595 | 0.552 | -0.016 | 0.029 |
| hushrs | -1.802e-05 | 2.79e-05 | -0.647 | 0.518 | -7.27e-05 | 3.67e-05 |
| husage | -0.0155 | 0.004 | -3.522 | 0.000 | -0.024 | -0.007 |
| husedu | 0.0073 | 0.007 | 1.021 | 0.308 | -0.007 | 0.021 |
| faminc | -9.608e-06 | 9.89e-06 | -0.971 | 0.332 | -2.9e-05 | 9.81e-06 |
| mtr | 2.0558 | 0.206 | 9.964 | 0.000 | 1.651 | 2.461 |
| motheduc | -0.0033 | 0.006 | -0.538 | 0.591 | -0.015 | 0.009 |
| fatheduc | 0.0011 | 0.006 | 0.195 | 0.845 | -0.010 | 0.013 |
| unem | 0.0024 | 0.005 | 0.439 | 0.661 | -0.008 | 0.013 |
| city | -0.0186 | 0.037 | -0.508 | 0.612 | -0.090 | 0.053 |
| exper | -0.0042 | 0.006 | -0.690 | 0.491 | -0.016 | 0.008 |
| nwifeinc | 0.0220 | 0.010 | 2.233 | 0.026 | 0.003 | 0.041 |
| college | 0.1043 | 0.055 | 1.893 | 0.059 | -0.004 | 0.212 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 224.371 | **Durbin-Watson:** | 1.964 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 677.117 |
| **Skew:** | 1.455 | **Prob(JB):** | 9.25e-148 |
| **Kurtosis:** | 6.621 | **Cond. No.:** | 3.38e+05 |

***Note: Marked in red is significantat 0.1 and marked in blue is significant at 0.5***

By calculating the specific formula of the regression model in Python and combining the results obtained by the least squares method, the specific form of the model can be obtained:

Y= - 0.182 inlf - 0.000 hours - 0.053 kidsge6 + 0.000 expersq - 0.019 age + 0.013 edu+ 0.010 wage + 0.007 repwage - 0.000 hushrs - 0.015 husage + 0.007 husedu - 0.000 faminc + 2.055 mtr - 0.003 matheduc + 0.001 fatheduc + 0.002 unem -0.018 city - 0.004 exper+ 0.022 nwifeinc + 0.104 college + 1.031 (2)

### 4.2.2 Decision tree model:

A decision tree algorithm is a common machine learning algorithm used to solve classification and regression problems. The main idea is to construct a tree structure by recursively segmenting the dataset, where each node represents a feature, each branch represents the value of the feature, and each leaf node represents a category or a numeric value. The Classification and Regression Tree algorithm is suitable for modeling complex data with multiple feature variables, and has the advantages of simple extraction rules, high accuracy and strong interpretability. In view of this, this paper uses a simple decision tree model to predict that married women will have children under 6 years of age. The generation of decision trees is a binary recursive partitioning process, which adopts a splitting criterion that minimizes the square error. The splitting process starts from the root node, and each time the feature attribute with the smallest square error and its feature attribute value are taken as the optimal splitting attribute and the optimal splitting attribute value, all the features and eigenvalues are traversed, and the input is divided into two regions in turn. Then, the above partitioning process is repeated for each sub-region until the stop condition is satisfied, and a regression decision tree is generated.

The model compares the magnitude of the loss function error by the squared error, and uses the squared error minimization criterion to determine the optimal segmentation point and the predicted value. According to the principle of least squares, the mean of all outputs on the subspace can be expressed as, based on which the sharding attribute is selected to divide the input space, so that it can go through all the input eigenvariables and their eigenvalues to find the optimal sharding point, and then divide the space into two subregions according to the slicing point, until the two subregions cannot be divided, and the corresponding optimal output value can be obtained. The decision tree constructed in this paper uses the square error minimization as the splitting criterion for the selection of characteristic variables, the maximum depth is 10, the minimum number of samples required for internal node division is 2, and the maximum leaf node is 10. The following values about the accuracy of the model are obtained by the code implementation Accuracy_TAET which is 0.786, the mean result of cross-validation of the decision tree, and the cross-validation scores of the decision tree model are [0.74834437 ,0.79470199 ,0.69536424 ,0.7 ,0.64], and the average cross-validation score is 0.715。

