

Forecasting and Analyzing the Drivers of Rural Residents' Income Using Stacking Ensemble Learning: A Case Study of Zhuji City

Jiale He^{1,*}, Chuanming Ma², Shiya Pan³

¹School of Economics, Yunnan University, Kunming, China

²School of Law, Yunnan University, Kunming, China

³Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China

*Corresponding author: 20241040074@stu.ynu.edu.cn

Abstract. Forecasting rural income is vital for China's national strategies, yet traditional models falter on short, multi-feature time-series data, creating an accuracy-interpretability dilemma. This paper introduces a Stacking ensemble learning framework using 2012-2021 economic data from Zhuji City to address this. The framework integrates Ridge, SVR, and Random Forest base learners, with a Linear Regression meta-learner to optimally weigh their predictions. Empirical results demonstrate two key findings: 1) The Stacking model achieves significantly superior predictive accuracy over its base learners, proving its robustness on small-sample data. 2) An analysis of the meta-learner weights reveals that rural income growth follows a composite pattern, dominated by a linear trend and fine-tuned by nonlinear corrections. This research offers a robust forecasting paradigm for regional economies and a novel, interpretable method for analyzing the driving mechanisms of complex economic phenomena.

Keywords: Ensemble Learning; Stacking Model; Time Series Forecasting; Rural Residents' Income; Driving Factor Analysis.

1. Introduction

China's national strategies of Rural Revitalization and Common Prosperity set the agenda for its current economic and social development. Rural Revitalization aims to modernize the agricultural sector and holistically develop rural areas. This strategy is a foundational pathway to achieving the broader goal of Common Prosperity, which prioritizes shared development outcomes for all citizens. Within this framework, increasing farmers' income is a critical link, serving as a key metric for the success of both initiatives.

The academic community has produced a substantial body of research on the measurement and forecasting of residents' income. Liu Jiong et al. (2024) utilized the ARIMA model to forecast and analyze the urban-rural income gap in Anhui Province from 1980 to 2020, achieving good predictive results [1]. Li Yanfang (2024) employed a Grey Prediction Model to forecast the per capita income of rural residents in Hunan Province and analyzed its influencing factors [2]. Wang Yan (2019) utilized Ridge Regression to analyze the impact of the added value of the three industries on rural residents' income in Sichuan Province, finding a diversity in income sources [3]. Qian Li (2014) applied Quadratic Exponential Smoothing and ARMA models to forecast inter-provincial and inter-regional gaps in rural residents' income [4]. Chen Yushan (2023) utilized a coupled ARIMA-GM model to forecast the per capita disposable income of rural residents in China, finding that the hybrid model outperformed single models by overcoming potential long-term prediction deviations and thus enhancing accuracy [5][6].

In summary, the models prevalently adopted in this research area exhibit a common trade-off where predictive accuracy and model interpretability are difficult to achieve simultaneously, and there is a relative scarcity of studies that employ ensemble models to address this issue. Therefore, this paper aims to integrate the strengths of multiple models within a Stacking ensemble learning

framework to propose a viable ensemble learning approach for analyzing economic data characterized by short time series and multiple features.

2. Methodology

2.1. Data Sources and Indicator Selection

The objective of this study is to construct a hybrid forecasting model with high precision and practical utility to conduct an in-depth analysis of the driving factors of rural residents' income in Zhuji City, Zhejiang Province. The data for this study are sourced from the *Statistical Bulletin on the National Economic and Social Development of Zhuji City* and the *Monitoring Report on the Economic and Social Development of Zhuji City*. The sample period covers the years from 2012 to 2021. The specific target vector and feature variables are presented in Table 2-1.

