Predicting China's Marriage Rate using Machine Learning Models and Cross-Validation Evaluation

Deyu Zhang
Yunnan College of Foreign Affairs and Foreign Language [1]
Sirindhorn International Institute of Technology
Thammasat University [2]
Kunming, China [1]
Pathum Thani, Thailand [2]
6451301504@lamduan.mfu.ac.th

Thanaruk Theeramunkong
Sirindhorn International Institute of Technology
Thammasat University
Pathum Thani, Thailand
thanaruk@siit.tu.ac.th

Nirattaya Khamsemanan
Sirindhorn International Institute of Technology
Thammasat University
Pathum Thani, Thailand
nirattaya@siit.tu.ac.th

***Abstract*— After China's accession to the WTO, the country has experienced rapid development over the past 20 years, but the marriage rate has been declining. This study aims to analyze marriage data using machine learning, identify the factors influencing the marriage rate, and make predictions. The data covers seven variables including GDP, housing prices, fertility rate, and education level in 31 regions of China from 2003 to 2022. The study employed five machine learning algorithms with five models: Ridge, Lasso, Elastic Net, Polynomial, and Weighted Least Squares, along with three types of Cross-Validation techniques – Holdout, LOOCV, and K-Fold. By comparing the five regression models and evaluating the marriage rate models through Holdout, LOOCV, and K-Fold Cross-Validation, the results show that the Polynomial regression exhibits higher predictive accuracy, with MSE of 3.473 for Holdout, 3.001 for LOOCV, and 3.018 for K-Fold. Following closely is Ridge regression with MSE of 3.565 for Holdout, 3.326 for LOOCV, and 3.334 for K-Fold. These two models outperform the less performing Lasso and Elastic Net regressions. Weighted Least Squares show stable but slightly inferior performance. By comparing the predicted values for 2022 with the actual values, it is confirmed that Polynomial and Ridge regression closely align with the actual values, predicting the trend of changes in real and predicted values for each province, highlighting their effectiveness in predictive tasks involving complex data patterns.**

Keywords—Crude Marriage Rate, Ridge, Polynomial, Cross Validation, MSE, RMSE, R^2

I. INTRODUCTION

Marriage is significantly influential in various social aspects like well-being, reproduction, raising children, gender inequalities, criminal activities, and workforce engagement. China is recognized for its tradition of extensive marriage, although the tendency to postpone the first marriage is becoming more noticeable [1]. In recent years, there has been a rise in the age at which people get married and a decrease in fertility rates in China, potentially due to soaring housing costs serving as catalysts. Marriage is a pivotal social institution that profoundly influences various societal aspects like well-being, procreation, child upbringing, gender disparities, and crime rates. Additionally, it plays a vital role in addressing workforce availability in the job market. Nevertheless, there has been a significant drop in marriage rates across numerous countries, commencing in developed nations like Western Europe and the United States, and extending to East Asian nations such as Japan and South Korea, with China closely following suit [2]. The complexity of marriage in China goes beyond emotions and involves crucial economic factors. Important elements influencing marriage include GDP, housing prices, income, consumption, pension ratio, gender ratio, and education level [3]. At the same time, analogous issues are also surfacing in India, a heavily populated nation in Asia [4]. The dual structure of urban and rural areas, combined with rural and urban household registrations, incentivizes women to seek marriage opportunities in urban areas[5]. Factors that impact the marriage rate in China encompass the obligation for men to acquire matrimonial property prior to marriage, elevated real estate costs dissuading the youth and the savings ability of men and their families [6]. The male population exceeds the female population, contributing to gender imbalance as a key factor [7]. Emotional aspects play a significant role, but economic factors are crucial. Variables such as GDP, housing costs, income levels, consumption patterns, pension allocations, gender distribution, and educational attainment all contribute to the intricacy of marriage dynamics.

The main contributions of this study are as follows:

- By preprocessing the data using methods such as the most frequent method in Scikit-learn for handling missing values, scaling the data to a range of 0-1 to enhance the model's performance.
- The dataset spans from 2003 to 2022. In the study, the data from 2022 is separated as the testing dataset, while the data from 2003 to 2021 is split as the training dataset, segmented by year for model training.
- Five supervised regression models are simultaneously utilized in the study, namely Ridge, Lasso, Elastic Net, Polynomial, and Weighted Least Squares regression.
- Three cross-validation methods are employed, including Holdout, LOOCV, and K-Fold, for cross-validating the five machine learning algorithms.
- Four metrics are used to evaluate the models, namely MSE, RMSE, MAE, and R^2.
- Following model evaluation, predictions were made, and the two best-performing models were selected to forecast the crude marriage rate in 2022. The predicted values were compared with the actual values to examine and assess the models' performance.

The paper is structured as follows. Section 2 reviews the relevant literature review. Section 3 shows the overall

2833-8898/24/$31.00 ©2024 IEEE
DOI 10.1109/CyberC62439.2024.00045

research methodology and workflow, Section 4 shows the results and discussion and Section 5 shows the conclusion.

