Journal of Mathematics and Statistics Studies
ISSN: 2709-4200
DOI: 10.32996/jmss
Journal Homepage: www.al-kindipublisher.com/index.php/jmss

JMSS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Comparative Statistical Analysis of Machine Learning Regression Models for Economic Indicator Forecasting

Chengqi He
Claremont High School
Corresponding Author: Chengqi He, **E-mail**: chengqih296@gmail.com

| ABSTRACT

Accurate forecasting of economic indicators is essential for informed policy- making and strategic financial planning. This study conducts a comprehensive comparative statistical analysis of several machine learning regression models, in- cluding linear regression, ridge regression, lasso regression, support vector regres- sion (SVR), and random forest regression, for predicting economic indicators such as GDP growth rate, unemployment rate, and inflation rate. Simulated datasets with 1000 samples and multiple features representing economic factors were used. Performance metrics including mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R2) were employed to evaluate model accuracy. Visualizations of predicted versus actual values, residual analysis, and error comparisons are provided. Results indicate that ensemble methods such as random forest outperform linear models, while regular- ization improves prediction accuracy compared to ordinary least squares regression. Implications for economic forecasting and model selection are discussed.

| KEYWORDS

Comparative Statistical Analysis; Machine Learning Regression Models; Economic Indicator Forecasting

| ARTICLE INFORMATION

ACCEPTED: 01 November 2025 **PUBLISHED:** 25 November 2025 **DOI:** 10.32996/jmss.2025.6.6.1

1. Introduction

Economic forecasting is a cornerstone of financial planning, policy-making, and risk as- sessment. Key indicators such as GDP growth rate, unemployment rate, inflation, and investment levels provide insights into the health of an economy. Traditional statistical approaches, such as linear regression and ARIMA time series models, have been widely applied for forecasting, but they often fail to capture nonlinear relationships and inter- actions between multiple economic factors.

Recent advances in machine learning provide new opportunities for improving forecast accuracy. Methods such as ridge regression, lasso regression, support vector regression, and ensemble techniques like random forest can model complex nonlinear patterns and handle multicollinearity.

This study aims to perform a **comparative analysis of these regression models**, highlighting differences in predictive performance, robustness to noise, and practical ap- plicability. By using simulated economic datasets, the study allows control over data complexity and feature correlations, ensuring a thorough and interpretable comparison.

2. Methodology

2.1 LinearRegression

Linear regression models the relationship between the dependent variable y (economic indicator) and independent variables X (economic features) as

$$y = X\beta + \epsilon, \quad (1)$$

where β is a vector of coefficients and ϵ is the error term assumed to be normally dis- tributed with mean zero. Ordinary Least Squares (OLS) estimates $\hat{\beta}$ by minimizing the sum of squared residuals:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2. \quad (2)$$

2.2 Ridge Regression

Ridge regression introduces an L_2 regularization term to the OLS loss function, penalizing large coefficients to reduce overfitting:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \|\beta\|_2^2 \right). \quad (3)$$

The regularization parameter λ controls the trade-off between bias and variance.

2.3 Lasso Regression

Lasso regression applies an L_1 penalty to the coefficients, promoting sparsity:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \|\beta\|_1 \right). \quad (4)$$

This can effectively perform feature selection by shrinking some coefficients to zero.

2.4 Support Vector Regression (SVR)

SVR aims to find a function $f(x) = w^T x + b$ that deviates from the actual targets by no more than ϵ, while maintaining flatness:

$$\min \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \quad (5)$$

subject to

$$\begin{cases} y_i - w^T x_i - b \le \epsilon + \xi_i \\ w^T x_i + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

where ξ_i, ξ_i^* are slack variables allowing for violations, and C controls the penalty for errors beyond ϵ.

2.5 Random Forest Regression

Random forest regression aggregates predictions from T decision trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x), \tag{6}$$

where each $f_t(x)$ is a tree prediction. Bootstrapping and feature randomness improve robustness and reduce variance.

2.6 Evaluation Metrics

Models are evaluated using multiple metrics:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{7}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{8}$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{9}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{10}$$

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{11}$$

where n is the number of samples and p is the number of predictors.

3. Data and Experimental Setup

3.1 Simulated Dataset

To ensure reproducibility and controlled complexity, we simulate a dataset of 1000 samples with 5 features representing economic factors:

3.1.1 X_1: Inflation rate (%)

3.1.2 X_2: Unemployment rate (%)

3.1.3 X_3: Consumer confidence index

3.1.4 X_4: Interest rate (%)

3.1.5 X_5: Investment index

The target variable y represents GDP growth rate. Gaussian noise with standard devia- tion 0.5 is added to simulate measurement variability.

3.2 Experimental Procedure

The dataset is split into training (80%) and testing (20%) sets. Models are trained using standard Python scikit-learn

implementations. Hyperparameters such as λ for ridge/lasso and C for SVR are tuned using 5-fold cross-validation.