Table 1. Definitions and Explanations of Research Variables

Variable Type	Variable Name	Unit
Dependent Variable	Per Capita Disposable Income of Rural Residents	CNY
Feature Variable	Gross Domestic Product	100 Million CNY
	Tertiary Industry Ratio	%
	General Public Budget Expenditure	100 Million CNY
	Per Capita Disposable Income of Urban Residents	CNY

3. Stacking Ensemble Learning Model

This study employs representative non-ensemble models, specifically Ridge Regression and Support Vector Regression (SVR), as base learners. A Stacking ensemble model is subsequently constructed using a linear regression algorithm as the meta-learner. This framework integrates the predictions from the individual models, and its predictive accuracy and performance are then compared against those of the standalone base learners to validate its superiority.

3.1. Base Learners

The performance of a Stacking framework is highly contingent upon the selection of its base learners. Accordingly, this study employs Ridge Regression (Ridge), Support Vector Regression (SVR), and Random Forest (RF) as base learners. Each model was chosen for its distinct capabilities: Ridge Regression is a modification of traditional linear regression designed to address multicollinearity [8]; Support Vector Regression (SVR), derived from the classic Support Vector Machine algorithm, offers unique advantages in handling nonlinear problems by solving a convex optimization problem [9]; and Random Forest (RF) is a decision tree-based algorithm highly effective at revealing complex interaction effects among variables [10][11]. The use of these heterogeneous models ensures that information is captured from multiple perspectives, providing the upper-level meta-learner with a multi-dimensional and diverse basis for its decision-making [7].

3.2. Meta Learners

The primary function of the meta-learner is to learn how to intelligently organize and assign weights to the prediction results of the base learners [7]. For the M base learners in the first layer, the prediction vector obtained for the i -th sample in the training set is given by $z_i = [\hat{y}_{(i,1)}, \hat{y}_{(i,2)}, \dots, \hat{y}_{(i,m)}]$, where $\hat{y}_{(i,j)}$ is the prediction of the j -th base learner for the i -th sample. Subsequently, z_i is used as the new input features (i.e., meta-features) for the meta-learner. In this study, a linear regression algorithm is employed as the meta-learner, and its predictive function is shown in Equation (1).

$$\hat{y}_{\text{meta}} = \theta_0 + \sum_{j=1}^M \theta_j \hat{y}_j \quad (1)$$

Here, θ_0 is the intercept term, and θ_j is the weight assigned to the j -th base learner.

3.3. The Process of Model Conclusion

3.3.1. Data Processing

Since the SVR and Ridge models employed in this study are sensitive to the scale of the data, and this research utilizes multi-dimensional data inputs, the data is standardized prior to training. First, the mean μ_j and standard deviation σ_j of each feature variable are calculated using the training set. Subsequently, every feature value x_{ij} is transformed using Z-Score standardization.

3.3.2. Generating Meta-Features via Cross-Validation

Given the small size of the dataset in this study, Leave-One-Out Cross-Validation (LOOCV) is employed to mitigate the risk of overfitting. The number of base learners is set to K , denoted as M_1 , M_2 , and M_3 . The process consists of N iterations.

Given the small size of the dataset in this study, Leave-One-Out Cross-Validation (LOOCV) is employed to mitigate the risk of overfitting. The number of base learners is set to K , denoted as M_1 , M_2 , and M_3 , and the process consists of N iterations. In the i -th iteration, each base learner is trained on the training subset $D_{\text{train}}^{(-i)}$, which excludes the i -th sample. The formal expression for this process is shown in Equation (2).

$$D_{\text{train}}^{(-i)} = D_{\text{train}} \setminus (x_i, y_i) \quad (2)$$

After each base learner is trained on the subset $D_{\text{train}}^{(-i)}$, a temporary model, $M_k^{(-i)}$, is obtained. This model is then used to generate a prediction, z_{ik} , for the held-out test sample. The formula for this calculation is presented in Equation (3).

$$z_{ik} = M_k^{(-i)}(x_i) \quad (3)$$

Upon completion of these N iterations, the input feature matrix for the meta-learner, Z_{Train} , is obtained.

