Journal of Economic Theory and Business Management
ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online) | Vol. 1, No. 2, 2024

Scenarios Driving Economic Forecasts: Choosing Econometrics or Machine Learning

XIA, Shaoliang [1*] WU, Zhen [2] SONG, Qingqing [1]

[1] Belarusian State University, Belarus
[2] Nanjing University of Aeronautics and Astronautics, China

**** XIA, Shaoliang is the corresponding author, E-mail: fpm.sya@bsu.by***

Abstract: Economic forecasting is an in-depth understanding of the laws of economic operation. It is based on historical facts and past data to make possible inferences about future economic trends and development trends in related fields. Since the main operating rules of economic activities are different and contain randomness, economic forecast is not a deterministic forecast, but a possibility forecast. Therefore, there are various economic forecasting methods, and different forecasting methods are suitable for different application scenarios. This paper aims to explore the applicability of econometric forecasting methods and machine learning methods in different application scenarios in scenario-driven economic forecasting. For macroeconomic forecasting, the ARIMA model and VAR vector autoregressive model can process time series data and are suitable for predicting macroeconomic indicators. For industry economic forecasting, random forest regression and Adaboost regression can handle large amounts of data and complex data structures, and are suitable for predicting the development trends of specific industries. For product economic forecasting, linear regression and decision trees are capable of handling continuous-valued forecasts and are suitable for forecasting demand and price for a specific product or service. For prediction of the impact of major events, double difference DID (difference-in-difference method) can avoid the endogeneity problem of data to a large extent and is suitable for evaluating policy effects.

Keywords: Economic forecasting, Machine learning, Econometrics, Economic forecasting scenarios, Time series models, Advanced regression models

DOI: https://doi.org/10.5281/zenodo.10927145

1. Introduction

Economic forecasting plays a crucial role in our daily life and decision-making. It can not only depict the possible future directions of the economy and society, allowing people to respond to the future, but also predict the most likely scenarios in the future, thereby guiding current decisions. [1] At the government and enterprise levels, macroeconomic forecasting is an important basis for policy-making, planning, and business decision-making. Through the correction of economic forecasts, it is possible to calibrate expectations in the economy and guide decision-making adjustments. [2] Scientific economic forecasting can improve the economic benefits of enterprises and avoid or reduce predictable risks. At the household level, mastering certain macroeconomic forecasting knowledge can help households better arrange their consumption, savings, and investments. Economic forecasting is not a deterministic prediction, but a possible prediction. We cannot be certain that the future economy will develop in a specific way, but we can predict the likelihood and possible direction of future economic development. [3]

Due to the complexity and uncertainty of economic activities, there are various methods for economic forecasting. [4] These methods include but are not limited to time series analysis, regression analysis, machine learning, etc. Different prediction methods are suitable for different application scenarios. For example, time series analysis is commonly used to predict macroeconomic indicators (such as GDP, inflation rate, etc.); Regression analysis is commonly used to study the relationship between economic variables; Machine learning has advantages in handling big data and complex data structures. The purpose of this paper is to explore the applicability of factors such as prediction objectives, data characteristics, and model assumptions in scenario driven economic forecasting, in order to ensure the accuracy and reliability of prediction results.

2 The Concept and Importance of Economic Forecasting

Economic forecasting, as a scientific exploration, relies on survey and statistical data, data generated from social and economic activities, and information on the expected future prospects of different economic actors. This method cleverly integrates the changes in historical data of the economy, using scientific data analysis methods in fields such as mathematics and statistics to deeply study the operational

changes of corresponding objects, and accurately quantify the development prospects of future changes. The concept is to make full use of existing information and scientific methods to predict future economic activities, in order to gain a deeper understanding and respond to future economic changes. Its goal is to reduce the probability of similar negative scenarios occurring in history, depict the possible future direction of the economy and society, and predict the most likely future scenarios, thereby providing guidance for current decision-making. At the same time, economic forecasts can also be calibrated for exceeding expectations in the economy through correction, further guiding decision-making adjustments. Overall, economic forecasting is aimed at meeting people's needs, improving economic efficiency, avoiding or reducing predictable risks, and better arranging consumption, savings, and investment to achieve sustained and healthy economic development. [5]

3 Economic Forecasting Scenarios

According to the scope of activities, economic forecasts are mainly classified into macroeconomic forecasts, industry economic forecasts, product economic forecasts, and major event impact forecasts:

1. Macroeconomic forecasting: This is a prediction of the economic activity of the entire economic system or specific regions, including the prediction of macroeconomic indicators such as GDP, unemployment rate, inflation rate, etc.

2. Industry economic prediction: This is a prediction of the economic activities and development trends of a specific industry, such as predicting the trends of the automotive industry or the real estate market in the coming years.

3. Product economic prediction: This is a prediction of the demand, price, and other economic indicators for a specific product or service, such as predicting the sales volume of a certain car or the market demand for a certain service within the next year.

4. Prediction of the impact of major events: This is the prediction of the impact of specific events (such as policy changes, technological innovation, natural disasters, etc.) on economic activities, such as the prediction of the impact of new environmental protection policies on the oil industry, or the prediction of the impact of the COVID-19 epidemic on the global economy

4 The Difference Between Econometrics and Machine Learning

Econometrics and machine learning are both important methods for data analysis, but their goals and application methods differ. The main goal of econometrics is causal inference, which infers the causal effect of xi on yi. In other words, we ultimately want to know which independent variables will have an impact on the dependent variable and how much impact it will have. Econometrics usually starts from theory and constructs models to verify or refute the theory. The methods of econometrics usually include regression analysis, time series analysis, etc., which are mainly used to test economic theory. The main goal of machine learning is prediction, that is, predicting yi based on xi. Machine learning typically starts from data, constructs models that provide solutions through cross validation of data, and continuously adjusts them until satisfactory prediction accuracy is achieved. The prediction methods of machine learning include linear regression, decision tree, random forest regression, Adaboost regression, etc., which are mainly used for prediction. [6,7]

5 Introduction to Commonly Used Predictive Models in Econometrics and Machine Learning

5.1 Econometric Models

The econometric model is a method that combines economic theory and statistical methods to quantitatively analyze economic phenomena. It is the main method of empirical analysis in modern economics and can specifically derive empirical relationships between some real-world variables. These models can help us gain a deeper understanding and explanation of economic phenomena. [8]

5.1.1 Time Series Model

Prediction is the judgment of the outcome of future events, or the judgment of the expected outcome of a certain event under certain important influencing factors. Therefore, time series prediction method is a quantitative prediction method that uses variable indicator data arranged in chronological order and applies mathematical statistical methods. This includes ARIMA models, GARCH models, and VAR vector autoregressive models to process various complex time series data and make accurate predictions. [9,10]

(A) ARIMA Model

The ARIMA model, also known as the Autoregressive Integrated Moving Average Model, is a commonly used method for time series analysis and prediction. The ARIMA model combines the concepts of autoregressive model (AR), difference model (I), and moving average model (MA) to predict future values by linearly combining historical observations and lagged error terms. The ARIMA model is widely used in fields such as economics, finance, and meteorology to model and predict time series data. Especially suitable for analyzing and predicting volatility, such analysis can play a very important guiding role in investor decision-making. The general form of the ARIMA model is as follows:

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \quad (5\text{-}1)$$

Equation (5-1) $X_t$represents the observed values of the time series at time t, αi and βj is the model parameter, εt is the random error term.

The following experiment uses the World Bank Open Data from The World Bank dataset on China's GDP (in US dollars) from 1960 to 2022 for prediction, which includes time (year) and corresponding GDP (in US dollars) values for each year

Analysis steps:

(1)The ARIMA model requires the sequence to satisfy stationarity. Check the ADF test results and analyze whether it can significantly reject the assumption of sequence instability based on the analysis of t-values (P<0.05).

Table1 ADF Inspection Form

variable	Differential order	t	P	AIC	critical value 1%	5%	10%
GDP (USD)	0	0.207	0.973	2841.637	-3.563	-2.919	-2.597
	1	2.101	0.999	2785.402	-3.563	-2.919	-2.597
	2	-2.253	0.188	2728.06	-3.571	-2.923	-2.599

Table 1 shows the results of ADF test, including variables, difference order, T-test results, AIC values, etc., used to test whether the time series is stationary.

The model requires that the sequence must be a stationary time series data. By analyzing the t-value, examine whether it can significantly reject the null hypothesis of sequence instability.

If it shows significance (P<0.05), it indicates rejection of the null hypothesis and that the sequence is a stationary time series. Conversely, if it is not significant, it indicates that the sequence is an non-stationary time series.

The comparison between the statistical values and ADF Test results of different degrees of rejection of the original hypothesis at critical values of 1%, 5%, and 10%, where the ADF Test result is less than 1%, 5%, and 10%, indicates a very good rejection of the hypothesis.

Differential order: Essentially, it is the next numerical value. Subtracting the previous numerical value is mainly to eliminate some fluctuations and make the data tend to be stationary. Non stationary sequences can be transformed into stationary sequences through differential transformation.

AIC value: A standard for measuring the goodness of fit of statistical models, with smaller values being better.

Critical value: A critical value is a fixed value corresponding to a given level of significance.

