Modern Management Science & Engineering
ISSN 2052-2576
Vol. 8, No. 3, 2026
www.scholink.org/ojs/index.php/mmse

Original Paper

Forecasting China's Total Retail Sales of Consumer Goods: A Multi-Model Comparative Approach

Xianghao Meng[1] & Zeran Cao[1]

[1] Renmin University of China, Beijing, Beijing, China

Received: May 29, 2026 Accepted: July 07, 2026 Online Published: July 29, 2026

doi:10.22158/mmse.v8n3p145 URL: http://dx.doi.org/10.22158/mmse.v8n3p145

Abstract

An accurate forecast of the TRSCG is an important requirement for macroeconomic monitoring and policymaking since it represents a valuable index related to the domestic consumption of China. In the current research, 304 monthly values of TRSCG from January 2000 to April 2025 are obtained from the National Bureau of Statistics of China for comparing the accuracy of statistical, machine learning, and deep learning techniques. Data preprocessing, trend decomposition, and stationarity analysis (Augmented Dickey–Fuller test) have been performed, after which the original non-stationary series has been made stationary using first-order differencing. Four forecasting methods, including ARIMA, SARIMA, LightGBM, and Long Short-Term Memory (LSTM), are tested using RMSE, MAE, and MAPE measures of errors. It was found that the SARIMA method is much more accurate than ARIMA thanks to its ability to capture seasonal patterns; meanwhile, LightGBM gives good forecasting results when using macroeconomic factors. Among all other methods, LSTM is the most accurate one thanks to the ability to capture nonlinear patterns and long-term temporal dependency.

Keywords

Total Retail Sales of Consumer Goods (TRSCG), Time Series Analysis, SARIMA, LightGBM, Long Short-Term Memory (LSTM)

1. Introduction

1.1 Research Background and Significance

Total Retail Sales of Consumer Goods (TRSCG) is a very important indicator that measures the domestic consumption as well as the country's macroeconomic performance, and is considered one of the basic variables in the country's national accounting system. This variable directly determines the consumption tendency of the households and also the health condition of the consumer market [1]. Over the past few years, along with the continuous improvement of China's economic structure as well as the upgrading of consumption pattern, TRSCG has been growing steadily, but with more complicated structural features and strong cyclic volatility. Moreover, in the context of the rapid development of "Internet Plus" consumer goods ecosystem, various promotions and online shopping festivals are making the seasonality features of retail sales much more evident. Therefore, there is a high need to have reliable predictions of TRSCG, which is also of great practical significance [3].

From the academic viewpoint, classical time series forecasting methods like Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) models are commonly used for economic forecasting because of their solid theoretical background. Despite being good at estimating linear trends and seasonality, these models show poor results in case of dealing with nonlinear relations or external shock (e.g., COVID-19 shock). Thanks to machine learning and deep learning developments, there have been opportunities to apply new models such as Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) to time series forecasting. The main advantage of using these methods is the ability to estimate nonlinear dependencies and complicated temporal relations. However, despite the rapid development in forecasting methods, there has not been a lot of researches comparing forecasting methods in the field of consumer-related time series, especially under the unified framework taking into account forecasting performance, robustness, and interpretability [4]. Moreover, most of the studies are based on the use of lagged values for feature construction and do not take into account the impact of exogenous macroeconomic factors on consumer behavior.

Under such circumstances, this study uses monthly data on TRSCG for the period from January 2000 to April 2025 to build different forecasting models and evaluate them. The combination of both the explanatory power of traditional statistical models and the adaptability of machine learning and deep learning approaches is expected to add some novelty to time series forecasting techniques and improve their efficiency for macroeconomic analysis.

2. Data Exploration and Feature Engineering

2.1 Data Description

The dataset used for this study is sourced from the National Bureau of Statistics of China (NBS) and includes the monthly observations on Total Retail Sales of Consumer Goods (TRSCG) during January 2000 to April 2025, as shown in Table 1 below. One of the most significant economic indicators, TRSCG is an important gauge for the overall size and development level of the Chinese consumer market and a major tool for studying economic dynamics.

A total of 304 months of data is included in the dataset and provides sufficient data coverage to conduct reliable time series analysis and forecasting. To provide consistent and reliable data for this study, the officially released revised data of the National Bureau of Statistics was used.

Table 1. Monthly Total Retail Sales of Consumer Goods (Partial Sample) (Unit: RMB 100 million)

Time	Total Retail Sales of Consumer Goods	Time	Total Retail Sales of Consumer Goods	Time	Total Retail Sales of Consumer Goods
2025-4	37174.1	2024-4	35699.1	2023-4	34910.5
2025-3	40940.5	2024-3	39019.9	2023-3	37855.4
2025-2		2024-2		2023-2	
2025-1		2024-1		2023-1	
2024-12	45171.5	2023-12	43550.2	2022-12	40542.3
2024-11	43763	2023-11	42505	2022-11	38615.1
2024-10	45396.4	2023-10	43333	2022-10	40270.5
2024-9	41112.3	2023-9	39826	2022-9	37744.9
2024-8	38725.8	2023-8	37932.7	2022-8	36257.9
2024-7	37757.4	2023-7	36760.7	2022-7	35870.1
2024-6	40731.6	2023-6	39951.4	2022-6	38742.5
2024-5	39211	2023-5	37803.3	…	…

Source: National Bureau of Statistics of China.

2.2 Data Preprocessing and Descriptive Analysis

2.2.1 Data Preprocessing

The raw dataset downloaded from the National Bureau of Statistics contains a number of missing observations. As reported in Table 2, the monthly retail sales variable (RSCG) contains 28 missing values, while the cumulative retail sales variable (TRSCG) contains 14 missing values. These missing observations arise from the statistical reporting practice adopted by the National Bureau of Statistics, whereby certain monthly indicators are not separately reported for January and February. Instead, cumulative retail sales for January and February are released together at the end of February. Consequently, the cumulative series naturally lacks observations for January each year.