3.3.3. Meta-Learner Training

Based on the meta-feature matrix Z_{Train} generated in the previous step and the original target vector Y_{Train} , this study employs a linear regression model for further training. The objective is to obtain the optimal weights by minimizing a loss function, which is presented in Equation (4).

$$\min_{\theta} \sum_{i=1}^N \left(y_i - \left(\theta_0 + \sum_{k=1}^K \theta_k z_{ik} \right) \right)^2 \quad (4)$$

Through this subsequent training, the meta-learner, M_{Meta} , is finalized.

3.3.4. Final Prediction

First, we train the base learners on the complete training set, D_{Train} , to obtain the final, fully trained base models. These trained models are then used to make predictions for the test set, generating the meta-features $z_{\text{new}} = (M_1^*(x_{\text{new}}), M_2^*(x_{\text{new}}), \dots, M_K^*(x_{\text{new}}))$. Subsequently, these meta-

features are fed as input into the trained meta-learner to obtain the final predicted value, $\hat{y}_{new}$. The calculation process is formalized in Equation (5).

$$\hat{y}_{new} = M_{meta}(z_{new}) = \theta_0 + \sum_{k=1}^K \theta_k z_{new,k} \quad (5)$$

The overall framework of the model is illustrated in Figure 3-3.

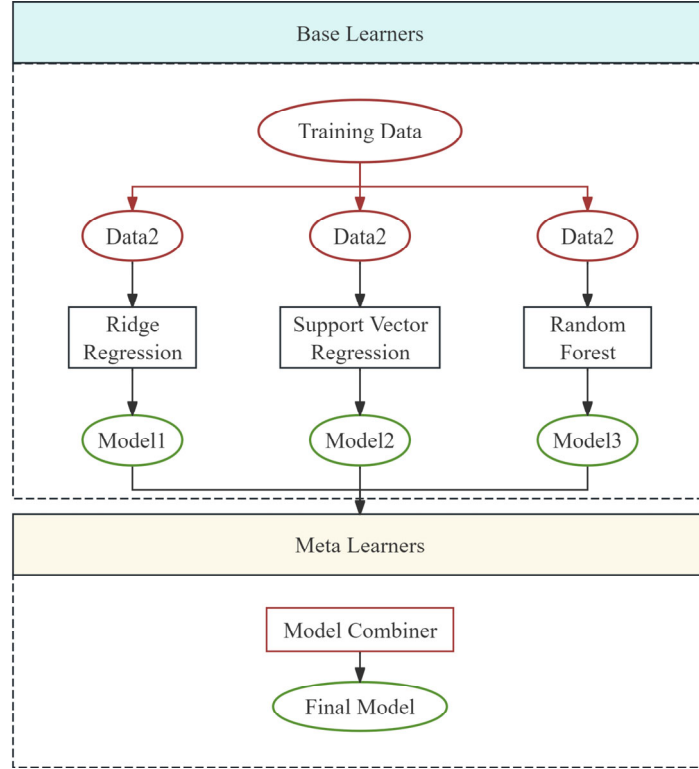


Figure 1. Overall Framework of Model Construction

3.4. Model evaluation metrics

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are two of the most common metrics for evaluating model performance in the field of machine learning. Therefore, this study employs these indicators to assess the accuracy and precision of the proposed Stacking ensemble learning model.

RMSE is particularly sensitive to large errors, making it a good indicator of the model's performance on outliers. In contrast, MAE reflects the average magnitude of the prediction errors. The combined use of these two metrics, therefore, offers a comprehensive evaluation of the model's overall predictive capability.