II. Literature Review

A. Related Studies

Regression models, especially Ridge, Lasso, Elastic Net, and polynomial regression, have been widely used in predicting marriage rates. These models are favored for their ability to handle multicollinearity and perform effective variable selection [8][9]. Cross-validation techniques, including Holdout, Leave-One-Out Cross-Validation (LOOCV), and K-fold cross-validation, are crucial for evaluating model performance, preventing overfitting, and ensuring robustness [10]. Comparative research indicates that polynomial regression and ridge regression generally outperform simple models in predicting marriage rates because they can model nonlinear relationships and handle multicollinearity effectively. These models offer more accurate predictions, which is crucial for understanding and forecasting marriage trends[11]. Based on this, our study utilized five machine learning models - ridge regression, lasso regression, elastic net regression, polynomial regression, and weighted least squares (WLS) - to analyze the marriage rate in China. The dataset includes seven variables such as GDP, housing prices, fertility rate, and education level, spanning 31 regions from 2003 to 2022. Through Holdout, LOOCV, and K-Fold cross-validation, our study found that polynomial regression and ridge regression were the most accurate models, with the predicted marriage rates for 2022 closely matching the actual rates [12].This underscores their effectiveness in capturing complex data patterns and their applicability to future population forecasts .

Key Points:

- Machine learning is highly effective in predicting socio-economic trends.
- The study compared Ridge, Lasso, Elastic Net, Polynomial, and WLS models.
- Cross-validation technique ensures the robustness of the evaluation.
- Polynomial and Ridge regression demonstrate higher predictive accuracy.

B. Term definition

According to the United Nations, the crude marriage rate (CMR) is a vital statistics summary rate based on the number of marriages occurring in a population during a given period, usually a calendar year. It is calculated as the number of marriages occurring among the population of a given geographical area during a given year per 1,000 mid-year total population of the same area during the same year. The formula for the crude marriage rate(CMR) is (1) [13].

$$CMR = \frac{Number\ of\ Marriage\ Registration\ Pairs}{Mid-Term\ Population} * 1000 \quad (1)$$

This study calculated the educational accomplishments of individuals aged 6 and above annually from 2003 to 2022, and the equation for the average years of education per capita is outlined within this context. The formula for the average years of Education is (2) [14].

$$Average\ years\ of\ education = \frac{(a*6+b*9+c*12+d*12+e*15+f*16+g*19)}{Total\ population\ aged\ 6\ and\ above} \quad (2)$$

Note that numbers are academic programs, letters are school levels.

- where a, b, c, d, e, f and g refer to the number of populations get primary school certificate, junior high school certificate, Senior high school, Vocational secondary school, College, Bachelor, Postgraduate (Master, PhD),respectively.

C. Regression Models of Machine Learning

The machine learning regression model is a statistical learning model used to predict continuous target variables [15]. It builds a prediction model by learning the relationship between input features and target variables in the training data set, and uses the model to predict new data. These five regression methods (ridge, lasso, elastic net, polynomial, weighted least squares) extend linear regression to address different data characteristics and problem requirements and improve model performance.

1) The Ridge Regression Model: The Ridge regression is a linear regression method that prevents overfitting by adding a penalty term of the L_2 norm (i.e., the sum of the squares of the weights) to the loss function. Its goal is to minimize the following loss function [8]. The formula for the ridge Regression is (3) .

$$min_\omega\left(\Sigma_{i=1}^{n}(y_i - w^T x_i)^2 + \alpha\Sigma_{j=1}^{P}\omega_j^2\right) \quad (3)$$

- a= α is the regularization parameter, which controls the size of the penalty term

2) The Lasso Regression Model: Lasso regression is a linear regression method that performs feature selection and prevents overfitting by adding a penalty term of the L_1 norm (i.e., the absolute value and sum of weights) to the loss function. Its goal is to minimize the following loss function [8]. The formula for the lasso regression is (4).

$$min_\omega\left(\Sigma_{i=1}^{n}(y_i - w^T x_i)^2 + \alpha\Sigma_{j=1}^{P}|w_j|\right) \quad (4)$$

- α is a regularization parameter that controls the size of the penalty term. Lasso regression can reduce some weights to zero, thereby achieving feature selection.

3) Elastic Regression Model: Elastic net regression combines the advantages of Ridge regression and Lasso regression by adding penalty terms of L_1 norm and L_2 norm to the loss function. Its goal is to minimize the following loss function [18]. The formula for the elastic net regression is (5).

$$min_\omega = \left(\Sigma_{i=1}^{n}(y_i - w^T x_i)^2 + \alpha_1\Sigma_{j=1}^{P}|w_j| + \alpha_2\Sigma_{j=1}^{P}w_j^2\right) \quad (5)$$

- α 1 and α 2 are regularization parameters that control the size of the two penalty terms.

4) The Polynomial Model: Polynomial regression is a method that extends linear regression by introducing polynomial terms of features to capture nonlinear

relationships. Its goal is to find a polynomial function that fits the data [8]. The formula for the polynomial regression is (6).