No duplicate observations are found in the dataset; therefore, duplicate sample processing is unnecessary.

Table 2. Summary of Missing and Duplicate Observations

Variable	Abbreviation	Missing Values	Duplicate Values
Monthly Retail Sales of Consumer Goods	RSCG	28	0
Cumulative Retail Sales of Consumer Goods	TRSCG	14	0

The missing observations in the target variable (RSCG) must be properly imputed before model development. Since the monthly and cumulative retail sales satisfy the relationship expressed in Equation (1),

$$RSCG_t = \begin{cases} TRSCG_t - TRSCG_{t-1} & t \neq 1月 \\ TRSCG_t & t = 1月 \end{cases} \tag{1}$$

where the value of t ranges from 1 to 12, directly treating missing values as zero would result in the February monthly retail sales becoming identical to the cumulative value and the January retail sales being equal to zero, which is economically implausible. Therefore, the missing values are imputed according to Equation (2):

$$RSCG_t = \begin{cases} TRSCG_t - TRSCG_{t-1} & TRSCG_t \neq NaN \ \& TRSCG_{t-1} \neq NaN \\ \dfrac{TRSCG_2}{k} & t = 1 \ \& \ t = 2 \end{cases} \tag{2}$$

where k denotes an empirical coefficient. When k=2, the missing observation is replaced by the average value implied by the cumulative data, which is the imputation strategy adopted throughout this study.

2.2.2 Trend Analysis

Before constructing forecasting models, descriptive statistical analysis is conducted to examine the distributional characteristics and variability of the monthly retail sales series.

Table 3. Descriptive Statistics of Monthly Retail Sales of Consumer Goods

Statistic	**Description**	**Monthly Retail Sales**
Count	Number of observations	304.00
Mean	Average	19,410.89
Standard deviation	Dispersion	13,343.02
Minimum	Lowest value	2,571.50
25th percentile	First quartile	6,151.05
Median	50th percentile	17,620.75
75th percentile	Third quartile	31,810.95
Maximum	Highest value	45,396.40

From the descriptive statistics, it can be seen that there are 304 observations, which imply that the data represents a complete continuous series on monthly basis. The average monthly retail sales are at RMB 1,941.09 billion, which is quite an impressive figure for the magnitude of the market of China. Large standard deviation (at RMB 1,334.30 billion) implies temporal variation; thus, the consumer expenditure may vary depending upon the business cycle and macroeconomic policies and also due to any unexpected event like the outbreak of the COVID-19 disease.

The large gap between the maximum (at RMB 4,539.64 billion) and minimum (at RMB 257.15 billion) figures shows the variation in the monthly retail sales figures. In addition, the median (at RMB 1,762.08 billion) figure is less than the mean, which means that the data is moderately positively skewed, with few months having exceptionally high consumption. First, second and third quartiles are at RMB 615.11 billion, RMB 1,762.08 billion and RMB 3,181.10 billion, respectively. This means that most of the months fall within this range, and those months above the third quartile figure have very strong consumer demand because of holidays and promotions in the whole nation.

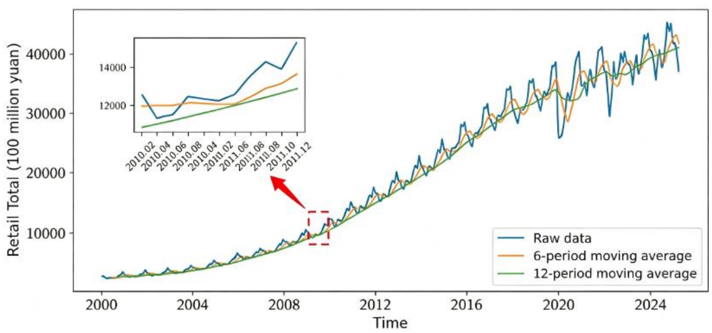

Figure 1. Trend of Monthly Total Retail Sales of Consumer Goods

As shown in Figure 1 below, retail sales in China have followed a striking stepwise upward pattern since 2000 up to 2025, generally in keeping with macroeconomic development in China.

In the initial stage (2000-2006), the growth in light manufacturing industries has led to stable growth since then when China has entered the World Trade Organization (WTO). After 2007 and until 2012, China's economic stimulus package worth RMB 4 trillion has driven the increase in domestic demands for retail products, resulting in accelerated growth when retail sales exceeded the trillion-yuan mark. During the period 2013 to 2019, continuous improvements in consumer spending and rapid expansion of online retailing such as Pinduoduo, Meituan, and JD.com have facilitated higher growth rates of over 10% per year. However, the year 2020 was characterized by the emergence of COVID-19.

There is also noticeable seasonality in the series. Month-on-month fluctuations reach 15% during January and February because of the Spring Festival holiday period. Besides, there are seasonal peaks in June and November because of China's big e-commerce festivals like "618 Shopping Festival" and "Double Eleven (Singles' Day)".

Two specific structural turning points are especially remarkable. First, about 2012, the 12-month moving average shows a decrease in the rate of growth, reflecting the change in the stage of economic development in China – the stage where the tertiary sector became more valuable in terms of economic contribution, while the consumption type switched from being necessity driven to developmental consumption. The second turning point happened in 2020 due to the pandemic, which resulted in a clear V-shaped drop in retail sales. The six-month moving average shows that in February 2020, retail sales decreased by about 20.5% compared to the same period of last year, while the 12-month moving average was still positive, indicating that the pandemic was an extreme and temporary shock rather than a trend-reversal event.

Moreover, the pace of recovery after the pandemic in 2021 was higher than the historical average, proving the phenomenon known as "revenge consumption." Starting from 2023, the seasonality pattern becomes consistent with the pre-pandemic one.

2.2.3 Seasonal Analysis

Given the presence of the long-term trend in the data, it becomes important to study the seasonality of the given monthly retail sales time series. As can be seen in Figure 1, while there is an evident increasing trend present in the data, retail sales display strong cyclic variations on the yearly basis. To detect and analyze the seasonality in the data, the current paper utilizes two popular time-series decomposition models [5]: classical additive decomposition and Seasonal Trend decomposition using Loess (STL).