After training and testing the model data, the algorithm related to cross-validation scores was used to perform five-fold cross-validation, and each cross-validation used a different subset of data to train and test the model, and the results are shown above, and the decision tree cross-validation results can be known, where the first cross-validation score is 0.74834437, the second score is 0.79470199, the third score is 0.69536424, and the fourth score is 0.7, with a score of 0.64 for the fifth time. According to the results of the five cross-validations, the average cross-validation score is obtained, and the value is the average cross-validation score = ( 0.74834437 + 0.79470199 + 0.69536424 + 0.7 + 0.64 )/5 = 0.71568. This average cross-validation score represents the average performance of the model over multiple cross-validations. A higher score usually indicates a better performance of the model. In this example, the average cross-validation score is 0.715, which means that the model has achieved an average accuracy of about 71.6% across multiple cross-validations, and the model performs well.

### 4.2.3 Random forest model:

Random forest is an ensemble learning model, and each decision tree in the random forest is built based on random samples and random features, which can reduce the risk of overfitting and improve the generalization ability of the model. Random forests can combine multiple decision trees for classification or regression, which can reduce the error of a single decision tree and improve the accuracy of the model. Random forests can process a large number of features and samples, making them suitable for datasets of all sizes. Random forests can assess the importance of features and help with feature selection. Random forests can be computed in parallel, speeding up model training. In the process of training, the data of the training set and the test set were divided into 8:2 and a random forest classifier containing 100 trees was created, and the seed of the random number generator was set to 42. Finally, the accuracy of the model is 0.814.

### 4.2.4 Gradient boosting regression tree:

Gradient Boosting Regression Tree is a machine learning method for regression problems. Unlike random forests, gradient boosted regression trees train models by fitting residuals stepwise rather than by building multiple parallel models to integrate predictions. The basic idea of gradient boosted regression trees is to combine multiple weak learners into one strong learner. At each iteration step, the new decision tree attempts to correct the errors of the previous tree. At the heart of this approach is the use of a gradient boosting strategy, which optimizes the gradient of the loss function with respect to the parameters. Based on this, a model was also selected for comparison in the problem study, and the following settings were made in the training

process, and the data of the training set and the test set were divided into a ratio of 8:2, and the number of random seeds was 2. After training, the accuracy on training set is 0.99003, and the accuracy on test set is 0.834.

#### 4.2.5 XGboost model:

XGBoost is a machine learning algorithm based on gradient boosting decision trees. It fits the data by iteratively adding trees and uses a gradient boosting strategy to optimize the loss function. XGBoost uses a forward distribution algorithm for greedy learning during training, and each iteration learns a CART tree to fit the residuals of the prediction results of the previous t-1 trees and the true values of the training samples. The core principle of XGBoost is that weak classifiers are integrated together to form strong classifiers. XGBoost adds a regular term to the cost function to control the complexity of the model. XGBoost supports multi-threading (parallel) gain calculation of each feature when splitting nodes, so the algorithm is faster and has a relatively higher accuracy. During the training process, the following settings were made to divide the data of the training set and the test set in an 8:2 ratio. 100 trees are selected with a learning rate of 0.1, a random sampling ratio of 0.3 for each tree training, and a maximum depth of 5 for each tree. After training, the accuracy of the model is 0.827.

#### 4.2.6 Logistic model:

Logistic Regression is a machine learning method for classification problems. It uses the logistic function (also known as the Sigmoid function) to convert the results of a linear regression into a probability value, enabling the classification of data. The basic principle of logistic regression is to construct a regression equation from training data, and then use this equation to predict the category of new unknown data. During training, the algorithm looks for the best combination of parameters to minimize the error between the predicted and true categories. In the problem study, the following settings were made in the training process, and the data of the training set and the test set were divided into a ratio of 8:2, and the number of iterations was set to 1000, and finally the accuracy of the model was 0.79.

#### 4.2.7 K nearest neighbor model:

The K-Nearest Neighbor (KNN) model is a very simple and intuitive model, and it is important to choose an appropriate value of k (k is an integer) in the K-nearest neighbor model. The choice of k-value will determine how the model classifies or predicts. In the problem study, the test data is set to 0.25 of the original data, the number of random seeds is 10, and the number of neighbors in the KNN regressor is 502. After the model is trained, the accuracy of the model is 0.830.