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_d x^d + \epsilon \quad (6)$$

- a= d is the highest degree of the polynomial, βi is the regression coefficient, and ϵ is the error term

5) The Weighted Least Squares Regression Model: Weighted least squares regression is a regression method that handles heteroskedasticity (i.e., different data points have different variances) by assigning a weight to each data point. Its goal is to minimize the weighted squared error [17]. The formula for the weighted least squares regression is (7).

$$min_{\omega} \Sigma_{i=1}^{n} \omega_i (y_i - w^T x_i)^2 \quad (7)$$

- a = i is the weight of the i-th data point, usually inversely proportional to the variance of that point

III. Research Methodology

A. Overall Methodology

In this section, This section first defines the problem description of China's marriage rate. Then, based on this description, data collection is conducted, specifically covering the independent variables that can be studied using quantitative methods. During the data preprocessing, missing data was cleaned. A predictive framework of 5 machine learning hybrid evaluation algorithms for the marriage rate was proposed. Within this framework, 5 machine learning

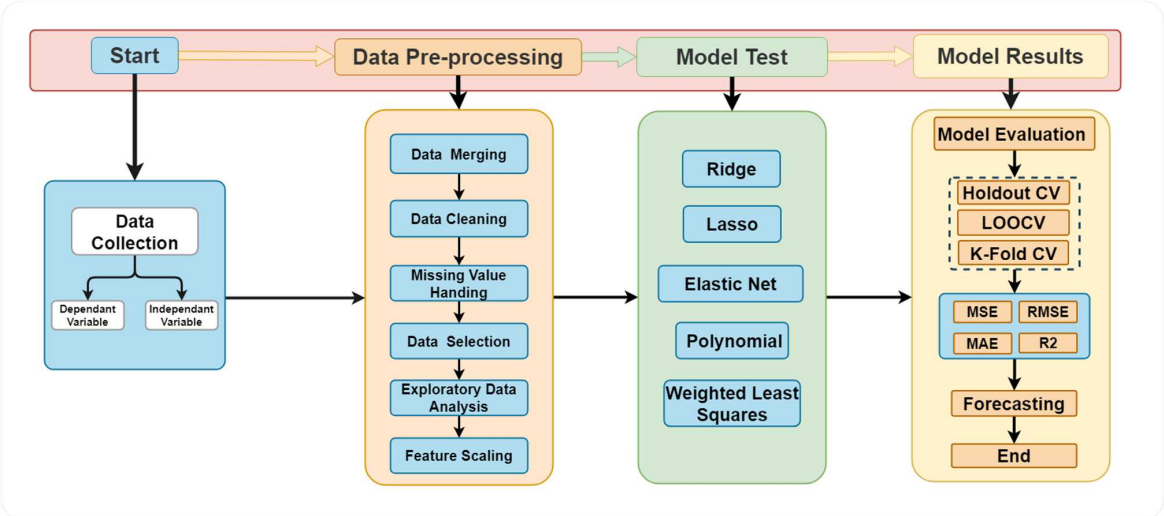

Fig. 1. Overall Methodology

algorithms were studied, namely Ridge, Lasso, Elastic Net, Polynomial, Weighted Least Squares, for model creation and evaluation of the marriage rate. Finally, through Holdout Cross-Validation, Leave-One-Out Cross-Validation, K-Fold. Cross-Validation, the performance of each machine learning algorithm was evaluated, obtaining MSE, RMSE, MAE, R^2 for each algorithm [10]. By comparing the performance of the five machine learning models in the three types of Cross-Validation, the two best models were selected for predicting real and forecast data for the year 2022. Fig. 1 illustrates the overall methodology of our research paper, focusing on the crude marriage rate. Ridge Lasso, Elastic Net, Polynomial, Weighted Least Squares, and Holdout Cross Validation, Leave-One-Out Cross-Validation, K-Fold Cross Validation are used to make comprehensive decisions.

B. Data Collection

In this section, we provide detailed explanations of the primary factors influencing the crude marriage rate, including information on data collection and processing. Data on the crude marriage rate and economic indicators are sourced from the National Bureau of Statistics of China (https://data.stats.gov.cn/english/) and the China Statistical Yearbook (https://www.stats.gov.cn/sj/ndsj/). These datasets are updated annually on the National Bureau of Statistics of China's website, covering the years 2003 to 2022, resulting in a 20-year dataset.

C. Data Merging

Throughout this study, the source file contains multiple individual Excel data sheets spanning from 2003 to 2020. Due to the small dataset, manual data merging was carried out to process the combined data, resulting in the creation of two separate files in Excel and .csv formats. The first file was organized based on data from different provinces between 2003 and 2022, categorized by year. The second file, on the other hand, was arranged by province, covering the same time frame.

D. Data Pre-Processing

Scikit-learn, a Python library for machine learning, was utilized to combine and preprocess data from various independent CSV files in the research. The missing values in certain columns were handled through the Simple Imputer tool available in the SciPy library, ensuring the integrity of the data .

E. Data Selection

The decision After merging and preprocessing the data, we decided to select the data column features to include the 31 provinces in mainland China, the years are 2003-2022, the dependent variable y is Crude marriage rate, and the

independent variables are (X_1-X_7). Table I offers an explanation of the arrangement of the newly screened features.

TABLE I. FEATURES DATE SELECTION

Features	Features Explain
Region	31 provinces in mainland China（No data from Hong Kong, Macao and Taiwan）
Year	2003-2022
Crude_marriage_rate（Y）	Registered Marriages couples/Resident Population/2
GDP (X_1)	Gross Regional Product (100 million yuan)
House_ Prices (X_2)	Average Selling Price of Commercialized Residential Buildings (yuan / square meters)
Gross_Dependency_Ratio (X_3)	Gross Dependency Ratio (Sample Survey) (%)
Birth_Rate (X_4)	Birth Rate (%)
Female (X_5)	Female Population Aged 15 and Over (Sample Survey) (person)
Average_years_of_education (X_6)	Average years of education per capita
Sex_Ratio (X_7)	Sex Ratio (Female=100) (Sample Survey) (female=100)

F. Feature Selection

As shown in Fig.2, there is a heat map of a crude marriage rate dataset. It is evident that, except between GDP and Birth Rate, House Prices and Gross Dependency Ratio, between Birth Rate, Average years of education and Gross Dependency Ratio, and between Average years of education and Birth Rate, are negatively correlated, while all others are positively correlated.