The main idea behind both models is that the initial time series is decomposed into three components: trend, seasonality, and residual. The advantage of the STL model compared to classical additive decomposition is its increased flexibility, which allows for time-changing seasons. Thus, the STL model is particularly suitable for capturing the changes in the frequency and magnitude of retail sales fluctuations during the COVID-19 pandemic.

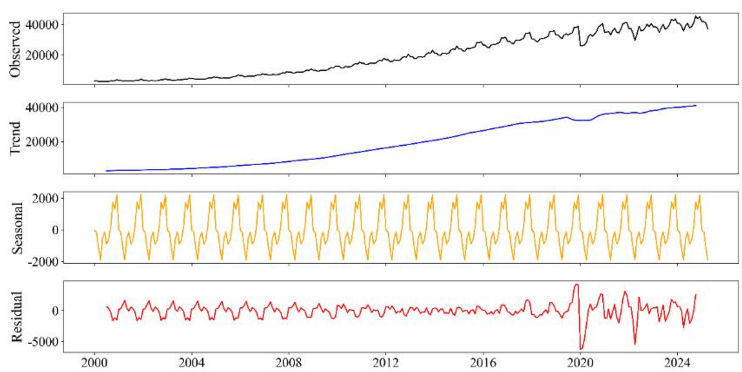

Figure 2. Classical Additive Decomposition of the Monthly Retail Sales Series

As is evident from the decomposition illustrated in Figure 2, there are three main features of the sales time series in question. First, the seasonal component displays a very stable pattern repeated annually, characterized by two peaks – during the Spring Festival (January–February) and two e-commerce shopping festivals – in June and November. The seasonal pattern became significantly distorted due to the coronavirus pandemic that occurred in 2020 but restored its historical stability starting from 2022.

Second, the residual component represents unusual events that cannot be captured through trend and seasonal decomposition. Specifically, the exceptionally low residual value for February 2020 is caused by the effects of quarantine, while the positive residuals during 2021 result from the increased consumption activity after the pandemic restrictions were lifted.

Overall, the decomposition shows that the long-term growth trend of the Chinese consumer market is preserved. External shocks affect only the short-term retail performance by disturbing seasonal pattern and increasing irregularity. Additionally, the small size of the residual component proves the successful capturing of systematic features of the original time series through decomposition.

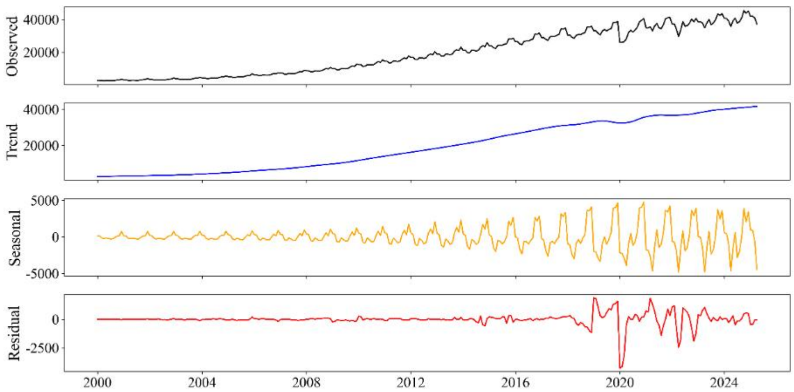

Figure 3. STL Decomposition of the Monthly Retail Sales Series

The STL decomposition shown in Figure 3 offers further proof of the highly seasonality of the retail sales time series data. It is possible to see cyclical movements during each year, and the ability of STL to capture the seasonal pattern is evident. Once the seasonal and irregular components are taken out of the data, the trend component that has been estimated shows consistent upward movement over time, supporting the idea that there has been steady growth in China's consumer market.

Apart from that, the residual component stays in a rather narrow interval for the majority of the observations, which suggests that STL decomposition is efficient in capturing the main trend and seasonal components, having very little noise left.

2.3 Stationarity Testing

A preliminary visual examination of the RSCG series indicates that the initial series is probably non-stationary. But in order to confirm stationarity before estimation of the model, it becomes necessary since the classical time series models such as ARMA and VAR depend upon the assumption that the initial series is stationary [6]. This section therefore provides a systematic examination of the stationarity of the RSCG series using graphical as well as formal methods of testing.

Mathematically, the concept of first order differencing can be written as follows:

$$\Delta Y_t = Y_t - Y_{t-1} \tag{3}$$

where Yt is the value observed for the original time series at period t. Through the application of first order difference, the deterministic trend can easily be eliminated, making the original time series weakly stationary with a relatively constant mean and variance.

Figure 4 below shows the comparison of the original RSCG time series and its first order differenced series. It is evident from the figure that the original time series has an upward trend over time, with increasing magnitude of fluctuations, thus a feature of non-stationary time series.

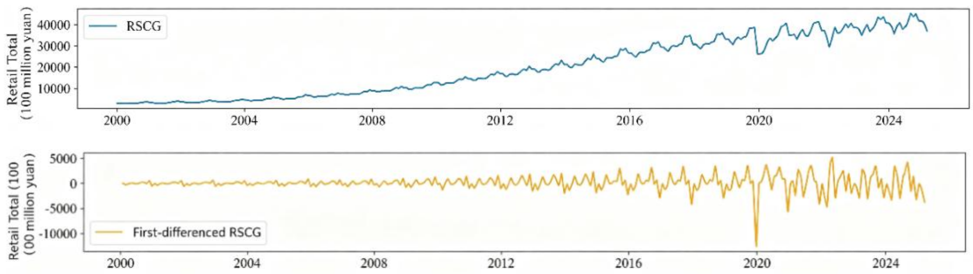

Figure 4. Comparison of the Original and First-Order Differenced RSCG Series

In order to offer statistical backing to our graphical findings, the Augmented Dickey-Fuller (ADF) unit root test is carried out on the original as well as the differenced series. The findings are depicted in Table 4.