(2)Check the comparison chart of data before and after the difference to determine if it is stable (with small fluctuations), and perform partial (autocorrelation analysis) on the time series. Estimate its p and q values based on the truncation situation.

GDP (USD)

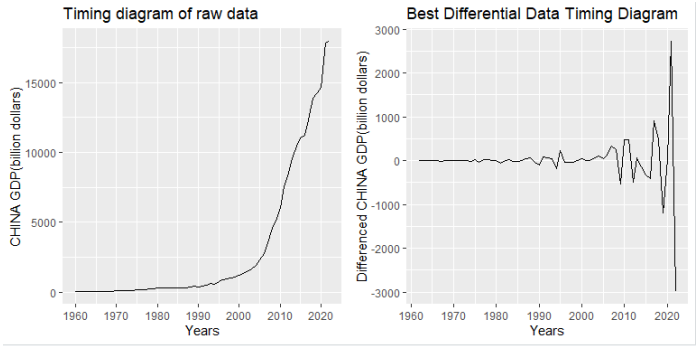

Figure 1 Differential sequence diagram

Figure 1 shows the timing chart of the 0-order difference and the best difference of the original data.

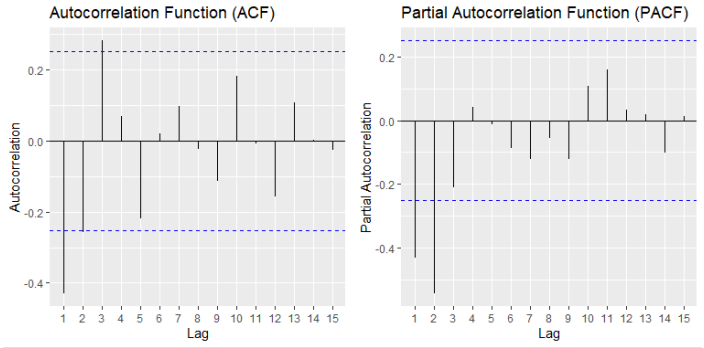

Figure 2 Differential data autocorrelation and differential data partial autocorrelation graphs

Figure 2 includes coefficients, upper and lower confidence limits. The textual explanation is as follows:

The horizontal axis represents the number of delays, and the vertical axis represents the autocorrelation coefficient.

The autocorrelation (ACF) graph is truncated in the qth order, while the partial autocorrelation (PACF) graph is trailing. The ARMA model can be simplified into the MA (q) model.

Partial autocorrelation (PACF) plots are truncated at the p-th order, while autocorrelation (ACF) plots are trailing. The ARMA model can be simplified into an AR (P) model.

Table 2 ARIMA Model Test Table

item	symbol	value
	Df Residuals	57
Number of samples	N	63
Q statistics	Q6(P value)	0.01(0.919)
	Q12(P value)	3.267(0.775)

	Q18(P value)	5.438(0.942)
	Q24(P value)	7.268(0.988)
	Q30(P value)	7.901(0.999)
information guidelines	AIC	3413.168
	BIC	3423.722
goodness of fit	R^2	0.996

Table 2 shows the results of the model validation, including sample size, degrees of freedom, Q-statistics, and the goodness of fit of the information criterion model.

The ARIMA model requires that the residual of the model does not have autocorrelation, that is, the model residual is white noise. Check the model validation table and test the white noise of the model based on the P-value of the Q-statistic (P value greater than 0.1 is white noise).

According to the information criteria, AIC and BIC values are used for multiple model comparisons (lower is better).

R^2 The degree of fitting representing the time series, the closer it is to 1, the better the effect.

The model result is Table 2, based on the variable: GDP (US dollars). From the analysis of the Q-statistic results, it can be concluded that Q6 does not show significance at the horizontal level, and the assumption that the residual of the model is a white noise sequence cannot be rejected. At the same time, the goodness of fit of the model R^2 At 0.996, the model performs well and basically meets the requirements.

(3)The ARIMA model requires the model to have pure randomness, that is, the residual of the model is white noise. Check the model validation table and test the white noise of the model based on the P-value of the Q-statistic (P>0.05). It can also be analyzed by combining the information criterion AIC and BIC values (the lower the better), or by analyzing the residual ACF/PACF graph of the model. Based on the model parameter table, obtain the model formula combined with the time series analysis graph for comprehensive analysis, and obtain the order result of backward prediction.

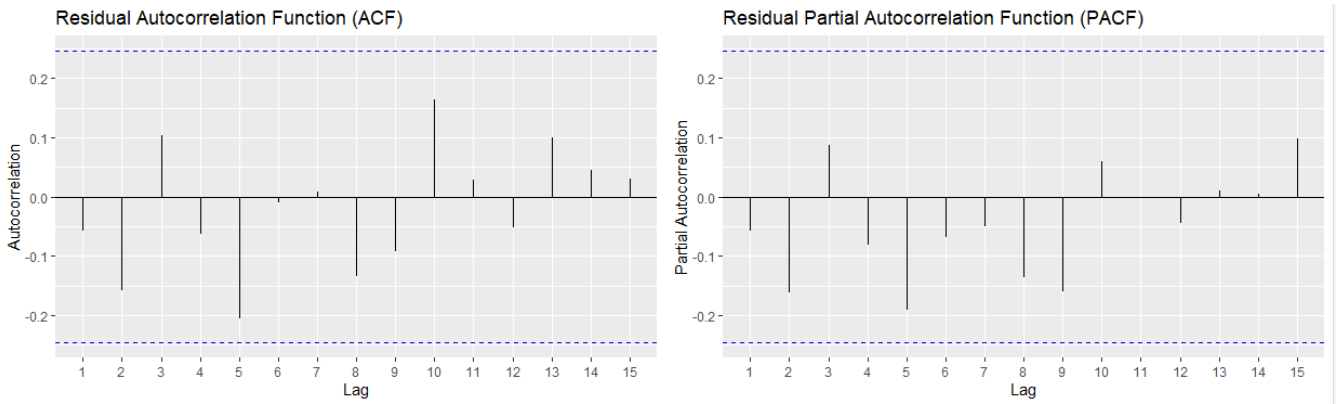

Figure3 Residual autocorrelation and residual partial autocorrelation graphs

Figure3 Including coefficients, upper and lower confidence limits

The horizontal axis represents the number of delays, and the vertical axis represents the autocorrelation coefficient.

If the correlation coefficients are all within the dashed line, the autoregressive model (AR) residual is a white noise sequence, and the time series requires the model residual to be a white noise sequence.

Table 3 Model Parameter Table

	coefficient	standard deviation	t	P>|t|	0.025	0.975
constant	52092727085.026	0	8.005250277364784e+23	0	52092727085.026	52092727085.026
ar.L1	-0.784	0.084	-9.353	0	-0.948	-0.619
ar.L2	-1	0.076	-13.117	0	-1.149	-0.851
ma.L1	-0.142	0.145	-0.981	0.327	-0.426	0.142
sigma2	9.378168853605432e+22	0	7.72409746255837e+46	0	9.378168853605432e+22	9.378168853605432e+22

Table 3 shows the results of the model parameters, including coefficients, standard deviation, T-test results, etc., for analyzing the model formulas. The model results are presented in Table 2, and the model formula is as follows:

$$y(t) = 52092727085.026 - 0.784 * y(t-1) - 1.0 * y(t-2) - 0.142 * \varepsilon(t-1)$$

Table 4 Predicted Value Table

Order (time)	forecast result
1	17872817318719.344
2	21007246455808.28

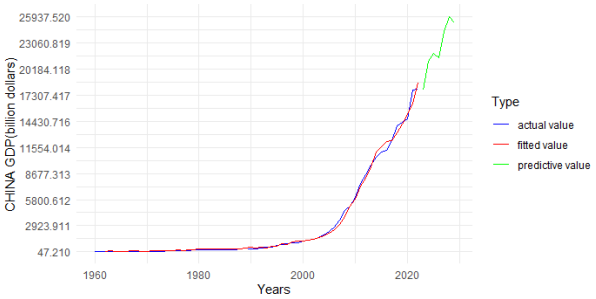

Figure4 time series diagram

Figure 4 shows the original data graph, model fitting values, and model prediction values of the time series model.

(B) GARCH Model

The GARCH model, also known as the Generalized Autoregressive Conditional Heteroscedasticity Model, is a method used for analyzing and predicting time series data. The GARCH model not only considers the autoregressive term, but also introduces the moving average term, which can better describe high-order ARCH processes. The GARCH model is widely used in fields such as economics and finance to model and predict time series data. Especially suitable for analyzing and predicting volatility, such analysis can play a very important guiding role in investor decision-making.

The general form of the GARCH model is as follows:

$$\sigma_t^2 = a_0 + a_1\varepsilon_{t-1}^2 + \ldots + a_q\varepsilon_{t-q}^2 + b_1\sigma_{t-1}^2 + \ldots + b_p\sigma_{t-p}^2 \quad (5\text{-}2)$$

Equation (5-2) σ_t^2 represents the conditional variance of the time series at time t, εt represents the residual term of the model, and ai and bj are model parameters.