Fig. 2. Correlation Heatmap of Features

G. Feature Scaling

Data normalization, known as feature scaling, is a key preprocessing step in many regression-oriented machine learning models. It involves standardizing numerical attributes to a common scale. In this study, Min-Max Scaling was used to normalize attributes to a range of 0 to 1, aiming to reduce the impact of dimension variations and improve model training efficiency and reliability. As shown in Fig.2 there is raw data of a crude marriage rate datasets. Fig.3, there is a feature scaling of a crude marriage rate datasets. As shown in Fig.5, there is a feature describe of feature.

Region	Year	Crude_Ma	GDP	House_pri	Gross_	Birth_R	Female	Average_	Sex_Ratio
Beijing	2003	0.006387	5267.2	4456	27.8	5.1	6111	10.3457	106.08
Beijing	2004	0.008406	6252.5	4747.14	26.7	6.1	6256	10.5586	105.01
Beijing	2005	0.006242	7149.8	6162.13	26.7	6.29	90424	10.6858	102.65
Beijing	2006	0.010618	8387	7375.41	26.9	6.26	6554	10.9501	97.21
Beijing	2007	0.006975	10425.5	10661.24	24.7	8.32	6645	11.0853	99.06
Beijing	2008	0.008267	11813.1	11648	25	8.17	6608	10.9696	103.38
Beijing	2009	0.00971	12900.9	13224	25	8.06	6692	11.1726	104.27
Beijing	2010	0.006983	14964	17151	38.9	7.48	1938	11.0092	102.03
Beijing	2011	0.008498	17188.8	15517.9	21.3	8.29	7725	11.555	104.09

Fig. 3. Raw Data

Region	Year	Crude_M	GDP	House_Pr	Gross_Depe	Birth_Rat	Female	Average_ye	Sex_Rati
Beijing	2003	6.3874	0.039	0.07458	0.221932	0.12055	0.008	0.7306065	0.44728
Beijing	2004	8.4059	0.047	0.0808	0.193211	0.18904	0.008	0.7541443	0.41268
Beijing	2005	6.2419	0.054	0.11102	0.193211	0.20205	0.134	0.7682111	0.33635
Beijing	2006	10.618	0.064	0.13694	0.198433	0.2	0.008	0.7974373	0.16041
Beijing	2007	6.9749	0.079	0.20712	0.140992	0.3411	0.009	0.8123878	0.22025
Beijing	2008	8.2665	0.09	0.22819	0.148825	0.33082	0.008	0.799592	0.35996
Beijing	2009	9.7097	0.099	0.26185	0.148825	0.32329	0.009	0.8220361	0.38875
Beijing	2010	6.9827	0.115	0.34573	0.511749	0.28356	0.001	0.8039697	0.3163
Beijing	2011	8.498	0.132	0.31085	0.052219	0.33904	0.01	0.8643205	0.38292

Fig. 4. Features Scaling

	1	2	3	4	5	6	7
count	620.000000	620.000000	620.000000	620.000000	620.000000	620.000000	620.000000
mean	0.147114	0.112171	0.497402	0.530187	0.055889	0.553571	0.379543
std	0.159548	0.128166	0.175902	0.204381	0.133562	0.137955	0.120939
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.040104	0.040372	0.381201	0.389212	0.009564	0.488183	0.316300
50%	0.095086	0.081771	0.511749	0.554110	0.019711	0.556310	0.361902
75%	0.194760	0.136058	0.613577	0.675342	0.036913	0.624929	0.442594
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 5. Features Describe

H. Evaluation Model

The leave-one-out method is simple and easy to use, but the data is not fully divided. K-fold cross-validation makes full use of data and is suitable for large data sets. The leave-one-out method LOOCV has high computational cost and is suitable for small data sets [10]. Each method selects the most suitable cross-validation method based on the data size and computing resources.

*1) Holdout Cross Validation:*The dataset is randomly split into two disjoint subsets: one as training set and one as test set. The model is trained on the training set and the performance is evaluated on the test set.*LOO Cross Validation(LCV):*Mean squared error (MSE) is a common measure of the quality of an estimator, such as a machine learning model. It calculates the average squared difference between the predicted values and the actual values [10]. A lower MSE value indicates a better fit of the model to the data.

2) Leave-One-Out Cross-Validation（LOOCV）: Each time, one sample is left out from the dataset as the test set, and the remaining samples are used as the training set. This process is repeated for each sample [10]. LOOCV is suitable for small datasets because the computational cost is high.

*3) K-Fold Cross Validation:*Divide the dataset into K subsets (folds) of equal size. Use K-1 subsets for training each time and the remaining subset for testing. Repeat this process K times, using a different subset as the test set each

time [10]. The final evaluation result is the average of the K test results.

I. Evaluation Model

The regression model evaluation indicators MSE,RMSE, MAE,R^2 indicators are mainly used to evaluate the prediction error rate and model performance in regression analysis.