Table 4. Comparison of ADF Test Results Before and After First-Order Differencing

Statistic	Original Series	First-Order Difference
ADF statistic	0.7734	-4.8410
p-value	0.9912	0.0000
Number of lags	15	14
Number of observations	288	288
1% critical value	-3.4533	-3.4533
5% critical value	-2.8716	-2.8716
10% critical value	-2.5721	-2.5721

According to Table 4, the ADF test statistics of the original series is 0.7734, while the p-value associated with it is 0.9912, which is relatively much higher compared to the usual level of significance. Hence, we fail to reject the null hypothesis of a unit root, thereby suggesting the non-stationarity of the original RSCG series.

After carrying out first order differencing on the original series, the ADF test statistics falls to −4.8410, which is significantly lower than the 1% critical value (−3.4533), while the p-value is nearly zero. Thus, the null hypothesis is firmly rejected, proving the stationarity of the series and its suitability for time-series analysis.

In order to establish the stationarity of the differenced series, the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are plotted. These are presented in Figure 5.

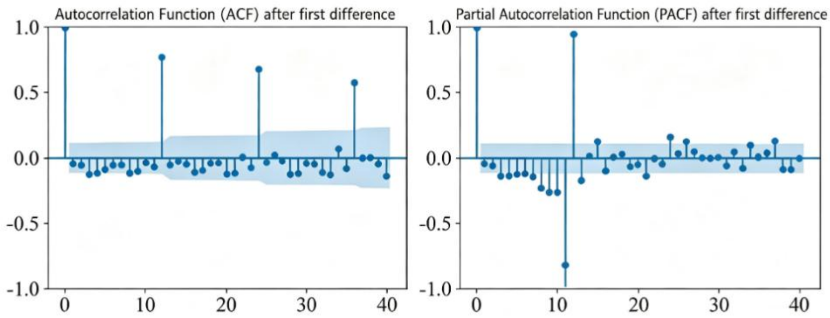

Figure 5. ACF and PACF of the First-Order Differenced RSCG Series

As is apparent in the plot of the ACF, except the one trivial correlation at lag zero, the other autocorrelations reduce quickly and stay within the statistical significance bands in subsequent lags. No signs of any form of long-term dependence have been detected, implying that the differenced series does not have the properties of a non-stationary series anymore.

The PACF plot indicates relatively high partial autocorrelations at the first two lags, after which the autocorrelations quickly fall to zero, displaying a short-tail pattern. Together with the results of the ADF test presented in Table 4, the autocorrelation analysis serves as further proof that the first-order difference of the RSCG series has met the stationarity assumption necessary for modeling.

3. Empirical Analysis of RSCG Forecasting

3.1 Model Selection and Evaluation Strategy

3.1.1 Model Selection and Experimental Environment

Selection of forecasting technique is an important determinant of predictive performance and interpretability in case of retail sales forecast. Due to the presence of clear long-term trend, seasonality and possible nonlinearity in the Total Retail Sales of Consumer Goods (RSCG) time series, the following forecasting paradigms have been chosen: statistical, machine learning and deep learning models. All of them are coded in RStudio with R version 4.4.3.

Preprocessing steps required for each of the models, including Z-score standardization and Min-Max normalization, were selected depending on the peculiarities of each particular model.

3.1.2 Performance Evaluation Metrics

Selection of appropriate performance metrics plays an important role in providing objective assessment of the forecast accuracy since various measures emphasize different aspects of prediction errors. In order to assess the accuracy of forecasts produced by alternative models, this research uses the following three popular measures: root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These measures are used as complements to each other since they cover different aspects of prediction accuracy [7].

3.2 Analysis of Forecasts Using Statistical Models

3.2.1 Basics of the ARIMA Model

One of the most popular techniques used to model non-stationary time series is the Autoregressive Integrated Moving Average model [8]. Its fundamental idea is to transform a non-stationary series into a stationary one through differencing and subsequently model the resulting stationary process using autoregressive (AR) and moving average (MA) components. The standard notation of the model is $ARIMA(p,d,q)$, where p, d, and q denote the autoregressive order, differencing order, and moving average order, respectively.

The ARIMA model can be written compactly as

$$\Phi(L)(1-L)^d y_t = \Theta(L)\varepsilon_t \tag{4}$$

where $(1-L)^d y_t$ denotes the d-order differencing operator applied to the original series. Model parameters are estimated using methods such as Maximum Likelihood Estimation (MLE). Owing to its solid theoretical foundation and transparent modeling framework, the ARIMA model remains one of the benchmark methods for time-series forecasting.

3.2.2 Basic Principles of the SARIMA Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model can be considered an extension of the ARIMA approach that incorporates the seasonal component directly into the model. In addition to non-seasonal patterns, SARIMA is able to describe periodic patterns; hence, this model is especially appropriate for the analysis of economic time series with seasonal behavior [9].

Let us suppose that we have the original time series y_t, where the seasonal period is s. For monthly observations, the value of s is equal to 12. The SARIMA model is marked by $SARIMA(p,d,q)(P,D,Q)_s$, where p, d, and q correspond to the order of the non-seasonal autoregressive process, differencing, and the order of the moving average, respectively.

3.2.3 Model Development

As stated from the stationarity test done in chapter three, the original RSCG is stationary after first-order differencing. Consequently, the order of differencing remains d=1.

From figure 5 above, the ACF of the differenced series is seen to have significant correlation at lags 1 and 2 which is beyond the 95% confidence interval, and this indicates that there are moving averages. Additionally, the correlations become insignificant quickly from lag 2 onwards and show a cutoff characteristic which suggests either an MA(1) or MA(2) model can be considered.

The PACF shows a significant spike at lag 1 and then quickly drops off, showing that there is an autoregressive part and thus AR(1) should also be considered.