The following experiment uses World Bank Open Data from The World Bank to predict China's inflation (annual inflation rate) measured by consumer price index from 1987 to 2022. The dataset includes time (year) and corresponding annual inflation rates for each year.

Analysis steps:

(1)Before establishing the GARCH model, it is necessary to perform stationarity tests on time series variables.

Table 5 Stationarity Check Table

variable	t	P	critical value		
			1%	5%	10%
annual deflation rate	-6.497	0.000***	-3.689	-2.972	-2.625

Note: *** represents a significance level of 5%

Table 5 shows the results of the ADF test, including variables, statistics, P-values, etc., used to test whether the sequence is stationary. The significance P-value is 0.000 * * *, showing significance at the horizontal level. Rejecting the null hypothesis, this series is a stationary time series.

(2)If the time series is a stationary series, the ARCH effect test can continue.

Table 6 ARCH Effect Test Table

Lag order (1-12)	X^2
1	11.531
2	17.393
3	14.618
4	16.516
5	21.865
6	22.645
7	23.519
8	35.101
9	24.619
10	11.554
11	14.258
12	17.389

Note: *** represents a significance level of 5%

Table 6 shows the results of the ARCH-LM Lagrange multiplier test (the system only takes the ARCH test with a lag of 1-12 orders).

(3)If there is an ARCH effect in the time series, a GARCH model can be established to estimate the parameters of the GARCH model.

Table 7 Model Parameter Estimation Results Table

	estimated value	standard error	Statistics	P	maximum likelihood value	AIC value
Constant	3.303	1.614	2.047	0.041**		
RESID(-1)^2	1	0.303	3.306	0.001***	-105.09	5.949

Note: *** represents a significance level of 5%

Table 7 presents the parameter estimation results of the GARCH model, which are used to test the model's fit. A stable GARCH model needs to meet the following requirements: the values of each parameter must be greater than zero; The sum of all parameters for the RESID term (i.e. ARCH term) and GARCH term must be less than 1.

(4)Perform a pure randomness test on the standardized residual sequence of the GARCH model. If it satisfies pure randomness, it indicates that the GARCH model is effective.

Table 8 Standard Residual Pure Randomness Test Table

Lag order	LB statistics	P
1	3.623	0.057*
2	3.846	0.146
3	7.043	0.071*
4	7.433	0.115
5	9.65	0.086*
6	10.112	0.120
7	10.735	0.151
8	14.927	0.061*
9	17.585	0.040**
10	19.041	0.040**
11	20.686	0.037**
12	20.729	0.054*

Note: *** represents a significance level of 5%

Table 8 presents the pure randomness test results of the GARCH model's standard residuals, which are used to test the effectiveness of the GARCH model.

(5)Finally, create a volatility chart to visually demonstrate the GARCH model's fitting of the original sequence's volatility characteristics.

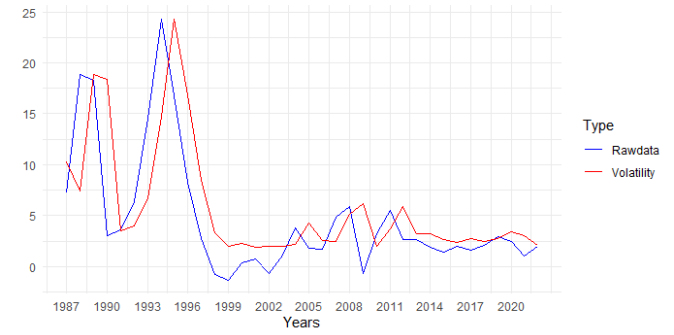

Figure5 Volatility chart

Figure 5 shows the conditional variance plot of the original sequence, intuitively demonstrating the GARCH model's fitting of the fluctuation characteristics of the original sequence.

(C) VAR Vector Autoregressive Model

The Vector Autoregressive Model (VAR model) is a commonly used econometric model. [13] The VAR model constructs the model by treating each endogenous variable in the system as a function of the lag value of all endogenous variables in the system, thereby extending the single variable autoregressive model to a vector autoregressive model composed of multiple time series variables. The VAR model is commonly used to predict interconnected time series systems and analyze the dynamic impact of random disturbances on variable systems. It is mainly applied in macroeconomics and is one of the easiest models to handle the analysis and prediction of multiple related economic indicators.

The VAR model can be represented as:

$$\mathbf{Y}_t = \mathbf{c} + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{A}_2\mathbf{Y}_{t-2} + \cdots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{u}_t \quad (5\text{-}3)$$

The following experiment uses Scientific Platform Serving for Statistics Professional to provide numerical data on manufacturing, agriculture, and tourism industries in a certain region.

Analysis steps:

(1)Before establishing the VAR model, it is necessary to perform stationarity tests on each time series variable. If all time series are stationary, a VAR model can be established; Otherwise, the obtained vector autoregressive model is a pseudo regression. If each data does not satisfy stationarity but passes the cointegration test, a vector autoregressive model can also be established.

Table 9 ADF inspection form

variable	t	P	critical value		
			1%	5%	10%
manufacturing	-4.658	0.000***	-3.501	-2.892	-2.583
agriculture	-3.966	0.002***	-3.499	-2.892	-2.583
tourism	-4.627	0.000***	-3.501	-2.892	-2.583

Note: *** represents a significance level of 5%

Table 9 shows the results of ADF test, including variables, T-test results, AIC values, etc., used to test whether the time series is stationary.

(2)Comparison of different lag orders. (Based on the results of various information criteria for different lag orders, find an optimal lag order and then rebuild the VAR model).

Table 10 Comparison table for different lag orders

Lag order	logL	AIC	SC	HQ	FPE
0	-502.011	1.587	1.665	1.618	4.887
1	-426.42	0.343	0.658*	0.471*	1.41
2	-410.329	0.289	0.843	0.513	1.336
3	-393.824	0.225	1.021	0.547	1.255
4	-374.25	0.096*	1.138	0.517	1.106*
5	-361.294	0.103	1.393	0.624	1.119
6	-349.879	0.143	1.686	0.766	1.174
7	-341.042	0.24	2.037	0.966	1.307
8	-330.017	0.291	2.347	1.121	1.395
9	-313.565	0.224	2.542	1.159	1.331
10	-307.996	0.397	2.981	1.439	1.625
11	-300.932	0.541	3.393	1.691	1.941

Table 10 shows the information criteria for vector autoregressive models with p-order lag, used to select the optimal lag order. Including logL, FPE, AIC, SC, HQ, where logL participates in the calculation of FPE, AIC, SC, HQ, and ultimately evaluates the indicators of FPE, AIC,

SC, and HQ. There are two rules for selecting the optimal lag order:

If there is the highest * for a certain lag order, it is recommended to select that lag order to establish a VAR model.

If there are orders with the same number of *, then choose the smallest possible order.

Based on the results of the four evaluation indicators FPE, AIC, SC, and HQ, it is recommended to choose the lag order as 1, which is to establish the VAR (1) model.

(3)Establish a VAR model and estimate the parameters

Table 11 Model parameter estimation table

parameter	estimator	manufacturing	agriculture	tourism
manufacturing(-1)	coefficient	-1.066	-0.435	-0.607
	standard deviation	0.507	0.242	0.348
	t	-2.102	-1.796	-1.741
manufacturing(-2)	coefficient	-0.422	-0.325	-0.128
	standard deviation	0.475	0.227	0.326
	t	-0.889	-1.431	-0.391
agriculture(-1)	coefficient	0.265	0.511	0.155
	standard deviation	0.263	0.126	0.181
	t	1.006	4.062	0.855
agriculture(-2)	coefficient	0.086	0.272	0.076
	standard deviation	0.273	0.131	0.188
	t	0.317	2.084	0.405
tourism(-1)	coefficient	2.452	0.677	1.447
	standard deviation	0.72	0.344	0.495
	t	3.406	1.968	2.925
tourism(-2)	coefficient	0.708	0.387	0.274
	standard deviation	0.709	0.339	0.487
	t	0.999	1.143	0.563
constant	coefficient	0.738	0.256	0.573
	standard deviation	0.379	0.181	0.26
	t	1.946	1.412	2.198

Table 11 shows the parameter estimation results of the VAR model. The following is the calculation process:

$$manufacturin = -1.066 * manufacturing(-1) - 0.422 * manufacturing(-2) + 0.265 * agriculture(-1) + 0.086 * agriculture(-2) + 2.452 * tourism(-1) + 0.708 * tourism(-2) + 0.738$$

$$agriculture = -0.435 * manufacturing(-1) - 0.325 * manufacturing(-2) + 0.511 * agriculture(-1) + 0.272 * agriculture(-2) + 0.677 * tourism(-1) + 0.387 * tourism(-2) + 0.256$$

$$tourism = -0.607 * manufacturing(-1) - 0.128 * manufacturing(-2) + 0.155 * agriculture(-1) + 0.076 * agriculture(-2) + 1.447 * tourism(-1) + 0.274 * tourism(-2) + 0.573$$

(4)After establishing the VAR model, stability testing of the model is required. Pulse response analysis and variance decomposition can only be performed after passing the test

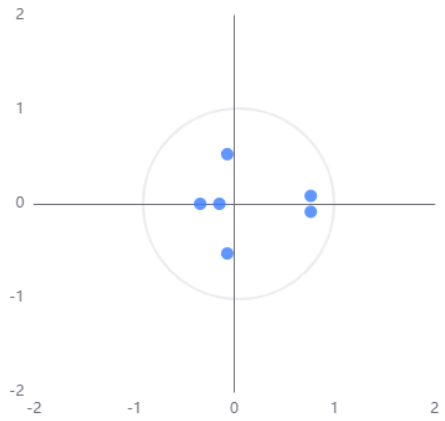

Figure 6 VAR Model Stability Test Chart

Figure 6 shows the AR root graph in the VAR model. If all points are within the unit circle, it can be determined that the VAR system is stable, and the model can further perform impulse response analysis and variance decomposition.