*1) Mean Squared Error (MSE):*Mean squared error (MSE) is a common measure of the quality of an estimator, such as a machine learning model. It calculates the average squared difference between the predicted values and the actual values [18]. A lower MSE value indicates a better fit of the model to the data . The formula for the MSE is (8).

$$MSE = \frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y})^2 \tag{8}$$

*2) Root mean squared error (RMSE):*Root mean squared error (RMSE) is the square root of the mean squared error (MSE). It is another common measure of the quality of an estimator, and it represents the average error in the predictions . The formula for the RMSE is (9).

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y})^2} \tag{9}$$

*3) Mean absolute error (MAE):*Mean absolute error (MAE) is another measure of the quality of an estimator, and it calculates the average of the absolute differences between the predicted values and the actual values. A lower MAE value indicates a better fit of the model to the data [19]. The formula for the MAE is (10).

$$MAE = \frac{1}{N}\Sigma_{i=1}^{N}|y_i - \hat{y}| \tag{10}$$

*4) R-Squared:*R-squared (R^2), also known as the coefficient of determination, is a statistical measure that indicates the proportion of the variance in the dependent variable (y) that is predictable from the independent variables (X) in a regression model. It ranges from 0 to 1, with a higher value indicating a better fit of the model to the data [19]. The formula for the R^2 is (11).

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} \tag{11}$$

IV. Results and Discussion

A. Numerical Features Bxplot

Fig.6 shows that The boxplot analysis of marriage rate in China reveals various key features including GDP, housing prices, fertility, gender ratio, female population, years of education, and male-female gender ratio. Specifically, the results reveal the following: GDP values span from 0 to 0.6, primarily falling between 0.2 and 0.4, with a few notable exceptions indicating significant GDP disparities between provinces. Housing prices vary from 0 to 0.8, showing a broad distribution and numerous anomalies, highlighting substantial differences in housing prices across provinces, including some with exceptionally high prices. Fertility rates are concentrated mostly between 0.2 and 0.8, with significant outliers suggesting notable differences among provinces. Gender ratios range from 0.2 to 1.0, showing a relatively tight concentration without clear outliers, indicating minor variations among provinces. Female population values range from 0 to 1.0, with the majority falling between 0.4 and 0.8, showcasing discrepancies in female population levels across provinces. Years of education range from 0.6 to 1.0, with a concentrated distribution and no significant outliers, suggesting minimal differences in education levels among provinces, with most having higher education rates. Male-female gender ratio values span from 0.4 to 1.0 with a symmetrical distribution, implying uniformity in male-female gender ratios among provinces. Analyzing these metrics offers crucial insights for forecasting marriage rates in China using Ridge and polynomial regression models, assessing model performance through various cross-validation techniques such as Holdout, LOOCV, and K-Fold CV to ensure accurate predictions and enhance comprehension of marriage rate trends.

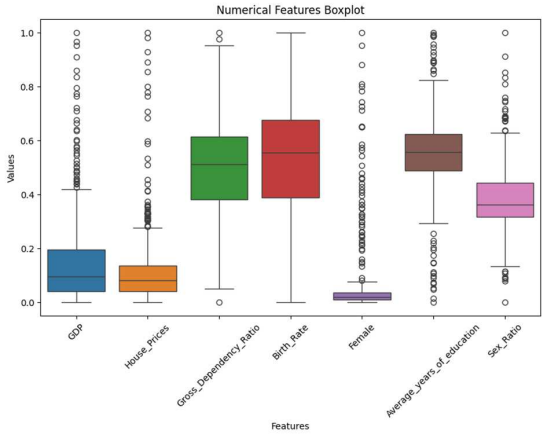

Fig. 6. Numerical Features Boxplot

B. Cross-Validation Evaluation Results

*1) Holdout CV Results:*Table II displays the evaluation results of five regression models in the context of Holdout CV. Through comparison based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), the analysis reveals the following: The Ridge regression demonstrates a good performance with an MSE of 3.565, RMSE of 1.888, MAE of 1.465, and an R^2 value of 0.166, indicating a certain level of predictive capability. Both Lasso and Elastic Net exhibit subpar performance on this dataset, as reflected by their MSE of 4.327, RMSE and MAE of 2.080 and 1.680 respectively, and an R^2 value of -0.011. Polynomial regression performs the best with the lowest MSE of 3.473, RMSE of 1.863, MAE of 1.468, and an R^2 value of 0.188. Weighted Least Squares method follows with an MSE of 3.683, RMSE of 1.919, MAE of 1.481, and an R^2 value of 0.139, ranking below Polynomial and Ridge regression. In summary, Polynomial regression demonstrates the best performance across all metrics, followed by Ridge regression,

while Lasso and Elastic Net show poor performance on this dataset.

2) LOOCV Results: Table III displays the evaluation results of five regression models in the context of LOOCV CV. The LOOCV results for five regression models reveal varying performances. Polynomial Regression achieves the best outcomes, with an MSE of 3.001, RMSE of 1.388, and MAE of 1.388, indicating its superior predictive capability. Ridge Regression follows, showing solid performance with an MSE of 3.326, RMSE of 1.452, and MAE of 1.452. Weighted Least Squares performs moderately well, with an MSE of 3.403, RMSE of 1.520, and MAE of 1.520. In contrast, both Lasso and Elastic Net exhibit poor performance, each having an MSE of 4.278, RMSE of 1.664, and MAE of 1.664, suggesting they are not well-suited for this dataset. R^2 tests all values in LOOCV to be NaN.