#### 4.2.8 Support vector machine model:

A Support Vector Machine (SVM) is a supervised learning model used for classification and regression analysis. It does this by finding decision boundaries that maximize the separation of different categories of data points. The basic idea of the SVM is to find a hyperplane that maximizes the separation of different classes of data points. In the case of nonlinear separability, SVMs can map the input space to a high-dimensional feature space by using a kernel function, and then look for a hyperplane in that feature space that maximizes the separation of data points. In model training, the test set is set to the original 0.3, the random number generator seed is 42, and the regularization parameter is 1. The model is trained and the accuracy of the model is 0.778.

## V. MODEL COMPARISON AND ANALYSIS

### 5.1 Model evaluation index standards:

The following evaluation indicators are used in problem research: (1) $R^2$ (R-squared) is a statistical index used to measure the goodness of fit of the regression model. $R^2$ represents the proportion of the variation in the dependent variable that can be explained by the independent variable. It has a value between 0 and 1, with closer to 1 indicating a better fit for the model and the more explanatory the independent variable is to the dependent variable. (2) The Average Cross-Validation Score is a metric that evaluates the performance of a machine learning model, and the average score is calculated by dividing the dataset into multiple subsets and training and validating the model multiple times on these subsets. This score reflects the model's performance on unknown data and helps to understand the model's generalization ability. (3) Accuracy is an important evaluation metric used to measure the proximity between the result of a measurement, calculation, or observation and the true value. Accuracy is a key evaluation criterion in fields such as scientific experiments, engineering, statistics, and more. (4) Accuracy on training set is limited to using the accuracy on the training set as an evaluation indicator. In machine learning, accuracy on the training set usually refers to the classification or regression accuracy of the model on the training data, which does not fully reflect the performance of the model in real-world applications. (5) Accuracy on test set is one of the important

metrics to evaluate the performance of machine learning models. A test set is a separate dataset from the training set that evaluates how the model performs on unknown data. By calculating the accuracy on the test set, you can get an idea of how the model performs in a real-world scenario.

### 5.2 Model comparison:

**TABLE 3**
**THE VALUE OF THE EVALUATION INDEX OF EACH MODEL**

| Model | Evaluation indicators | Value |
|---|---|---|
| Multiple regression models | $R^2$ | 0.452 |
| | Average cross-validation score | 0.102 |
| Decision tree model | Accueacy | 0.781 |
| | Average cross-validation score | 0.715 |
| Random forest model | Accueacy | 0.814 |
| Gradient boosting regression tree | Accuracy on training set | 0.99 |
| | Accuracy on test set | 0.834 |
| XGboost model | Accueacy | 0.827 |
| Logistic model | Accueacy | 0.79 |
| K-nearest neighbor model | Accueacy | 0.83 |
| Support vector machine models | Accueacy | 0.778 |

From the above table, it can be seen that the accuracy of gradient boosted regression tree model, K-nearest neighbor model, XGBoost model and random forest model is not much different, the accuracy of the improved regression tree model is the highest, the model is the best, the accuracy of the decision tree model, the support vector machine model and the logistic model is also higher, and the performance of the model is also better, but the fitting effect of the multiple regression model is the worst. It shows that the data used in this paper are better than the linear regression model compared with the regression machine learning algorithm.

### 5.3 Model analysis and interpretation:

Based on the training results obtained in the model training process, for the regression model, there are several significant factors in the model of the relationship between the comprehensive characteristics of married women and whether they have children aged 0-6 years, and the characteristics of married women are inlf, expersq, kidsge6, age, edu, wage, husage, husedu, faminc, mtr, exper, nwifeinc, college The relationship of this independent variable to the dependent variable is significant at different levels of significance, and it can be concluded that both hypothesis H1 and hypothesis H2 are true.