From the above initial considerations, a grid search is done to find the best possible ARIMA model in terms of (p,d,q) by considering all the possibilities exhaustively using the AIC as the criteria. It takes about 17.45 seconds to do the optimization, and the best model found is ARIMA(2,1,1). The results of this search are outlined in Table 5.

Table 5. Grid Search Results for ARIMA(p,d,q)

Parameter	Description	Search Space	Optimal Value
p	Autoregressive order	{1, 2, 3}	2
d	Differencing order	{1}	1
q	Moving average order	{1, 2, 3}	1

Consequently, the ARIMA(2,1,1) model will be taken as the baseline statistical model.

Analysis of the ACF indicates that there is a constant seasonal pattern of decay, while the PACF shows a sharp peak at lag 12, implying the presence of strong annual seasonality. In order to incorporate this seasonality into the analysis, the SARIMA model with seasonal period s=12 will be considered.

Just like the ARIMA model, the best combination of parameters for the SARIMA model will be determined using an AIC-based grid search. This is illustrated in Table 6.

Table 6. Grid Search Results for SARIMA(p,d,q)(P,D,Q)$_s$

Parameter	Description	Search Space	Optimal Value
p	Non-seasonal AR order	{0,1,2,3}	1
d	Non-seasonal differencing	{1}	1
q	Non-seasonal MA order	{0,1,2,3}	0
P	Seasonal AR order	{0,1,2}	1
D	Seasonal differencing	{0,1,2}	0
Q	Seasonal MA order	{0,1,2}	1
s	Seasonal period	{12}	12

Consequently, the final seasonal specification adopted in this study is SARIMA(1,1,0)(1,0,1)$_{12}$.

3.2.4 Empirical Results

Following model identification and parameter estimation, both ARIMA and SARIMA models are fitted to the training data and evaluated on the testing set. Forecasting performance is assessed using RMSE, MAE, and MAPE. The evaluation results are reported in Table 7.

Table 7. Forecasting Performance of the ARIMA and SARIMA Models

Model	RMSE	MAE	MAPE
ARIMA(2,1,1)	4,490.06	3,899.55	9.34%
SARIMA(1,1,0)(1,0,1)$_{12}$	978.11	794.70	1.91%

As can be seen from Table 7, the optimal ARIMA(2,1,1) model demonstrates an RMSE equal to 4,490.06, an MAE of 3,899.55, and a MAPE of 9.34% in the testing sample. In spite of the fact that the parameters of the model are estimated optimally according to the AIC criterion, the forecasting accuracy of the model is still rather low.

Another evidence of this is presented in Figure 6 where the forecasts made by the ARIMA model are too smooth and underestimate the real values. The model is biased negatively at several forecasting horizons and fails to react adequately to turning points, sudden jumps, and season peaks. As a result, the dynamic pattern of the retail sales series is not reproduced well.

The explanation of this drawback of the model can be given by its linear nature. Indeed, since the model uses only the autoregressive and moving average terms after differencing, it does not take into account the seasonality of the data. Therefore, the presence of periodic fluctuations in the retail sales series does not allow the forecast trajectory to go out of the certain narrow band.

On the other hand, the inclusion of the seasonal autoregressive and moving average makes the SARIMA(1,1,0)(1,0,1)12 model more accurate compared to other models. As evident from Table 7 below, the performance of the SARIMA model is better compared to that of the ARIMA model and has reduced the RMSE to 978.11, MAE to 794.70, and MAPE to 1.91%.

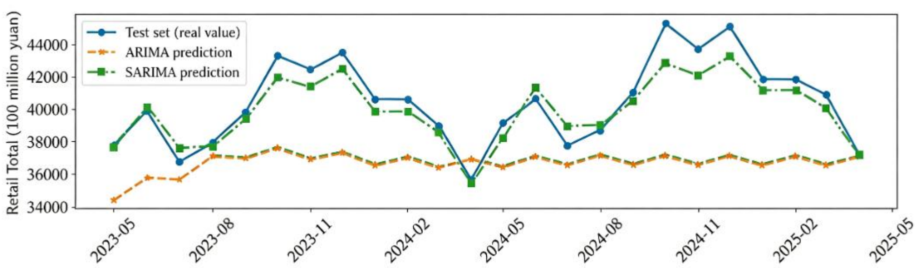

Figure 6. Forecast Comparison between the ARIMA and SARIMA Models

The graphical comparison shown in Figure 6 also provides additional proof of the dominance of the SARIMA model, with its forecast path being a close match to the actual values during the entire testing period and correctly representing the peaks of high consumption in mid-year and at the end of the year. Notably, following the third quarter of 2024, the SARIMA model is able to adequately capture the turning point and fluctuation patterns of the time series.

Contrary to the relatively straightforward linear nature of the ARIMA model, the SARIMA model proves to have significantly higher seasonal dynamic modeling capacity. The results obtained once again highlight the necessity of taking into account the seasonality effect when forecasting periodic macroeconomic variables.

3.3 Forecasting Analysis Based on the LightGBM Model

3.3.1 The Basic Principles of the LightGBM Algorithm

The LightGBM algorithm was invented to solve the problem of computationally intensive work of the conventional Gradient Boosting Decision Trees (GBDT) algorithm on big data. Traditional GBDT algorithms build decision trees iteratively, selecting samples from training data sets. This approach entails considerable computational expenses and poor usage of modern multi-core parallel computing architecture [10]. The invention of Light Gradient Boosting Machine (LightGBM) can be considered as an attempt to create an efficient ensemble learning model.

Comparing to the conventional GBDT algorithms, LightGBM implements the Histogram-based Decision Tree (HDT) algorithm that makes possible a discretization of continuous values of features into the histogram bins. Thus, the number of computations required is considerably reduced. Besides, LightGBM algorithm uses efficient parallel computations making fast training on large-scale datasets possible. Also, LightGBM has other optimizations such as the Leaf-wise tree growth with depth constraints, contributing to improved predictive accuracy and generalization. [11].