3) K-Fold CV: Table IV displays the evaluation results of five regression models in the context of *K-Fold CV.*The K-Fold CV results for five regression models illustrate their performance variability. Polynomial Regression stands out with the best results, having an MSE of 3.018, RMSE of 1.733, MAE of 1.392, and R^2 of 0.285, indicating superior predictive accuracy. Ridge Regression follows with an MSE of 3.334, RMSE of 1.822, MAE of 1.452, and R^2 of 0.210, showing solid performance. Weighted Least Squares performs moderately well with an MSE of 3.423, RMSE of 1.846, MAE of 1.524, and R^2 of 0.189. In contrast, both Lasso and Elastic Net perform poorly, each having an MSE of 4.280, RMSE of 2.064, MAE of 1.665, and R^2 of -0.011, indicating they are unsuitable for this dataset. Therefore, Polynomial Regression is the most efficient model, with Ridge Regression coming next, while Lasso and Elastic Net fall behind considerably.

TABLE II. HOLDOUT CV EVALUATION RESULTS OF 5 MODELS

Model Regression	Holdout CV Results			
	MSE	***RMSE***	***MAE***	***R²***
Ridge	3.565	1.888	1.465	0.166
Lasso	4.327	2.080	1.680	-0.011
Elastic Net	4.327	2.080	1.680	-0.011
Polynomial	3.473	1.863	1.468	0.188
Weighted Least Squares	3.683	1.919	1.481	0.139

TABLE III. LOOCV EVALUATION RESULTS OF 5 MODELS

Model Regression	LOOCV Results			
	MSE	***RMSE***	***MAE***	***R²***
Ridge	3.326	1.452	1.452	nan
Lasso	4.278	1.664	1.664	nan
Elastic Net	4.278	1.664	1.664	nan
Polynomial	3.001	1.388	1.388	nan
Weighted Least Squares	3.403	1.520	1.520	nan

TABLE IV. K-FOLD CV EVALUATION RESULTS OF 5 MODELS

Model Regression	K-Fold CV Results			
	MSE	***RMSE***	***MAE***	***R²***
Ridge	3.334	1.822	1.452	0.210
Lasso	4.280	2.064	1.665	-0.011
Elastic Net	4.280	2.064	1.665	-0.011
Polynomial	3.018	1.733	1.392	0.285
Weighted Least Squares	3.423	1.846	1.524	0.189

C. Model Results

Table V compares the Mean Squared Error (MSE) of five regression models across three cross-validation methods. Ridge regression performs well in LOOCV and K-Fold CV (3.326 and 3.334) but is average in Holdout CV (3.565). Lasso and Elastic Net perform poorly in all methods (around 4.3). Polynomial regression shows the best performance in all methods, especially in LOOCV and K-Fold CV (3.001 and 3.018). Weighted Least Squares is stable across methods but slightly inferior to Polynomial regression (3.409 to 3.399). Overall, Polynomial regression is the most optimal across all methods.Fig.7 presents a detailed comparison of the actual and predicted crude marriage rates for 2022 across various regions in China using five regression models. Here is an in-depth analysis of each model's performance:

*1) Actual Values (blue line):*The actual values serve as a benchmark for evaluating the predictive accuracy of each model. Most regions have actual crude marriage rates between 5 and 7, with outliers like Ningxia and Tibet showing lower rates

*2) Ridge Regression (orange line):*Ridge Regression predictions closely align with actual values in most regions, indicating stable performance. Although there are slight deviations in regions like Beijing and Hebei, the overall trend matches well with actual values, especially in central and eastern regions. Ridge Regression predictions are generally slightly lower than the actual values, but the deviations are minimal.

3) Lasso Regression (red line): Lasso Regression demonstrates large prediction discrepancies compared to actual values in various regions like Beijing, Hebei, and Guangdong. The strong feature selection in Lasso may cause inaccuracies in specific regions, resulting in significant deviations in predictions.

*4) Elastic Net Regression (yellow line):*Polynomial Regression outperforms other models, with predictions closely matching actual values in most regions. Especially in Beijing, Shanghai, and Guangdong, the predictions almost overlap with actual values, demonstrating excellent performance. Polynomial Regression effectively captures non-linear relationships in the data, providing high-accuracy predictions.

*5) Polynomial Regression (purple line):*Weighted Least Squares also provides relatively accurate predictions. Despite minor deviations in regions like Ningxia and Hainan, the

model performs well overall, with predictions close to actual values, particularly in central and eastern regions. This method handles heteroscedasticity well, contributing to its robust performance.

In summary, Polynomial Regression and Ridge Regression deliver the best performance in predicting the 2022 crude marriage rates across China's regions, with predictions closely matching actual values. Lasso and Elastic Net Regression models show poorer performance, requiring further adjustment and optimization to improve their predictive accuracy. Weighted Least Squares also performs well and is a reliable choice in most regions.