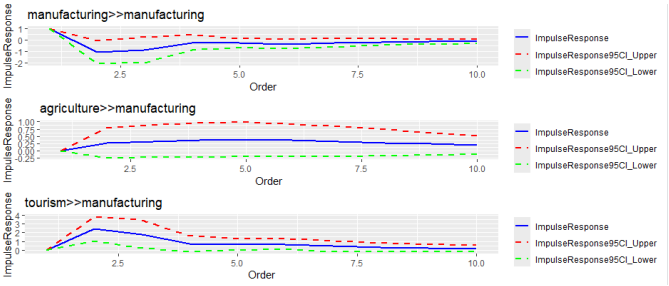

Figure 7 Pulse response analysis chart

Figure 7 shows the pulse response analysis graph. It describes the impact of an endogenous variable (shock variable) on another endogenous variable (affected variable) in the VAR model. For example, assuming that all endogenous variables in a vector autoregressive system have a positive impact on the manufacturing industry, it is a reflection of the manufacturing industry:

Manufacturing ->Manufacturing, the impact of manufacturing on itself appeared in the first five periods, and has gradually converged to zero since the sixth period.

Agriculture ->manufacturing industry, the response of agriculture to manufacturing industry has changed from negative to positive, and there is no significant convergence

in the 10th period, indicating that agriculture has had a stable and lasting impact on manufacturing industry.

Tourism industry ->manufacturing industry. In the second and third phases, the impact of tourism on manufacturing industry is relatively significant, and from the fourth phase onwards, the impact on manufacturing industry gradually converges.

Table 12 Variance decomposition results table

Order	standard deviation	manufacturing%	agriculture%	tourism%
1	2.697	100	0	0
2	3.367	92.06	0.445	7.495
3	3.636	89.002	1.029	9.969
4	3.769	88.318	1.933	9.749
5	3.861	87.507	2.809	9.684
6	3.917	86.651	3.532	9.817
7	3.947	86.048	4.118	9.834
8	3.966	85.621	4.57	9.809
9	3.977	85.31	4.895	9.795
10	3.984	85.095	5.12	9.784

Table 12 shows the results of variance decomposition. Variance decomposition is the analysis of the proportion of the standard deviation of predicted residuals affected by different shocks, as well as the corresponding contribution ratio of endogenous variables to the standard deviation.

5.1.2 Advanced Regression Model

Advanced regression model is a more complex statistical model that can be used for short-term prediction and long-term pattern research, as well as dynamic analysis of economic structure. This includes robust regression and double difference DID (double difference method) models.

(A) Robust Regression

RANSAC (Random Sample Consensus) is a robust regression method [14] that randomly selects data samples and fits the model, then calculates residuals for other samples based on the model. If a sample's residuals are less than a pre-set threshold, it is considered an inlier. This method has strong robustness to outliers, so the RANSAC regression model can play an important role in dealing with economic data with outliers. In the field of economics, the RANSAC regression model can be applied to various data samples, such as real estate price data, stock market data, and macroeconomic data.

The objective function of RANSAC can be expressed as:

$$\min_{\alpha} \sum_{i=1}^{n} \rho(x_i^T \alpha - y_i) \quad (5\text{-}4)$$

In equation (5-4) ρ It is a robust loss function used to reduce the impact of outliers.

The following experiment uses Scientific Platform Serving for Statistics Professional to provide a house price dataset for prediction, which includes house price, room square, floor, room age, and supporting elevator values.

Analysis steps:

(1)Analyze the significance of X, and if it shows significance (P<0.05), use it to explore the relationship between X and Y.

Table 13 The relationship between X and Y

	Unstandardized coefficient		Standardized coefficient	t	P	R^2	Adjustment R^2	F
	B	standard error	Beta					
constant	-43.795	16.44		-2.664	0.008***			
Room square meter	2.075	0.081	0.977	25.775	0.000***			
Floor	-0.802	0.284	-0.092	-2.823	0.005***	0.664	0.636	F=23.682, P=0.000***
House age	0.759	0.558	0.034	1.359	0.174			
Supporting elevator	48.33	12.218	0.176	3.956	0.000***			
dependent variable: House price								

Note: *** represents a significance level of 5%

Table 13 shows the analysis results of this model, including the standardization coefficients, t-values, and R of the model², Adjust R^2 Wait for the verification of the model and analyze the formula of the model. The significance P-value is 0.000 * * *, showing significance at the horizontal level, indicating that the Room square meter has a significant impact on the House price, and the model formula is determined.

- Based on the variable Room square meter, the significance P-value is 0.000 * * *, which shows significance at the horizontal level, indicating that Room square meter has a significant impact on House price.
- Based on the variable Floor, the significance P-value is 0.005 * * *, which shows significance at the horizontal level, indicating that Floor has a significant impact on

House price.

- Based on the variable House age, the significance P-value is 0.174, which is not significant at the level, indicating that House age does not have a significant impact on House price.
- Based on the variable Supporting elevator, the significance P-value is 0.000 * * *, which shows significance at the horizontal level, indicating that Supporting elevator has a significant impact on House price.

The formula of the model is as follows:

$$House\ price = -43.795 + 2.075 \times Room\ square\ meter - 0.802 \times Floor + 0.759 \times House\ age + 48.33 \times Supporting\ elevator$$

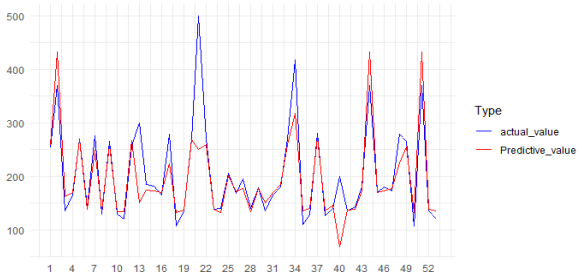

Figure 8 Robust regression result prediction chart

Figure 8 presents the raw data graph and model fitting values of this model in a visual form

Table 14 Robust regression result prediction table

variable	coefficient	test value
constant	-43.795	1
Room square meter	2.075	100
Floor	-0.802	
House age	0.759	
Supporting elevator	48.33	
forecast result:	163.70500000000004	

Table 14 is used for predicting robust regression (RANSAC), with a predicted price of RMB 1.63705 million when the house is 100 square meters

(B) Double difference DID (double difference method)

The Difference in Differences (DID) method is a method used to evaluate policy effects. The basic idea is to divide the survey sample into two groups: one group is the policy affected group, which is the experimental group; One group is the control group, which is not affected by the policy. The principle of DID is to use observational data to simulate experimental design, and to construct a double difference statistic that reflects the policy effect by comparing the differences between the control group and the treatment group before and after policy implementation. This method can largely avoid endogeneity issues in the data.

The benchmark DID model can be represented as:

$$y_{it} = \alpha + \beta \cdot \text{treated}_i + \gamma \cdot \text{after}_t + \delta \cdot (\text{treated}_i \times \text{after}_t) + \epsilon_{it} \quad (5\text{-}5)$$

Equation (5-5) y_{it} is the observed value of individual i at time t. treated_i is an indicator variable. If individual i is affected by policy implementation (i.e. belongs to the treatment group), the value is 1, otherwise it is 0. $\text{after}_t$is an indicator variable, which takes a value of 1 if it is implemented after the policy (i.e. in the later stage), and 0 if it is not. $\text{treated}_i \times \text{after}_t$is the interaction term between the processing group and the policy implementation dummy variable, and its coefficient δ Reflects the net effect of policy implementation.

The following experiment uses Scientific Platform Serving for Statistics Professional to provide GDP datasets for several large cities in China from 1990 to 1999 for prediction. The datasets include region, year, GDP changes, control variable 1, control variable 2, control variable 3, and experimental group values.

Analysis steps:

(1)Conduct descriptive statistics on the sample data based on whether the policy has an impact and whether it belongs to the control group.

Table 15 Data distribution table

	before the incident	after the incident	Summary
Control	20	20	40
Treated	15	15	30
Summary	35	35	70

Table 15 shows the distribution of data groups

(2)Display the fitting parameters of the double difference method model, such as coefficients, standard error, R^2 Wait to determine the fitting effect of the model.