3.3.2 Feature Engineering Driven by Macroeconomic Variables and Empirical Analysis

In contrast to traditional time-series models like ARIMA and SARIMA, which take into consideration the temporal dependency of the observations, LightGBM does not directly model such a relationship. As a result, using LightGBM only on the target time series will cause the algorithm to revert to the mean and be unable to identify turning points, seasonality or external economic impacts.

To solve the problem, the current research constructs a multidimensional feature set by introducing important macroeconomic variables in relation to the aspects of consumption supply, economic environment, consumer demand and monetary environment. Such additional explanatory variables contribute to the ability of the model to detect structural relations and provide better results in predictions.

Using the macroeconomic variable classification that is widely used in the literature, two representative indicators are chosen for each economic dimension. The obtained feature set can be seen in Table 8.

Table 8. Input Features for the LightGBM Model

Category	Feature	Abbreviation
Commodity supply	Cumulative Fixed Asset Investment	FAIAV
	Total Imports and Exports	TE
Consumption environment	Real GDP (Quarterly)	GDPQCV
	Consumer Price Index (Year-on-Year, Monthly)	CPIYTM

Consumer demand	Monthly Total Retail Sales of Consumer Goods	CRSMV
	Monthly Fiscal Revenue–Expenditure Balance	BDMV
Monetary and financial factors	M2 Money Supply	M2
	M1 Money Supply	M1

Further to improve the accuracy of predictions, Bayesian optimization through the Optuna library is used to optimize the important parameters of the LightGBM model. While optimizing, each set of parameter combinations is tested by five-fold cross-validation, where RMSE is the objective function. The search space and optimal parameters are listed in Table 9.

Table 9. Hyperparameter Search Space and Optimal Configuration for LightGBM

Hyperparameter	Search Space	Step Size	Optimal Value
n_estimators	50–1000	1	974
learning_rate	0.05–0.20	0.01	0.19
num_leaves	5~2*max_depth-1	1	11
max_depth	3–6	1	6
bagging_fraction	0.8–1.0	0.1	1.0
bagging_freq	2–4	1	2
feature_fraction	0.4–1.0	0.2	0.4

Following feature engineering and hyperparameter tuning, the LightGBM model is trained on the training set and validated on the testing set. The forecast accuracy of the model is described in Table 10.

Table 10. Performance of the LightGBM Model on the Training and Testing Sets

Dataset	RMSE	MAE	MAPE
Training	754.33	692.41	1.63%
Testing	1002.60	899.20	1.96%

As can be seen from Table 10, the model provides the RMSE equal to 754.33 on the training set and 1002.60 on the testing set, and the value of the testing MAPE stays as low as 1.96%. The small difference in the performance of the model on the training and testing sets is a good indication of the model generalization and absence of any signs of overfitting.

Moreover, as can be seen from the forecast results presented in Figure 7, the LightGBM model is capable of reproducing the dynamics of retail sales throughout most of the forecast horizon. Even in the months when there are quite high variations in retail sales, the values obtained from the forecast almost coincide with the actual dynamics, which means that the nonlinear dependencies in the data are reproduced well by the model. Small differences can be observed only in the regions of local maxima and minima.

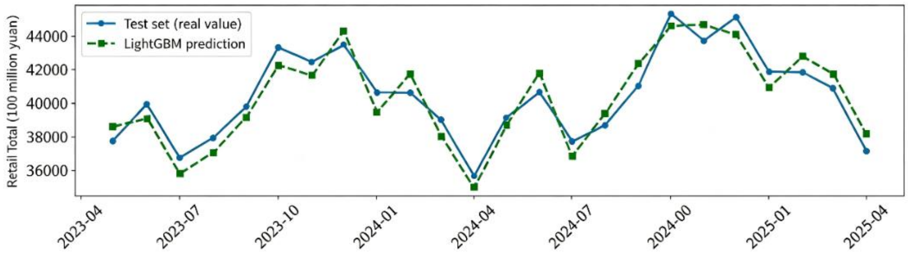

Figure 7. Forecasting Performance of the LightGBM Model

In summary, the LightGBM model has shown much more superior results than the conventional ARIMA model and has forecast accuracy that is close to the SARIMA model. Additionally, the capability to utilize a variety of macroeconomic indicators is another advantage of the proposed model.

3.4 Forecasting Analysis Based on the LSTM Model

3.4.1 Fundamentals of the LSTM Model

The Long Short-Term Memory (LSTM) network is an advanced form of the RNN that has been developed to model long-range temporal dependencies within sequences [12]. From Figure 8 below, the main components of the LSTM architecture include three gating mechanisms—the forget gate f_t, input gate i_t and output gate o_t —as well as the hidden state and memory cell that control the information retention, updating, and output processes. At each time step, the hidden state does not depend only on the current input x_t, but also on the previous hidden state and memory cell. Therefore, LSTM has an additive recurrent architecture with gated information processing [13].

Unlike the traditional RNNs, the LSTM network can effectively deal with the issues of gradient vanishing and gradient explosion, thus making it an ideal choice for modeling long range dependent time series data.

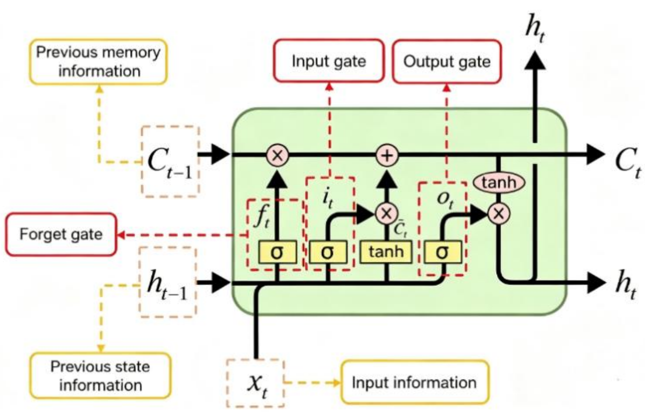

Figure 8. Architecture of a Single LSTM Cell

From Equation (5), it can be seen that the forget gate controls how much information in the previous memory cell will be preserved; the input gate adjusts the effect of the new information on the current memory; and the output gate controls the formation of the hidden state for generating predictions [14].

$$\begin{aligned} i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \odot c_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \odot c_{t-1} + b_f\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \odot c_t + b_o\right) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{5}$$

where $\odot$ a is the Hadamard (element-wise) product.