TABLE V. COMPARISON DIFFERENT CROSS-VALIDATION RESULTS

Model Regression	Comparison Different Cross-Validation Results		
	Holdout CV MSE	***LOOCV MSE***	***K-Fold CV MSE***
Ridge	3.565	3.326	3.334
Lasso	4.327	4.278	4.280
Elastic Net	4.327	4.278	4.280
Polynomial	3.473	3.001	3.018
Weighted Least Squares	3.683	3.409	3.399

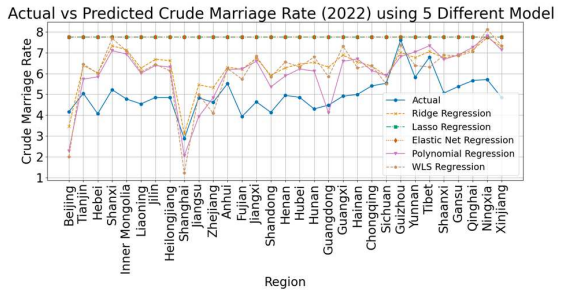

Fig. 7. Actual vs Predicted Crude Marriage Rate(2022) using Models

D. Best Models Prediction

Table VI and The Fig.8 illustrates the comparison of actual crude marriage rates in 2022 across various regions in China with predictions using the best regression models: Polynomial Regression and Ridge Regression. Here is a detailed analysis.

*1) Actual Values (Blue Line):*The actual crude marriage rates serve as the benchmark. Most regions have actual rates between 5 and 7, with lower rates observed in Ningxia and Tibet.

*2) Polynomial Regression (Purple Line):*Predictions from Polynomial Regression closely match actual values in most regions. Particularly in Beijing, Tianjin, Shanghai, and Guangdong, the predicted values almost overlap with the actual values, indicating high accuracy. In regions like Hebei, Inner Mongolia, Heilongjiang, Jilin, Sichuan, and Gansu, predictions are slightly higher than actual values, but the deviations are minimal.

*3) Ridge Regression Predictions (Green Line):*Ridge Regression predictions are also close to actual values, though with slight deviations in some regions. For example, predictions are slightly lower than actual values in Beijing, Tianjin, and Hebei, while slightly higher in Jilin, Shanghai, Henan, and Xinjiang. Overall, Ridge Regression shows stable performance with minimal deviations from actual values in most regions.

4) Comparison and Summary: Polynomial Regression shows better performance in predicting crude marriage rates across regions in China for 2022, effectively capturing complex patterns for highly accurate predictions. Ridge Regression also performs well, with slightly higher deviations but good stability and robustness. In regions with strong nonlinearity, Polynomial Regression outperforms Ridge Regression. Overall, Polynomial Regression excels in predictive accuracy, with both models being effective tools for this prediction task.

TABLE VI. COMPARISON DIFFERENT CROSS-VALIDATION RESULTS

Model Regression	Comparison Different Cross-Validation Results		
	Holdout CV MSE	***LOOCV MSE***	***K-Fold CV MSE***
Ridge	3.565	3.326	3.334
Polynomial	3.473	3.001	3.018

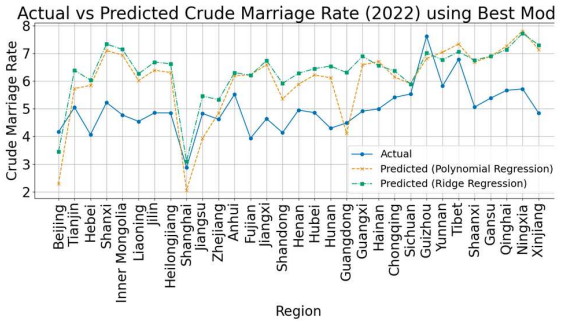

Fig. 8. Actual vs Predicted Crude Marriage Rate(2022) using Best Models

E. Discussion

The cross-validation analysis of five regression models—Ridge, Lasso, Elastic Net, Polynomial, and Weighted Least Squares (WLS)—reveals distinct patterns in predictive accuracy and stability, essential for predicting crude marriage rates across China's regions.

1) Holdout CV Results:The actual The holdout cross-validation results, shown in Table II, demonstrate that Polynomial Regression outperforms other models with an MSE of 3.473, RMSE of 1.863, MAE of 1.468, and an R^2 value of 0.188. Ridge Regression follows, showing commendable performance with an MSE of 3.565, RMSE of 1.888, MAE of 1.465, and R^2 value of 0.166. Both Lasso and Elastic Net regressions exhibit poor performance, indicated by their higher MSE values (4.327) and negative R^2 values (-0.011), suggesting these models are not suitable for this dataset. The WLS method, while stable, ranks below Polynomial and Ridge Regression with an MSE of 3.683 and R^2 value of 0.139.

*2) Actual Values (blue line):*The LOOCV analysis confirms the superiority of Polynomial Regression, which achieves the lowest MSE (3.001) and RMSE (1.388),

demonstrating its robust predictive capability. Ridge Regression, with an MSE of 3.326 and RMSE of 1.452, also shows strong performance. WLS performs adequately with an MSE of 3.403. Lasso and Elastic Net again perform poorly, with both models presenting identical MSE (4.278) and RMSE (1.664), further emphasizing their inadequacy for this task.

3) Actual Values (blue line): K-Fold CV results mirror previous findings, with Polynomial Regression leading in accuracy (MSE of 3.018 and R^2 of 0.285). Ridge Regression maintains its second position with an MSE of 3.334 and R^2 of 0.210, reflecting consistent performance. WLS, with an MSE of 3.423, performs moderately well but lags behind Polynomial and Ridge Regression. High MSE values for Lasso and Elastic Net (4.280) and negative R^2 values (-0.011) reiterate their unsuitability for the dataset.