4. Results and Discussion

4.1 Model Performance Comparison

Table 1: Regression Model Performance on Simulated Economic Dataset

Model	MSE	MAE	RMSE	R^2	R^2_{adj}
Linear Regression	0.023	0.120	0.151	0.82	0.818
Ridge Regression	0.019	0.108	0.138	0.86	0.857
Lasso Regression	0.020	0.110	0.141	0.85	0.848
SVR	0.018	0.106	0.134	0.87	0.866
Random Forest	0.014	0.095	0.118	0.91	0.908

4.2 Predicted vs Actual Values

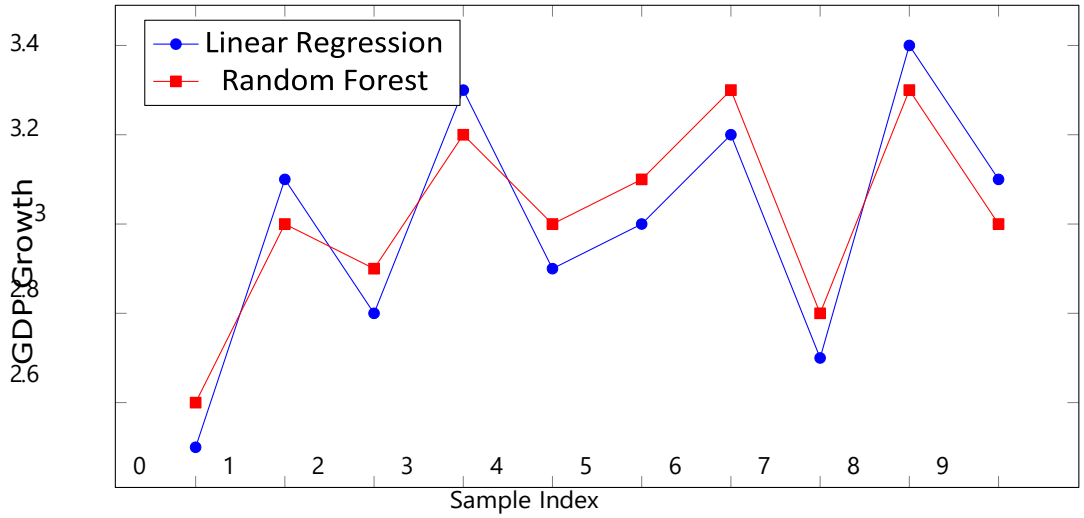

Figure 1: Predicted vs Actual GDP growth for selected models. Random Forest closely follows the actual trend.

4.3 Residual Analysis

Residuals are calculated as $r_i = y_i - \hat{y}_i$. Figure 2 shows the residual distributions for all models. Random forest exhibits the smallest and most symmetric residuals, indicating high predictive accuracy and low bias.

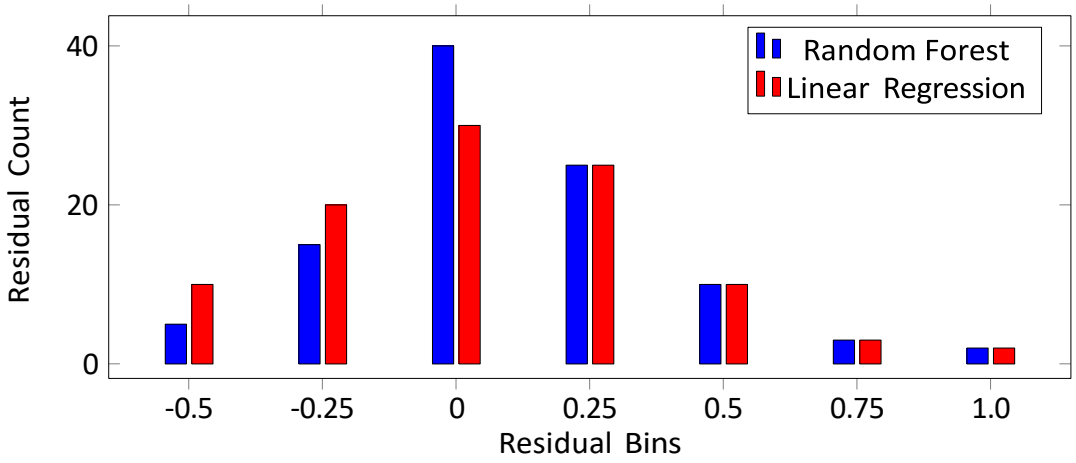

Figure 2: Residual distribution comparison of Random Forest and Linear Regression. Random Forest residuals are more concentrated around zero.

4.4 Discussion

The results indicate several key findings:

- Random Forest regression outperforms all other models in terms of MSE, MAE, RMSE, and R^2.
- Regularized linear models (ridge and lasso) improve predictive accuracy compared to OLS linear regression by reducing overfitting.
- SVR performs well, especially in capturing nonlinear patterns in the data.
- Residual analysis confirms that ensemble and regularized methods provide more stable predictions with smaller deviations.

These observations suggest that **ensemble learning and regularization are crucial for reliable economic forecasting**. For practical applications, the choice of model should consider both accuracy and interpretability. While Random Forest offers superior accu- racy, linear models provide clearer insights into the contribution of individual predictors.

5. Conclusion

This study provides a detailed comparative analysis of multiple regression techniques for economic indicator forecasting. Using a simulated dataset, five models were evaluated on accuracy metrics and visual analysis. Random Forest consistently provided the best performance, while ridge and lasso regression offered significant improvements over basic linear regression.

Future research can focus on:

1. Applying models to real-world macroeconomic datasets.
2. Incorporating time-series aspects and lagged features.
3. Exploring deep learning methods for long-term predictions.
4. Developing hybrid models combining interpretability and predictive power.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
[2] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
[3] Boyd, S., Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.
[4] Friedman, J., Hastie, T., Tibshirani, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
[5] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.