Table 16 Double difference DID model result table

result		coefficient	standard error	t	P
	Control	654692288.5			
Before	Treated	2003742142.783			
	Diff (T-C)	1349049854.283	1062151115.892	1.27	0.209
	Control	2010512217.495			
After	Treated	877404803.783			
	Diff (T-C)	-1133107413.712	918323173.866	-1.234	0.222
Diff-in-Diff		-2482157267.995	1498367382.736	-1.657	0.103

Note: *** represents a significance level of 5%

Table 16 shows the results of the double difference DID (double difference method) model. The DID results showed a significant P-value of 0.103, which is not significant at the level and does not reject the null hypothesis. The coefficient is -2482157267.995, indicating that the policy intervention is ineffective and negative.

(3)The double difference method requires the assumption of parallel trends, rough estimates can be determined using time trend graphs, and accurate estimates can be determined using event study methods.

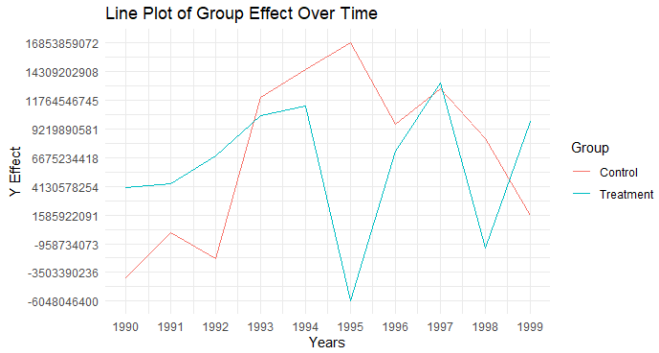

Figure9 Time trend chart of experimental group and control group

Figure 9 shows the time trend graphs of the experimental group and the control group, with the y-axis representing the Y effect value and the x-axis representing the time point at which the event occurred.

Table 17 The results of the regression of the interaction term

interaction term	coefficient	standard error	t
const	3871104	941401687.405	0.004
time	1623615216	1331343033.969	1.22
Before-3.0	617680394.167	2196603937.279	0.281
Before-2.0	-1771465629.333	2196603937.279	-0.806
Before-1.0	-2124945986.667	2196603937.279	-0.967
current	-8465170520	2196603937.279	-3.854
After1.0	-2223262789.333	2196603937.279	-1.012
After2.0	-998066114.667	2196603937.279	-0.454
After3.0	-4749045260	2196603937.279	-2.162
1991.0	617263698.286	1331343033.969	0.464
1992.0	381804123.929	1630555552.917	0.234
1993.0	3957022387.429	1630555552.917	2.427
1994.0	4569576787.429	1630555552.917	2.802
1995.0	3544131427.429	1630555552.917	2.174
1996.0	1742299243.429	1630555552.917	1.069
1997.0	2525176291.429	1630555552.917	1.549
1998.0	1405636119.429	1630555552.917	0.862

Note: *** represents a significance level of 5%

Table 17 shows the results of regression using the interaction term between year variables and whether group variables are treated. The coefficient of the interaction term reflects the difference between the treatment group and the control group for a specific year (Before1-3 means three years before the year of the event, After1-3 means three years after the event). It is generally hoped that the coefficient before the event is not significant, indicating that the parallel trend assumption is met, and then significant again after the event, reflecting the effectiveness of the policy

Figure10 Chart of positive and negative policy effects

The content of Figure 10 is a visualization of the table content, which can more intuitively show the positive and negative effects of policies. If the 95% confidence interval (dashed in the figure) contains a value of 0, it indicates that the difference between the treatment group and the control group is not significant, and it is generally expected to occur before the event occurs. If it exceeds the 95% confidence interval, it indicates that the policy effect is significant. It is generally hoped that this situation will occur after the event occurs (After), and its values above and below 0 represent the positive and negative effects of the policy.

5.2 Machine Learning Models

Supervised learning is the most common and widely used method in machine learning, which predicts target variables through training datasets, where each training sample has a known label or output value. [16] Supervised learning has also been widely applied in the financial field, such as stock price prediction, risk assessment, credit scoring, etc. By training models, patterns and trends in historical financial data can be analyzed, helping investors and financial institutions make wiser decisions. Here are some common supervised learning algorithms:

5.2.1 Linear Regression (gradient descent method)

Linear regression is a regression analysis that uses linear regression equations to model the relationship between one or more independent variables and the dependent variable. This is an algorithm used to predict continuous values, such as predicting house prices or stock prices.

The following experiment uses the Scientific Platform Serving for Statistics Professional to provide a house price dataset. Based on the house type, elevator, area, age, decoration level, plot ratio, and greenery, the gradient boosting tree (GBDT) regression method is used to estimate the house price of the house. [18]

Analysis steps:

(1)Establishing a linear regression model through training set data

Table 18 Feature importance table

Feature name	Feature importance
House age	1.155
Distance to subway station	-0.005
Number of convenience stores in walking area	-0.321

Table 18 shows the weights of each feature in the model, where negative signs represent negative influences and positive signs represent positive influences.

(2)Calculate feature importance through the established linear regression model.

Table 19 Model Evaluation Results Table

	MSE	RMSE	MAE	MAPE	R^2
Training set	66.518	8.156	6.45	20.022	0.636
Test set	40.679	6.378	5.33	15.308	0.766

Table 19 shows the predictive evaluation metrics for the cross validation set, training set, and test set, which measure the predictive performance of the linear regression model through quantitative metrics. Among them, the evaluation indicators of the cross validation set can continuously adjust hyperparameters to obtain reliable and stable models.

MSE (Mean Square Error): The expected value of the square of the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the model.

RMSE (Root Mean Square Error): is the square root of MSE, and the smaller the value, the higher the accuracy of the model.

MAE (Mean Absolute Error): The average value of absolute error, which can reflect the actual situation of prediction error. The smaller the value, the higher the accuracy of the model.

MAPE (Mean Absolute Percentage Error): is the deformation of MAE, which is a percentage value. The smaller the value, the higher the accuracy of the model.

R^2: Compared to using only the mean, the closer the result is to 1, the higher the accuracy of the model.

(3)Apply the established linear regression model to training and testing data to obtain model evaluation results.

Table 20 Forecast situation table

Predict test set results Y	Price per unit area	House age	Distance to subway station	Number of convenience stores in walking area
53.5735049347449	59	6.6	90.45606	9
34.41644748845873	40.8	35.5	640.7391	3
42.31150170215638	36.3	32.5	424.5442	8
19.50610766163648	20	13.8	4082.015	0
53.11704545660477	54.4	6.8	379.5575	10
35.70881206445965	29.5	12.3	1360.139	1
34.41821234220269	36.8	35.9	616.4004	3
30.9698169253532 75	25.6	20.5	2185.128	3
32.86683861257207	29.8	38.2	552.4371	2
33.58522074380241	26.5	18	1414.837	1
43.75795104009253	40.3	11.8	533.4762	4
40.79793508910778	36.8	30.8	377.7956	6
48.82010203887957	48.1	13.2	150.9347	7
26.6331527189736	17.7	25.3	2707.392	3
46.96642129293321	43.7	15.1	383.2805	7

Table 20 shows the prediction of linear regression models on test data.

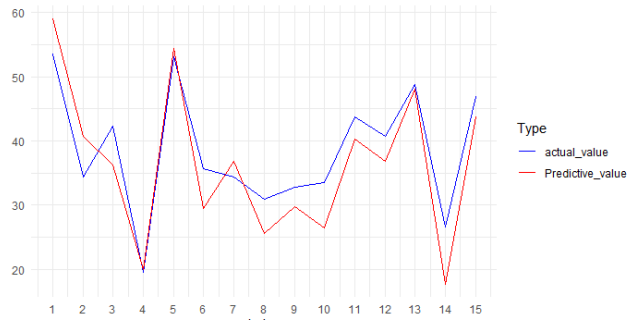

Figure11 Result prediction chart

Figure 11 shows the prediction of the linear regression model on the test data, with the blue line representing the true value and the green line representing the predicted value.

5.2.2 Decision Tree Regression

Decision tree regression is a predictive model that represents instances as points in space, using a straight line to separate data points. For example, evaluating and predicting the impact of environmental factors on housing prices. [19]

The following experiments were conducted using the sklearn boston datasets dataset for prediction, which includes:

CRIM: Urban per capita crime rate

Zn: The proportion of residential land exceeding 25000 square feet

INDUS: The proportion of non retail commercial land in urban areas

CHAS: Charles River dummy variable (1 if the boundary is a river; 0 otherwise)

NOX: Nitric oxide concentration

RM: Average number of residential rooms

AGE: The proportion of self owned houses built before 1940

DIS: Weighted distance to the five central areas of Boston

RAD: Proximity index of radiative highways

TAX: Full value property tax rate for every $10000

PTRATIO: Urban teacher-student ratio

B: 1000 (Bk -0.63) ^ 2, where Bk refers to the proportion of black people in the town

MEDV: is the target variable of the Boston housing price dataset, which represents the median value of houses occupied by homeowners. This value is in thousands of dollars

LSTAT: The proportion of people with lower status in the population

Analysis steps:

(1)Establish a decision number regression model using training set data to obtain a decision tree structure.