3.4.2 LSTM Model Development and Empirical Analysis

A fixed-length sliding-window strategy is adopted to reconstruct the original time series into supervised learning samples. Specifically, an input window of length window_size and a one-step forecasting horizon are employed, with the window moving along the time axis using a stride of one observation. After Min–Max normalization, this procedure generates a complete set of input-output sample pairs.

To evaluate out-of-sample forecasting performance, the final 24 samples are reserved as the testing set, while the remaining observations are used for model training.

The LSTM network architecture and training configuration are summarized in Table 11.

Table 11. Hyperparameter Settings of the LSTM Model

No.	Hyperparameter	Description	Value
1	lstm_cell	Number of LSTM units	64
2	activation	Activation function	tanh
3	window_size	Input window length	6
4	input_shape	Input dimension	(input_length, 6, 1)
5	optimizer	Optimization algorithm	Adam
6	loss	Loss function	MSE
7	learning_rate	Learning rate	0.001
8	batch_size	Batch size	5
9	epoch	Number of training epochs	100

The LSTM model used here includes only one LSTM layer. There are four time windows $\{4,6,8,12\}$ available, and the optimal $window\_size = 6$ is selected through calling the basic model LSTM model. The whole training configuration is listed in Table 11.

In order to test the effectiveness of the designed LSTM model for monthly retail sales forecast, the performance measures such as RMSE, MAE, and MAPE will be used for measuring the prediction accuracy of the model on both training and testing data sets. The prediction performance of the model will be reported in Table 12.

Table 12. Performance of the LSTM Model on the Training and Testing Sets

Dataset	RMSE	MAE	MAPE
Training	771.33	695.64	1.57%
Testing	540.64	536.29	1.32%

From Table 12, it can be seen that the RMSE, MAE, and MAPE of LSTM model are 771.33, 695.64, and 1.57% on the training data set, respectively. These measures become even better on the testing data set, reaching 540.64, 536.29, and 1.32%, respectively. It can be concluded that there is no overfitting phenomenon in the model due to the similarity between errors on the training and testing data sets.

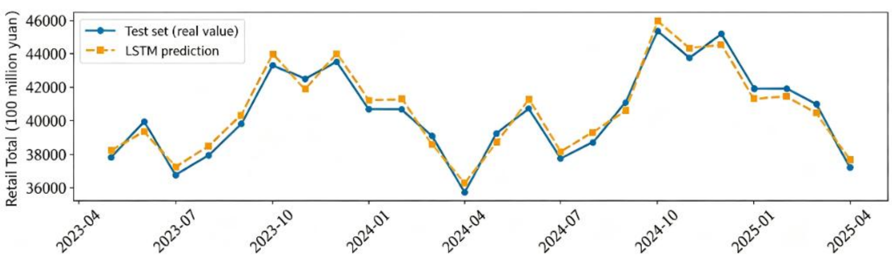

Figure 9. Forecasting Performance of the LSTM Model

As shown in Figure 9, the forecast curve of LSTM model well follows the real retail sales data in the testing period. Especially, it can well reflect the obvious seasonality and the long-term trend.

3.5 Comprehensive Comparison of Forecasting Models

In order to assess the performance of the competing methods, the ARIMA, SARIMA, LightGBM and LSTM models will be analyzed with respect to forecasting accuracy, robustness, and the ability to capture long-term temporal patterns.

Table 13. Overall Performance Comparison of the Forecasting Models

Model	RMSE (Train)	RMSE (Test)	MAE (Train)	MAE (Test)	MAPE (Train)	MAPE (Test)
ARIMA(2,1,1)	—	4490.06	—	3899.55	—	9.34%
SARIMA(1,1,0)(1,0,1)$_{12}$	—	978.11	—	794.70	—	1.91%
LightGBM	754.33	1002.60	692.41	899.20	1.63%	1.96%
LSTM	771.33	**540.64**	695.64	**536.29**	1.57%	**1.32%**

According to Table 13, the conventional ARIMA model performs the worst in terms of forecasting accuracy, obtaining the testing MAPE of 9.34% owing to its inability to incorporate the prominent seasonal and cyclical nature of the retail sales data.

Since it takes into account the seasonal factors, the SARIMA model shows better results, lowering the testing MAPE to 1.91%. This indicates the necessity to include seasonal trends into the analysis of monthly macroeconomic indicators.

In comparison to the other models, the LightGBM model based on machine learning algorithms has a better fitting performance on the training data set. Nevertheless, its testing performance is somewhat worse than that of the SARIMA and LSTM models, obtaining the testing MAPE of 1.96%.

Out of all the alternative models considered, the LSTM model provides the most accurate forecast results. It yields the lowest RMSE value of 540.64, MAE of 536.29 and MAPE of 1.32% on the test set, which means that it is more efficient at capturing nonlinearity and long dependencies of the considered economic series.