In addition, the accuracy of other models is due to the degree of fit of multiple regression models, and the best model is the gradient boosted regression tree model, as it is a machine learning method for regression problems. Unlike random forests, gradient boosted regression trees train models by fitting residuals stepwise rather than by building multiple parallel models to integrate predictions. The basic idea of gradient boosted regression trees is to combine multiple weak learners into one strong learner. At each iteration step, a new decision tree corrects the mistakes of the previous tree. At the heart of this approach is the use of a gradient boosting strategy, which optimizes the gradient of the loss function with respect to the parameters. In turn, a high accuracy rate can be obtained during the training process. The K-nearest neighbor model is second, and the XGboost model also performs well, i.e., the classification or ensemble model is better than the linear regression model. There are several reasons why the linear regression model is not well fitted: first, there is no multicollinearity, heteroskedasticity, autocorrelation and other tests and corrections for each variable during model training, and there is no stationarity test for the model; Second, the dependent variable is a categorical data, which is regressed with continuous data, and the effect is not significant, and there

is no good reference in the selection of variables, and the multiple linear regression model does not perform as well as the gradient to improve the regression tree and other machine learning models.

## VI. CONCLUSIONS AND RECOMMENDATIONS

Based on the research on the comprehensive characteristics of married women and whether they have children aged 0-6 years old and other methods such as machine learning, this paper draws the following conclusions: (1) A series of comprehensive characteristics such as employment status, age level, and mother's education level of married women have a significant negative correlation with the decision of whether to have children aged 0-6. In other words, these characteristics will influence to some extent whether married women choose to have children. Specifically, as married women increase in employment, age, and mothers become more educated, their willingness to have children appears to be decreasing, leading to a corresponding decrease in the number of children aged 0-6 years. (2) The characteristics of married women's husbands and the characteristics of the family as a whole had a positive and significant impact on whether they had children aged 0-6 years. This means that when married women have good husbands and good family characteristics, they are more inclined to want married women to have children aged 0-6 years, and it is worth noting that the higher the education level of the father of a married woman, the positive impact on his or her childbearing. This finding seems to hint at a general psychological tendency that men, whether as husbands or fathers, expect married women to be able to have young children to some extent. (3) Based on the results of multiple regression model, it is found that when conducting research or selecting models, it is necessary to fully consider various possible influencing factors and conduct a comprehensive analysis. Avoid subjective selection of models, but there is a real possibility that they do not fully match the characteristics of the data. Therefore, before making a decision, it is necessary to carefully evaluate the pros and cons of various models to ensure that the research or prediction can achieve better results. (4) After comparing various models, it is found that the gradient boosted regression tree model, K-nearest neighbor model, XGBoost model and random forest model are significantly better than linear regression models. The results suggest that when dealing with complex data problems, more diverse models and algorithms should be tried to find the most suitable solution for the data characteristics, so as to better capture the nonlinear relationships and latent features in the data, so as to provide deeper and more comprehensive insights for research. The research in this paper has the following three policy implications:

First, policies tend to encourage childbearing. In order to effectively encourage childbearing, policymakers should focus on improving the family situation of married women, thereby motivating them to have children. Specifically, a series of measures such as information platform sharing and sharing sessions can be used to promote married women and their husbands to receive a higher level of university education, so as to improve the education level of the whole family. In order to better broaden their horizons and enhance the efficiency and competitiveness of families in production and life. At the same time, by improving family productivity, such as providing a family-friendly working environment and a flexible work system, the pressure on married women between home and the workplace can be further reduced, and a more relaxed and suitable environment for them to have children can be created.

Second, to promote women's employment, we should take practical and effective measures to protect women's equal rights and interests in the workplace based on the reality of their family situation. This includes, but is not limited to, promoting equity in employment in society and ensuring that women are not discriminated against or excluded on the basis of their gender. At the same time, the principle of fairness should be promoted within the family to avoid placing the responsibility of raising young children entirely on the shoulders of women, thereby limiting their career choices and career development. Men should be encouraged to participate more in family life, share childcare responsibilities and create a more equal and free employment environment for women.

Third, married women should fully consider their age when choosing to have children. Because as you get older, the risks and difficulties you face in having children also increase. Therefore, it is advisable for married women to plan their family plans as early as possible within the appropriate age range to ensure the health and safety of themselves and their children. At the same time, society should also provide more reproductive support and protection for married women, such as providing high-quality medical resources and parenting guidance, so as to help them better cope with various challenges in the reproductive process.