Table 21 Model parameter table

parameter name	Parameter value
training time	0.009s
Data segmentation	0.7
Data shuffling	no
Cross-validation	no
Node Split Evaluation Criteria	friedman_mse
Feature dividing point selection criteria	best
Maximum feature proportion considered when dividing	None
Minimum number of samples for internal node splitting	2
Minimum number of samples for leaf nodes	1
Minimum weight of samples in leaf nodes	0
maximum depth of tree	10
maximum number of leaf nodes	50
Threshold for node division impurity	0

Table 21 shows the configuration of various model parameters and the training duration of the model.

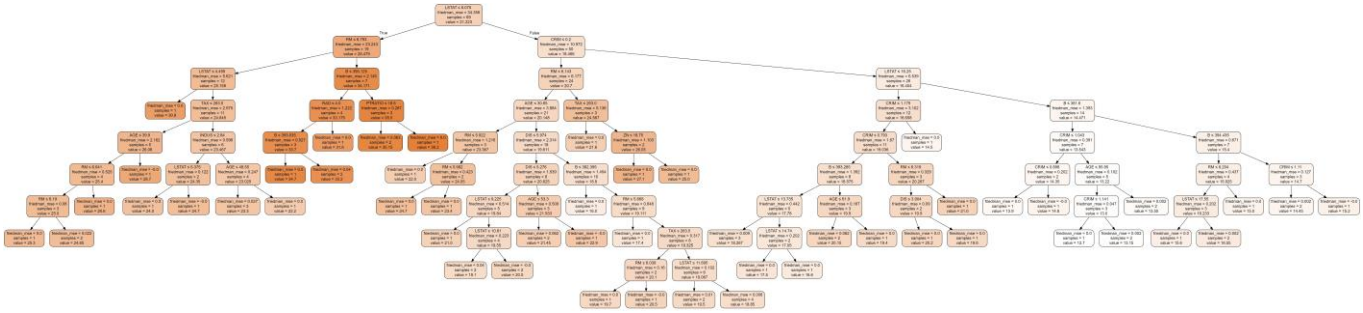

Figure12 Decision tree structure diagram

Figure 11 shows the decision tree structure, where the internal nodes provide the specific segmentation of the branched features, that is, partitioning based on a certain segmentation value of a certain feature.

(2) Calculate feature importance through established decision trees.

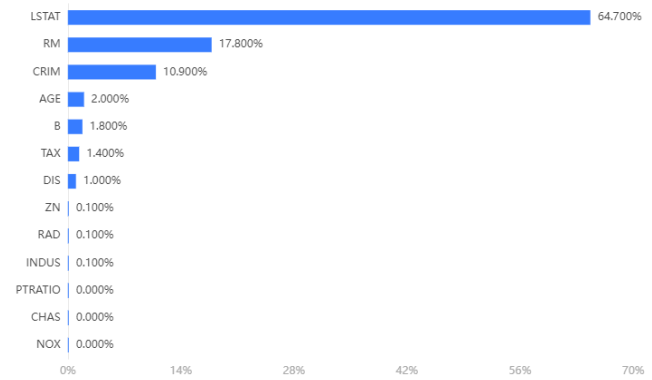

Figure13 importance ratio chart

Figure 13 shows the importance ratio of each feature (independent variable).

(3)Apply the established decision tree regression model to training and testing data to obtain model evaluation results.

Table 22 Predictive evaluation index table

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.011	0.105	0.06	0.28	1
Test set	20.745	4.555	3.535	13.117	0.194

Table 22 shows the predictive evaluation metrics for the cross validation set, training set, and test set, which measure the predictive performance of the decision tree through quantitative metrics. Among them, the evaluation indicators of the cross validation set can continuously adjust hyperparameters to obtain reliable and stable models.

MSE (Mean Square Error): The expected value of the square of the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the model.

RMSE (Root Mean Square Error): is the square root of MSE, and the smaller the value, the higher the accuracy of the model.

MAE (Mean Absolute Error): The average value of absolute error, which can reflect the actual situation of prediction error. The smaller the value, the higher the accuracy of the model.

MAPE (Mean Absolute Percentage Error): is the deformation of MAE, which is a percentage value. The smaller the value, the higher the accuracy of the model.

R^2: Compared to using only the mean, the closer the result is to 1, the higher the accuracy of the model.

(4)Due to the randomness of decision trees, the results of each operation are different.

Table 23 Classification results table

Predict test set result Y	MEDV	CRIM	ZN	PTRATIO	LSTAT	RAD	B	INDUS	TAX	CHAS	NOX	RM	DIS	AGE
19.5	20.9	0.12816	12.5	18.9	8.79	4	396.9	6.07	345	0	0.409	5.885	6.498	33
23.3	24.2	0.08826	0	19.2	6.72	4	383.73	10.81	305	0	0.413	6.417	5.2873	6.6
24.7	21.7	0.15876	0	19.2	9.88	4	376.94	10.81	305	0	0.413	5.961	5.2873	17.5
23.3	22.8	0.09164	0	19.2	5.52	4	390.91	10.81	305	0	0.413	6.065	5.287	37.8
23.3	23.4	0.19539	0	19.2	7.54	4	377.17	10.81	305	0	0.413	6.245	5.2873	6.2
23.3	24.1	0.07896	0	18.7	6.78	5	394.92	12.83	398	0	0.437	6.273	4.2515	6
27.1	21.4	0.09512	0	18.7	8.94	5	383.23	12.83	398	0	0.437	6.286	4.5026	45
27.1	20	0.10153	0	18.7	11.97	5	373.66	12.83	398	0	0.437	6.279	4.0522	74.5
19.1	20.8	0.08707	0	18.7	10.27	5	386.96	12.83	398	0	0.437	6.14	4.0905	45.8
27.1	21.2	0.05646	0	18.7	12.34	5	386.4	12.83	398	0	0.437	6.232	5.0141	53.7
21	20.3	0.08387	0	18.7	9.1	5	396.06	12.83	398	0	0.437	5.874	4.5026	36.6
23.3	28	0.04113	25	19	5.29	4	396.9	4.86	281	0	0.426	6.727	5.4007	33.5
22.2	23.9	0.04462	25	19	7.22	4	395.63	4.86	281	0	0.426	6.619	5.4007	70.4
23.3	24.8	0.03659	25	19	6.72	4	396.9	4.86	281	0	0.426	6.302	5.4007	32.2
23.3	22.9	0.03551	25	19	7.51	4	390.64	4.86	281	0	0.426	6.167	5.4007	46.7

Table 23 shows the classification results of the decision tree model on the test data, with the classification result value being the classification group with the highest prediction probability.

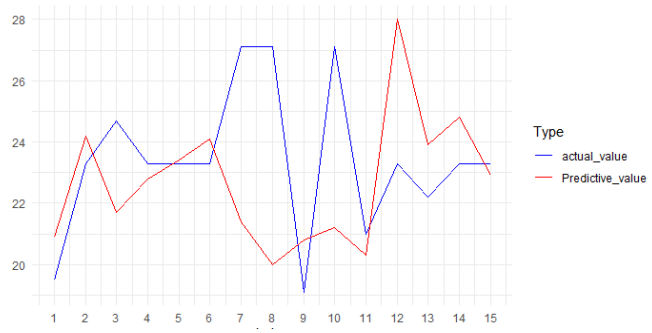

Figure14 Result prediction chart

Figure 14 shows the prediction of the decision tree on test data, with the blue line representing the true value and the green line representing the predicted value.

5.2.3 Random Forest Regression

Random forest regression is the process of generating numerous decision trees by randomly sampling the sample observations and feature variables of the modeling dataset. Each sampling result is a tree, and each tree generates rules and judgment values that match its own attributes. The forest ultimately integrates the rules and judgment values of all decision trees to achieve the regression of the random forest algorithm. [20]

Analysis steps:

(1)Establish a random forest regression model using training set data.

Table24 Model parameter configuration table

Parameter name	Parameter value
training time	0.084s
Data segmentation	0.7
Data shuffling	no
Cross-validation	no
Node Split Evaluation Criteria	mse
Maximum feature proportion considered when dividing	None
Minimum number of samples for internal node splitting	2
Minimum number of samples for leaf nodes	1
Minimum weight of samples in leaf nodes	0
maximum depth of tree	10
maximum number of leaf nodes	50
Threshold for node division impurity	0
Number of decision trees	100
Sampling with replacement	true
Out-of-bag data testing	false

Table 24 shows the configuration of various model parameters and the training duration of the model

(2)Calculate feature importance through the established random forest.

Table25 Feature importance ratio table

Feature name	Feature importance
area	75.80%
House age	10.60%
Volume rate	8.70%
House type_numeric code	2.00%
greening rate	1.50%
Elevator_Digital coding	1.10%
Decoration level	0.40%

Table 25 shows the importance ratio of each feature (independent variable).

(3)Apply the established random forest regression model to training and testing data to obtain model evaluation results.

Table 26 Predictive evaluation index table

	MSE	RMSE	MAE	MAPE	R^2
Training set	23.087	4.805	2.905	3.578	0.969
Test set	239.586	15.479	12.735	14.272	-0.33

Table 26 shows the predictive evaluation metrics for the cross validation set, training set, and test set, which measure the predictive performance of random forests through quantitative metrics. Among them, the evaluation indicators of the cross validation set can continuously adjust hyperparameters to obtain reliable and stable models.

MSE (Mean Square Error): The expected value of the square of the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the model.