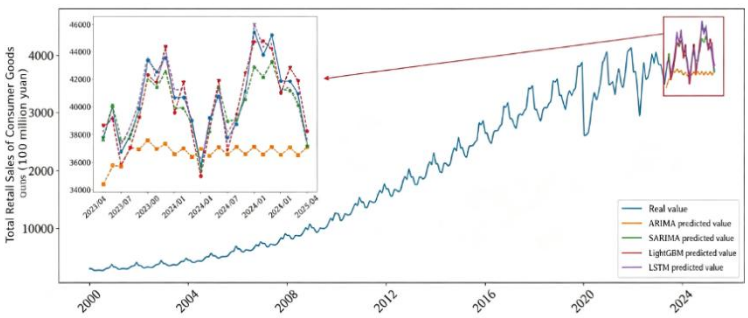

Figure 10. Comparison of Actual Values and Model Forecasts over the Final 24 Months

The graphical comparison provided in Figure 10 proves these claims. The forecast trajectories produced by both the LSTM and SARIMA models follow the actual values rather closely, especially during the episodes of high volatility, like in late 2023 and late 2024. Their ability to reproduce the local maxima and minima is much better than those of the ARIMA model, while the LightGBM model experiences some deviations during several sub-periods. Hence, the empirical analysis has shown that the LSTM model provides the most accurate forecasts of monthly consumer goods retail sales in China.

4. Conclusion

The current paper presents the systematic examination of the forecast of China's Total Retail Sales of Consumer Goods (RSCG) on the basis of monthly data from January 2000 to April 2025, applying both conventional statistical methods and contemporary machine learning algorithms. In the course of exploratory data analysis, descriptive statistics show that there is an obvious upward trend, significant cyclical movements, and a somewhat right-skewed distribution of RSCG. Seasonal-Trend Decomposition using Loess (STL) demonstrates that there is a considerable impact of major consumption periods, such as Chinese Spring Festival and mass e-commerce promotions, on the seasonality of the series. Moreover, through the application of the Augmented Dickey–Fuller (ADF) test for unit roots and first-order differencing, it becomes possible to make the series stationary.

In order to conduct comparative analysis of other approaches to forecasting, we will compare three models from different model classes: ARIMA and SARIMA statistical models; LightGBM – gradient boosting machine learning algorithm; and LSTM – deep learning approach to forecasting. Performance of all three models on test data set will be quantitatively compared by using three commonly used measures of forecast accuracy: RMSE, MAE, and MAPE. Empirical results show that the ARIMA model performs the worst in the group of competing models, with MAPE value equal to 9.34% on test data set. It is explained by the fact that ARIMA model cannot incorporate the highly seasonal nature of the sales time series. In turn, by including seasonal autoregressive and moving average components, the SARIMA model significantly increases forecasting accuracy by reducing testing MAPE to 1.91%. The LightGBM model additionally incorporates the influence of multiple macroeconomic factors as explanatory variables. Despite the fact that its testing MAPE value is equal to 1.96% (slightly higher than SARIMA's one), this model exhibits significant flexibility in incorporating relationships between macroeconomic variables and consumption dynamics.

In all the different models compared, it is the LSTM model that always comes out with the best forecasting performance. This is because this model has the lowest testing MAPE value of 1.32% while performing well in terms of RMSE and MAE. This model is capable of capturing long-term temporal dependence, nonlinearity and seasonality of the retail sales time series. In particular, this model has a very good ability to detect turning points and cope with sudden change in the underlying time series data, indicating the robust nature of this model for macroeconomic forecasting.

Generally, the empirical results imply that the SARIMA model is very effective in forecasting situations where there is stability and interpretability in the underlying time series due to its strong statistical interpretability and robustness. In contrast, the LSTM model will perform well in forecasting where there is complexity in terms of the nonlinear dynamics and structural changes.

References

[1] Liu, S. Y. Total retail sales of consumer goods in the city exceeded RMB 48.5 billion in the first quarter [N]. *Gannan Daily*, May 28, 2025 (in Chinese).

[2] Cheng, K. M., Liu, S. C., & Lei, L. Forecasting total retail sales of consumer goods based on web search data and deep neural networks[J]. *Operations Research and Management Science*, 2024, 33(12): 203–209 (in Chinese).

[3] Nakamura, E. Inflation forecasting using a neural network[J]. *Economics Letters*, 2004, 86(3): 373–378.

[4] Bu, P. D. Forecasting total retail sales of consumer goods based on web search data[D]. Chengdu: *Southwestern University of Finance and Economics*, 2022 (in Chinese).

[5] Dong, M. R. Research on CPI forecasting based on machine learning theory[D]. Beijing: *University of International Business and Economics*, 2022 (in Chinese).

[6] He, C. R. Empirical study on RMB exchange rate forecasting based on ARCH and XGBoost models[D]. Jinan: *Shandong University*, 2021: 36–37 (in Chinese).

[7] Jia, M. Q., Cai, Z. J., Hu, J., et al. Research on grain yield forecasting models based on machine learning[J]. *Journal of Hebei Agricultural University*, 2021, 44(3): 103–108 (in Chinese).

[8] Zhang, Y. Analysis and prediction of influencing factors of total retail sales of consumer goods in China: Based on multiple linear regression models and ARIMA models[C]//*Proceedings of the 4th International Conference on Global Economy and Business Management (GEBM 2022)*. Wuhan Zhicheng Times Cultural Development Co., Ltd., 2022: 13.

[9] Wei, S. Y. Forecasting China's GDP based on time series models[D]. Jinan: *Shandong University*, 2023 (in Chinese).

[10] Xu, M. Y. GDP forecasting for Jiangsu Province based on ARIMA and BP neural network models[D]. Jinan: *Shandong University*, 2020 (in Chinese).

[10] Yang, Y. P. Analysis and forecasting of factors influencing total retail sales of consumer goods[D]. Beijing: *Beijing University of Technology*, 2023 (in Chinese).

[12] Wang, Z. J., & Ning, Z. Y. Forecasting China's employment trends based on intervention time series models[J]. *Statistics and Management*, 2023, 38(6): 13–23 (in Chinese).

[13] Yıldırım, D. C., Toroslu, I. H., & Fiore, U. Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators[J]. *Financial Innovation*, 2021, 7(1): 1–20.

[14] Wang, J. N. Research on the application of combined forecasting models to China's total retail sales of consumer goods[D]. Lanzhou: *Lanzhou Jiaotong University*, 2018 (in Chinese)