Fourth, when analyzing the data used in this paper, we should take a comprehensive and detailed comparative analysis approach. By comparing the similarities and differences between different data, we can gain a deeper understanding of the nature and laws of the problem. At the same time, in order to find the most suitable analysis model, we also need to explore

and try in many aspects based on the actual situation. The same principle applies to other research work. We should always maintain objectivity and rigor when analysing and processing data to ensure the accuracy and reliability of the research results.

## REFERENCES


[1] Huang Qian, Cao Shurui. The impact of the number of children on the employment choice of married women[J].China Economic Issues,2022(06):67-81.)
[2] Song Alun, Zhang Feng. Analysis of the influence of family characteristics on the employment quality of married women[C]//Sichuan Labor and Social Security Magazine Publishing Co., Ltd. Proceedings of the Labor and Social Security Research Conference (15), 2022:4.
[3] Yan Yu. Age of the youngest child and married women's labor force participation[D].Southwestern University of Finance and Economics,2022.
[4] Luo Yuan. An empirical study on the labor supply decision of married women[D].Nanjing Agricultural University,2009.
[5] Wang Guanghui, Zhang Shiwei. An empirical analysis of the impact of education on the employment and income of married women in China[J].Science, Economy & Society,2010,28(04):83-87+93.
[6] Dou Zhangtianzi. Analysis of influencing factors of local and nearby employment of rural married female labor force: Based on Logit model[J].Tianjin Agricultural Sciences,2023,29(01):45-48+61.
[7] Zhang Kangzi, Wang Yadi. Research on the influence of childbearing on married women's employment choice[J].Public Management Review,2021,3(03):53-75.
[8] Wu Xiaogang, Zhou Ping, Peng Wenhui. Application of decision tree algorithm in the evaluation of college students' mental health[J].Computer Applications and Software,2011,28(10): 240-244.
[9] Zhang Guijie, Wang Shuai. Research on ID3 algorithm for decision tree classification[J].Journal of Jilin Normal University(Natural Science Edition,2008,29(3):135-137.
[10] Zhang Guijie, Wang Xiaocan, Xing Weikang, et al. Research on psychological assessment model based on decision tree classification algorithm[J].Journal of Jilin Normal University(Natural Science Edition),2023,44(04):123-130.
[11] Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature selection algorithm based on random forest[J].Journal of Jilin University(Engineering Science),2014,44(01):137-141.
[12] BreimanL.Randomforests[J]. Machine Learning,2001,45(1):5-32.
[13] Wu Xiaogang, Zhou Ping, Peng Wenhui. Application of decision tree algorithm in the evaluation of college students' mental health[J].Computer Application and Software,2011,28(10) : 240-244.
[14] Zhang Guijie, Wang Shuai. Research on ID3 algorithm for decision tree classification[J].Journal of Jilin Normal University(Natural Science Edition) ,2008,29(3) : 135-137.
[15] Hu Nan, Xue Fujing, Wang Haonan. Does Short-sightedness of Managers Affect Long-Term Investment? Management World,2021,37(05):139-156+11+19-21.
[16] Yang Jianfeng, Qiao Peirui, Li Yongmei, et al. Statistics and Decision,2019,35(06):36-40.
[17] Wu Changqi, Zhang Kunxian, Zhou Xinyu, et al. Digital transformation, competitive strategy choice and high-quality development of enterprises: Evidence based on machine learning and text analysis[J].Economic Management,2022,44(04):5-22.
[18] Li Bin, Shao Xinyue, Li Yueyang. China Industrial Economics,2019,(08):61-79.
[19] Huang Bowen, Liu Tianli. Population and Development,2024,30(03):38-50.
[20] Huang Qian, Cao Shurui. The impact of the number of children on the employment choice of married women[J].China Economic Issues,2022,(06):67-81.
[21] Zhang Kangzhi, Wang Yadi. Research on the influence of childbearing on married women's employment choice[J].Public Management Review,2021,3(03):53-75.
[22] Zheng Yuxin, Wu Yuxia, Mi Hong. Analysis of influencing factors of married female floating population having two children: Based on data analysis of Jiangsu, Zhejiang and Shanghai[J].Northwest Population,2023,44(06):100-114.