RMSE (Root Mean Square Error): is the square root of MSE, and the smaller the value, the higher the accuracy of the model.

MAE (Mean Absolute Error): The average value of absolute error, which can reflect the actual situation of prediction error. The smaller the value, the higher the accuracy of the model.

MAPE (Mean Absolute Percentage Error): is the deformation of MAE, which is a percentage value. The smaller the value, the higher the accuracy of the model.

R^2: Compared to using only the mean, the closer the result is to 1, the higher the accuracy of the model.

oob_score: For regression problems, oob_score is the R of out of bag data ^{2}。 If we choose to have resampling during the tree building process, approximately one-third of the records were not sampled. A control dataset is formed naturally without being extracted, which can be used for model validation. So random forest does not require additional data to be reserved for cross validation. Its algorithm is similar to cross validation, and the out of bag error is an unbiased estimate of the prediction error (only

when the algorithm parameter selects "out of bag test data" will the generalization ability of the model be tested through oob_score).

(4)Due to the randomness of random forests, the results of each operation are different. If the current training model is saved, the data can be directly uploaded to the current training model for calculation and prediction.

Table27 Result prediction table

Predict test set resultsY	house price	area	House age	Decoration level	Volume rate	Greening rate	House type	elevator
95.2923999999999	95	55	6	3	2.5	0.4	2	2
77.395	85	55	3	3	2.9	0.35	2	2
79.41	82	55	5	3	2.9	0.35	2	1
78.705	92	55	9	3	2.9	0.35	2	2
91.61	90	58	7	3	2.29	0.4	2	2
87.9208	115	59	3	3	2.5	0.4	2	2
94.5034666666666	115	59	4	3	2.5	0.4	2	2
85.2795000000001	97	59	7	3	3.24	0.45	2	1
86.095	97	59	9	3	3.24	0.45	2	2
85.7705666666667	94	59	6	3	3.24	0.45	2	1
86.74	82	59	9	2	2.8	0.43	2	2
86.76	96	59	9	3	2.8	0.43	2	1
97.0933333333333	98	59	9	3	2.58	0.41	2	2
94.9440666666666	116	59	5	3	2.5	0.4	2	2
94.8511666666667	128	59	5	2	2.5	0.4	2	2

Table 27 shows the prediction performance of the random forest model on test data.

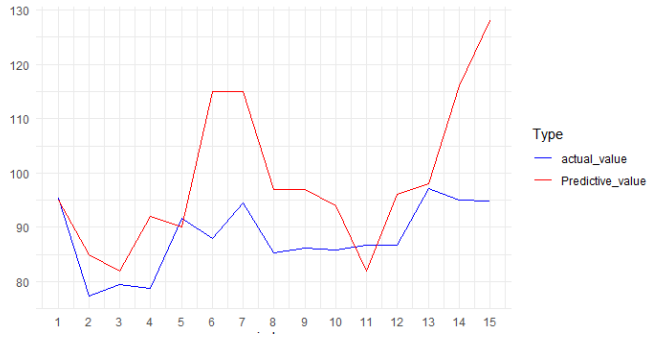

Figure15 Result prediction chart

Figure 15 shows the prediction of the random forest model on test data, with the blue line representing the true value and the green line representing the predicted value.

5.2.4 Adaboost Regression

Adaboost assigns a high weight to learners with low error rates and a low weight to learners with high error rates, combining weak learners with corresponding weights to generate strong learners. The difference between regression problems and classification problems lies in the way the error rate is calculated. Classification problems generally use a 0/1 loss function, while regression problems generally use a squared loss function or a linear loss function.

Analysis steps:

(1)Establish an Adaboost regression model using training set data.

Table28 Model parameter configuration table

Parameter name	Parameter value
training time	0.086s
Data segmentation	0.7
Data shuffling	no
Cross-validation	no
Number of base classifiers	100
loss function	linear
base classifier	Decision tree classifier
learning rate	1

Table 28 shows the configuration of various model parameters and the training duration of the model.

(2)Calculate feature importance through the established Adaboost.

Table29 Feature importance ratio table

Feature name	Feature importance
area	74.80%
House age	11.60%
Volume rate	5.90%
House type_numeric code	3.20%
greening rate	2.20%
Elevator_Digital coding	1.90%
Decoration level	0.40%

Table 29 shows the importance ratio of each feature (independent variable).

(3)Apply the established Adaboost regression model to training and testing data to obtain model evaluation results.

Table30 Predictive evaluation index table

	MSE	RMSE	MAE	MAPE	R^2
Training set	2.562	1.601	0.941	1.264	0.997
Test set	252.942	15.904	12.683	14.127	-0.404

Table 30 shows the predictive evaluation metrics for the cross validation set, training set, and test set, which measure the predictive performance of Adaboost through

quantitative metrics. Among them, the evaluation indicators of the cross validation set can continuously adjust hyperparameters to obtain reliable and stable models.

MSE (Mean Square Error): The expected value of the square of the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the model.

RMSE (Root Mean Square Error): is the square root of MSE, and the smaller the value, the higher the accuracy of the model.

MAE (Mean Absolute Error): The average value of absolute error, which can reflect the actual situation of prediction error. The smaller the value, the higher the accuracy of the model.

MAPE (Mean Absolute Percentage Error): is the deformation of MAE, which is a percentage value. The smaller the value, the higher the accuracy of the model.

R^2: Compared to using only the mean, the closer the result is to 1, the higher the accuracy of the model.

(4)Due to the randomness of Adaboost, the results of each operation are different. If the current training model is saved, the data can be directly uploaded to the current training model for calculation and prediction.

Table31 Result prediction table

Predict test set results Y	House price	area	House age	Decoration level	House type	Volume rate	elevator	Greening rate
93	95	55	6	3	2	2.5	2	0.4
80	85	55	3	3	2	2.9	2	0.35
80	82	55	5	3	2	2.9	1	0.35
80	92	55	9	3	2	2.9	2	0.35
93	90	58	7	3	2	2.29	2	0.4
93	115	59	3	3	2	2.5	2	0.4
93	115	59	4	3	2	2.5	2	0.4
93	97	59	7	3	2	3.24	1	0.45
93	97	59	9	3	2	3.24	2	0.45
93	94	59	6	3	2	3.24	1	0.45
96	82	59	9	2	2	2.8	2	0.43
97	96	59	9	3	2	2.8	1	0.43
97	98	59	9	3	2	2.58	2	0.41
93	116	59	5	3	2	2.5	2	0.4
93	128	59	5	2	2	2.5	2	0.4

Table 31 shows Adaboost's prediction of test data.

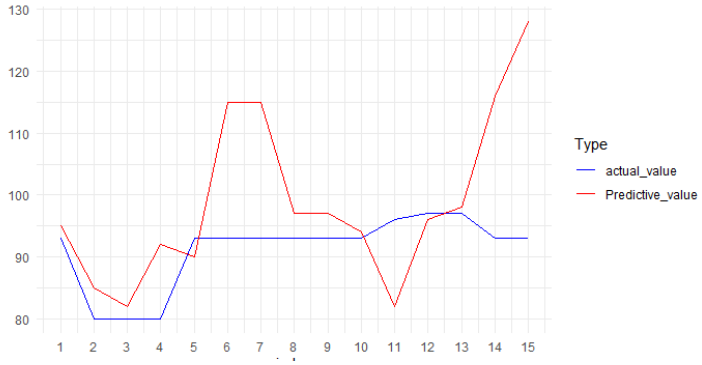

Figure16 Result prediction chart

Figure 16 shows Adaboost's prediction of test data, with the blue line representing the true value and the green line representing the predicted value.

5.2.5 Gradient Boosting Tree (GBDT) Regression

The GBDT model is an additive model that sequentially trains a set of CART regression trees, ultimately summing up the prediction results of all regression trees to obtain a strong learner, where each new tree fits the negative gradient direction of the current loss function. Finally, output the sum of this set of regression trees to obtain the regression result. [22]

Analysis steps:

(1)Establish a Gradient Boosting Tree (GBDT) regression model using training set data.

Table32 Model parameter configuration table

Parameter name	Parameter value
training time	0.034s
Data segmentation	0.7
Data shuffling	no
Cross-validation	no
loss function	friedman_mse
Node Split Evaluation Criteria	friedman_mse
Number of base learners	100
learning rate	0.1
Sampling ratio without replacement	1
Maximum feature proportion considered when dividing	None
Minimum number of samples for internal node splitting	2
Minimum number of samples for leaf nodes	1
Minimum weight of samples in leaf nodes	0
maximum depth of tree	10
maximum number of leaf nodes	50
Threshold for node division impurity	0

Table 32 shows the configuration of various model parameters and the training duration of the model.

(2)Calculate feature importance through the

established Gradient Boosting Tree (GBDT).

Table33 Feature importance ratio table

Feature name	Feature importance
area	76.90%
House age	10.10%
Volume rate	9.60%
House type_numeric code	1.60%
greening rate	1.30%
Elevator_Digital coding	0.30%
Decoration level	0.10%

Table 33 shows the importance ratio of each feature (independent variable).

(3)Apply the established Gradient Boosting Tree (GBDT) regression model to training and testing data to obtain model evaluation results.