4. Results

4.1. Analysis of Model Performance Comparison

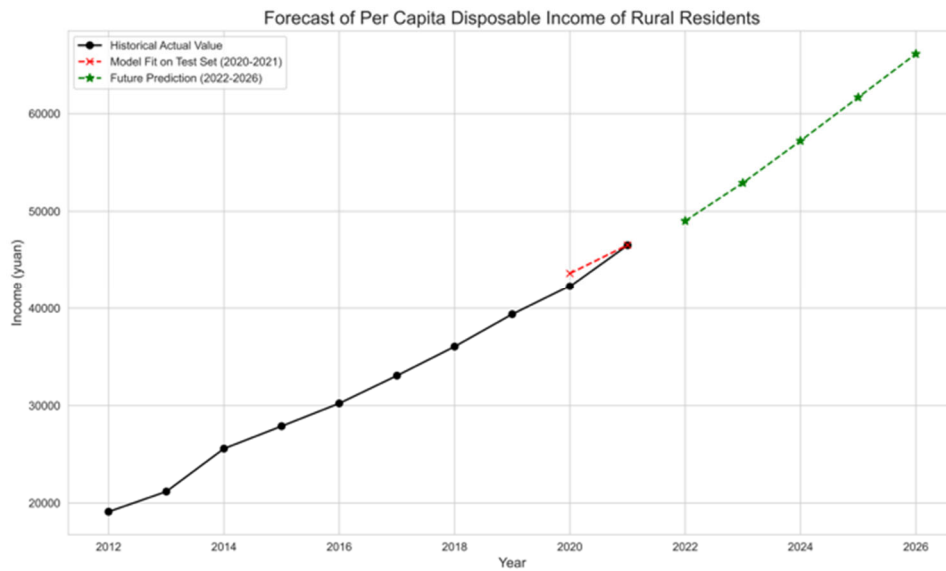
To evaluate the effectiveness of the Stacking ensemble learning model, this study conducts a comparative analysis of the RMSE and MAE values between the ensemble model and the three individual models. The results are presented in Table 4-1.

Table 2. Comparison of RMSE and MAE

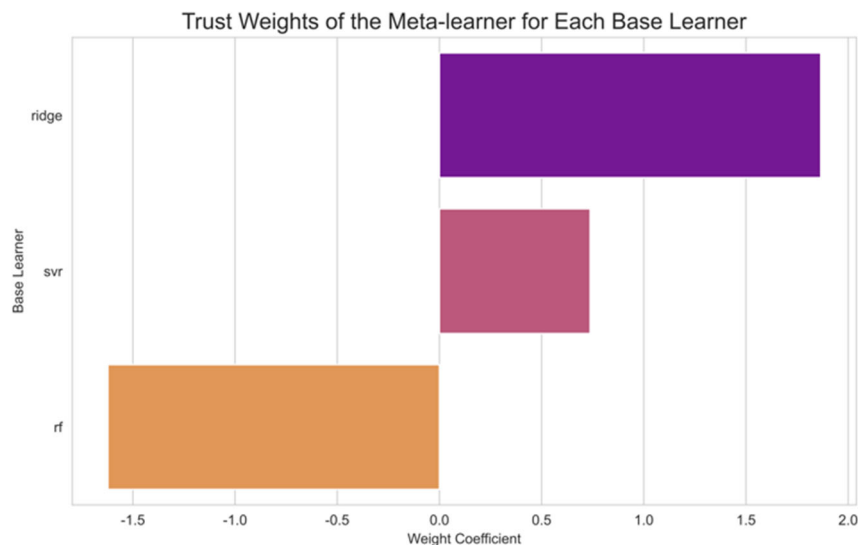
Model	RMSE	MAE
Stacking	956.06	687.95
Ridge Regression	978.29	957.25
Support Vector Regression	12178.74	11781.19
Random Froest	6926.80	6594.56

The quantitative results presented in Table 4-1 further substantiate this, showing that the proposed Stacking model integrating Ridge, SVR, and RF achieves lower RMSE and MAE values of 956.06 and 687.95, respectively, which signifies a smaller prediction error.