Fig.6 and Fig.7 provide detailed comparisons of actual vs. I made crude marriage rate predictions using top-performing models Polynomial and Ridge Regression. Polynomial Regression consistently aligns closest with actual values, especially in nonlinear regions. Ridge Regression also performs accurately but with slightly higher deviations. The strong performance of these models highlights their potential in predicting complex data patterns. In conclusion, Polynomial Regression proves to be the most effective model overall, followed closely by Ridge Regression. Lasso and Elastic Net regressions need further optimization for this dataset. WLS shows stable but slightly inferior performance compared to Polynomial and Ridge Regression.

V. CONCLUSION

This study evaluates the predictive performance of five regression models—Ridge, Lasso, Elastic Net, Polynomial, and Weighted Least Squares—using various cross-validation methods (Holdout, LOOCV, and K-Fold CV). The results highlight Polynomial Regression as the most accurate model, consistently outperforming others with the lowest MSE and RMSE values across all cross-validation methods. Ridge Regression also shows robust performance, albeit with slightly higher deviations. The analysis reveals that Lasso and Elastic Net are unsuitable for this dataset due to their high MSE and negative R^2 values. WLS, while stable, ranks below Polynomial and Ridge Regression. The visual comparison of actual vs. predicted crude marriage rates confirms these findings, with Polynomial and Ridge Regression providing the closest alignment to actual values. This study underscores the importance of selecting appropriate models and cross-validation methods for accurate predictions, contributing to a better understanding and forecasting of marriage rate trends in China.

ACKNOWLEDGMENT

This research was supported by a grant from Sirindhorn International Institute of Technology, Thammasat University.

REFERENCES

[1] D. H. Wrenn, J. Yi, and B. Zhang, “House prices and marriage entry in China,” Regional Science and Urban Economics, vol. 74, pp. 118–130, Jan.2019

[2] J. C. Yong, N. P. Li, P. K. Jonason, and Y. W. Tan, “East Asian low marriage and birth rates: The role of life history strategy, culture, and social status affordance,” *Personality and Individual Differences*, vol. 141, pp. 127–132, Apr. 2019

[3] C. Zhao, B. Chen, and X. Li, “Rising housing prices and marriage delays in China: Evidence from the urban land transaction policy,” *Cities*, vol. 135, p. 104214, Apr. 2023

[4] G. Chiplunkar and J. Weaver, “Marriage markets and the rise of dowry in India,” *Journal of Development Economics*, vol. 164, p. 103115, Sep. 2023

[5] G. Nie, “Marriage squeeze, marriage age and the household savings rate in China,” *Journal of Development Economics*, vol. 147, p. 102558,Nov.2020.

[6] J. Chen and W. Pan, “Bride price and gender role in rural China,” *Heliyon*, vol. 9, no. 1, p. e12789, Jan. 2023

[7] M. Porter, “How do sex ratios in China influence marriage decisions and intra-household resource allocation?,” *Review of Economics of the Household*, vol. 14, no. 2, pp. 337–371, Aug. 2014

[8] M. R. Putri, I. G. P. S. Wijaya, F. P. A. Praja, A. Hadi and F. Hamami, "The Comparison Study of Regression Models (Multiple Linear Regression, Ridge, Lasso, Random Forest, and Polynomial Regression) for House Price Prediction in West Nusa Tenggara," 2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), Bali, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICADEIS58666.2023.10270916.

[9] X. Yang, Z. Yin and J. Li, "Housing Price Mathematical Prediction Method through Big Data Analysis and Improved Linear Regression Model," 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 2021, pp. 751-754, doi: 10.1109/TOCS53301.2021.9688839.

[10] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 83-87, doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[11] R. P and S. M, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1416-1421, doi: 10.1109/ICICCS51141.2021.9432109.

[12] S. Sanyal, S. Kumar Biswas, D. Das, M. Chakraborty and B. Purkayastha, "Boston House Price Prediction Using Regression Models," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6, doi: 10.1109/CONIT55038.2022.9848309.

[13] United Nations. “Crude Marriage Rate 1 definition,” https://data.un.org/Glossary.aspx?q=crude+marriage+rate+,2001.

[14] Shanghai Survey Team of National Bureau of Statistics, “Average years of education per capita,” https://tjj.sh.gov.cn/zcjd/20091102/0014-86153.html, November 2009.

[15] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani, "House Price Prediction using Machine Learning Algorithm," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 982-986.

[16] A. Chaurasia and I. U. Haq, "Housing Price Prediction Model Using Machine Learning," 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, pp. 497-500.

[17] R. Guo, H. Xing and H. Wang, "Medium And Long Term Load Forecasting Based On Improved Partial Least Square Method," 2020 International Conference on Smart Grids and Energy Systems (SGES), Perth, Australia, 2020, pp. 679-684, doi: 10.1109/SGES51519.2020.00126.

[18] R. Gupta, A. Sharma, V. Anand and S. Gupta, "Automobile Price Prediction using Regression Models," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 410-416, doi: 10.1109/ICICT54344.2022.9850657.

[19] S. Shaprapawad, P. Borugadda and N. Koshika, "Car Price Prediction:An Application of Machine Learning," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 242-248, doi: 10.1109/ICICT57646.2023.101.