Table34 Predictive evaluation index table

	MSE	RMSE	MAE	MAPE	R^2
Training set	0.231	0.481	0.124	0.137	1
Test set	406.421	20.16	17.034	21.059	-1.256

Table 34 shows the predictive evaluation metrics for the cross validation set, training set, and test set, which measure the predictive performance of GBDT through quantitative metrics. Among them, the evaluation indicators of the cross validation set can continuously adjust hyperparameters to obtain reliable and stable models.

MSE (Mean Square Error): The expected value of the square of the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the model.

RMSE (Root Mean Square Error): is the square root of MSE, and the smaller the value, the higher the accuracy of the model.

MAE (Mean Absolute Error): The average value of absolute error, which can reflect the actual situation of prediction error. The smaller the value, the higher the accuracy of the model.

MAPE (Mean Absolute Percentage Error): is the deformation of MAE, which is a percentage value. The smaller the value, the higher the accuracy of the model.

R^2: Compared to using only the mean, the closer the result is to 1, the higher the accuracy of the model.

(4)Due to the randomness of Gradient Boosting Tree (GBDT), the results of each operation are different.

Table 35 Result prediction table

Predict test set resultsY	House price	area	Greening rate	House age	Decoration level	elevator	Volume rate	House type
94.13345940153744	95	55	0.4	6	3	2	2.5	2
79.93477479574626	85	55	0.35	3	3	2	2.9	2
75.42911509011981	82	55	0.35	5	3	1	2.9	2
75.35024846182505	92	55	0.35	9	3	2	2.9	2
88.28886840106675	90	58	0.4	7	3	2	2.29	2
72.40401926583225	115	59	0.4	3	3	2	2.5	2
93.01221316453396	115	59	0.4	4	3	2	2.5	2
75.27687558848596	97	59	0.45	7	3	1	3.24	2
75.22960526931816	97	59	0.45	9	3	2	3.24	2
75.30847189761289	94	59	0.45	6	3	1	3.24	2
75.23054458772327	82	59	0.43	9	2	2	2.8	2
75.28787395982539	96	59	0.43	9	3	1	2.8	2
97.5544098951945	98	59	0.41	9	3	2	2.58	2
93.71652491792278	116	59	0.4	5	3	2	2.5	2
93.82176100894657	128	59	0.4	5	2	2	2.5	2

Table 35 shows the prediction of gradient boosting tree regression (GBDT) on test data.

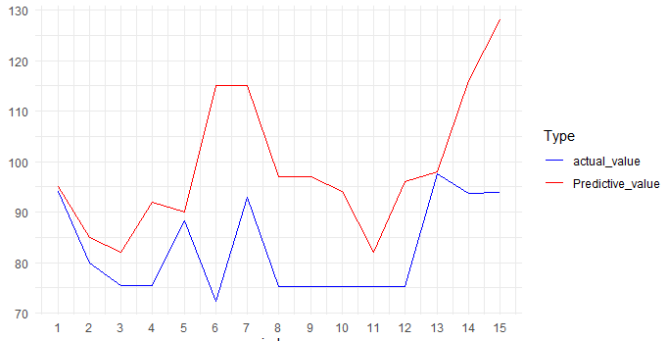

Figure17 Result prediction chart

Figure 17 shows the prediction of Gradient Boosting Tree Regression (GBDT) on test data, with the blue line representing the true value and the green line representing the predicted value.

6 Discussion

This paper uses econometric models and machine learning models to predict economic related activities. In econometrics, ARIMA model, GARCH model, VAR vector autoregressive model, robust regression, and double difference DID (double difference method) model are used to predict economic activities. Using algorithms such as linear regression (gradient descent), decision tree regression, random forest regression, Adaboost regression, and gradient boosting tree (GBDT) regression in machine learning models for economic activity prediction experiments. The experimental results are used to explore the applicability of econometric prediction methods and machine learning methods in different application scenarios in scenario driven economic forecasting.

However, there are still many shortcomings. Firstly, due to the insufficient amount of data in the experimental dataset, it may affect the accuracy of predictions. In future research, larger datasets can be considered for experiments to improve the accuracy of predictions. Secondly, some models used, such as ARIMA and linear regression, assume that the data is linear, which may limit the ability of these models to handle non-linear data. In future research, models that can handle nonlinear data can be considered. Some complex machine learning models, such as random forests and Adaboost, may have poor interpretability, which may make it difficult to understand the model's prediction results. In future research, it is possible to consider using more explanatory models or developing new methods to improve the interpretability of the models. Then, some machine learning models used may overfit the training data, which may reduce the predictive performance of the model on new data. In future research, regularization methods can be considered to prevent overfitting, or methods such as cross validation can be used to evaluate the model's generalization ability.

7 Conclusion

This paper explores how to choose suitable econometric or machine learning methods in different economic forecasting scenarios. For macroeconomic forecasting, ARIMA models and VAR vector autoregressive models can handle time series data and are suitable for predicting macroeconomic indicators. For industry economic forecasting, random forest regression and Adaboost regression can handle large amounts of data and complex data structures, making them suitable for predicting the development trends of specific industries. For product economy prediction, linear regression and decision trees can handle continuous value prediction, which is suitable for predicting the demand and price of specific products or services. For predicting the impact of major events, the double difference DID (double difference method) can largely avoid endogeneity issues in data and is suitable for evaluating policy effects. The selection and application of these models can help improve the accuracy and reliability of economic forecasting. Future researchers need to optimize and improve the data volume, model assumptions, model interpretability, and model generalization ability.

Acknowledgments

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential 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. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

XIA, Shaoliang

Faculty of Applied Mathematics and Computer Science; Belarusian State University; 4 Nezavisimosti Avenue, Minsk 220030, Belarus; e-mails: fpm.sya@bsu.by

WU, Zhen

Nanjing University of Aeronautics and Astronautics; No. 29, Yudao Street, Qinhuai District, Nanjing City, Jiangsu Province 210016, China; e-mails: 13wuzhen@sina.com

SONG, Qingqing

Faculty of Applied Mathematics and Computer Science; Belarusian State University; 4 Nezavisimosti Avenue, Minsk 220030, Belarus; e-mails: fpm.sunC@bsu.by

References

[1] Elliott, G., & Timmermann, A. (2008). Economic forecasting. Journal of Economic Literature, 46(1), 3-56.

[2] Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In Economic forecasts and expectations: Analysis of forecasting behavior and performance (pp. 3-46). NBER.

[3] Clements, M., & Hendry, D. (2002). An overview of economic forecasting. A Companion to Economic Forecasting. Oxford: Blackwell, 1-18.

[4] Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. Journal of Business & Economic Statistics, 4(1), 39-46.

[5] Makridakis, S., & Bakas, N. (2016). Forecasting and uncertainty: A survey. Risk and Decision Analysis, 6(1), 37-64.

[6] Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. Journal of macroeconomics, 24(4), 435-468.

[7] Diebold, F. X. (1998). The past, present, and future of macroeconomic forecasting. Journal of Economic Perspectives, 12(2), 175-192.

[8] D'Agostino, A., Gambetti, L., & Giannone, D. (2013). Macroeconomic forecasting and structural change. Journal of applied econometrics, 28(1), 82-101.

[9] Tsai, M. C., Cheng, C. H., Tsai, M. I., & Shiu, H. Y. (2018). Forecasting leading industry stock prices based on a hybrid time-series forecast model. PloS one, 13(12), e0209922.

[10] Batchelor, R. A., & Dua, P. (1990). Product differentiation in the economic forecasting industry. International Journal of Forecasting, 6(3), 311-316.

[11] Lai, Y., & Dzombak, D. A. (2020). Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation. Weather and Forecasting, 35(3), 959-976.

[12] Nyoni, T. (2018). Modeling and forecasting inflation in zimbabwe: a generalized autoregressive conditionally heteroskedastic (GARCH) approach.

[13] Lütkepohl, H. (2013). Vector autoregressive models. In Handbook of research methods and applications in empirical macroeconomics (pp. 139-164). Edward Elgar Publishing.

[14] El-Melegy, M. T. (2014). Model-wise and point-wise random sample consensus for robust regression and outlier detection. Neural networks, 59, 23-35.

[15] Zhou, G., Liu, C., & Luo, S. (2021). Resource allocation effect of green credit policy: based on DID model. Mathematics, 9(2), 159.

[16] Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised machine learning: algorithms and applications. Fundamentals and methods of machine and deep learning: algorithms, tools and applications, 1-16.

[17] Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, 234-240.

[18] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11) [Online Application Software]. Retrieved from https://www.spsspro.com.

[19] Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. Future generation computer systems, 13(2-3), 197-210.

[20] Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., & Lei, G. (2021). A data-driven shale gas production forecasting method based on the multi-objective random forest regression. Journal of Petroleum Science and Engineering, 196, 107801.

[21] Wang, F., Li, Z., He, F., Wang, R., Yu, W., & Nie, F. (2019). Feature learning viewpoint of AdaBoost and a new algorithm. IEEE Access, 7, 149890-149899.

[22] Jun, M. J. (2021). A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: The case of the Seoul metropolitan area. International Journal of Geographical Information Science, 35(11), 2149-2167.