Consequently, the proposed Stacking model, which integrates Ridge, SVR, and RF, is subsequently used to forecast the per capita income of rural residents in Zhuji City for the next five years. The results of this forecast are presented in Figure 4-1.

**Figure 2.** Forecast of Per Capita Disposable Income of Rural Residents

4.2. Analysis of Internal Mechanism of the Stacking

**Figure 3.** Trust Weights of the Meta-Learner for Each Base Learner

Analysis of the meta-learner's weights reveals a clear division of labor. The Ridge model receives a dominant weight (1.87), confirming that a macroscopic linear trend is the primary driver of rural income growth. Most notably, the Random Forest (RF) model is assigned a significant negative weight (-1.62). This indicates the meta-learner uses the RF's nonlinear predictions not as a direct forecast, but as a corrective term. It effectively subtracts the systematic deviations captured by RF from the primary linear forecast, performing a fine-grained calibration that enhances overall accuracy.

5. Conclusions

The Stacking ensemble learning model offers a dual advantage: it integrates the strengths of multiple models to achieve superior predictive accuracy, while its architecture allows for the interpretation of underlying economic mechanisms by analyzing the meta-learner's weights, thus enhancing model explainability. The primary conclusions of this research are as follows:**1)Effectiveness of the Stacking Ensemble Learning Model.**The empirical results demonstrate that the Stacking ensemble model achieves significantly lower prediction error than any of its single base learners. This validates its effectiveness and offers a new research paradigm for predictive tasks on small-sample datasets.**2)Intrinsic Mechanism of Rural Resident Income Growth in Zhuji City.**Through an analysis of the meta-learner's weight distribution, this study deduces that the growth of rural residents' income in Zhuji City follows a composite pattern: "dominated by a linear trend and corrected by nonlinear perturbations.

References

- [1] J. Liu, M. Zhou, and Y. Wu, "Predictive analysis of urban-rural resident income gap based on ARMA model: A case study of Anhui Province," *Journal of Jiangsu Shipping College*, vol. 23, no. 1, pp. 89-95, 2024.
- [2] Y. Li, "Forecast of per capita disposable income of rural residents in Hunan Province based on grey prediction model," *Economic Research Guide*, no. 5, pp. 32-34, 2024.
- [3] Y. Wang, "Analysis of influencing factors of rural residents' income in Sichuan Province based on ridge regression," *Journal of Aba Teachers University*, vol. 36, no. 2, pp. 35-40, 2019.
- [4] L. Qian, "Forecasting the development trend of regional disparities in rural residents' income: An analysis based on quadratic exponential smoothing and ARMA models," *Journal of Central University of Finance and Economics*, no. 7, pp. 78-82, 2014.
- [5] Y. Chen, "Forecasting research on per capita disposable income of rural residents in China based on ARIMA-GM coupled model," *Rural Economy and Science-Technology*, vol. 34, no. 19, pp. 221-224, 2023.
- [6] P. Cao and C. Li, "Forecast and realization path of doubling rural residents' income: A case study of Hebei Province," *Economic Research Reference*, no. 63, pp. 79-86, 2015.
- [7] J. Xu and Y. Yang, "Ensemble learning methods: A research review," *Journal of Yunnan University (Natural Sciences Edition)*, vol. 40, no. 6, pp. 1082-1092, 2018.
- [8] N. Yang, "The unique role of ridge regression analysis in solving the problem of multicollinearity," *Statistics & Decision*, no. 3, pp. 14-15, 2004.
- [9] D. Sun, "Research on support vector machine for classification and regression," Ph.D. dissertation, Central South University, 2004.
- [10] K. Fang, J. Wu, J. Zhu, et al., "A review of random forest methods," *Statistics & Information Forum*, vol. 26, no. 3, pp. 32-38, 2011.
- [11] X. Li, "Application of random forest model in classification and regression analysis," *Chinese Journal of Applied Entomology*, vol. 50, no. 4, pp. 1190-1197, 2013.