Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 7, No. 2, 2023
www.scholink.org/ojs/index.php/asir

Original Paper

Contrasting Univariate and Multivariate Time Series Forecasting Methods for Sales: A Comparative Analysis

Feng Wang[1] & Joey Aviles[1*]

[1] Graduate School, Angeles University Foundation, 2009 Angeles City, Philippines

[*] Corresponding author, Joey Aviles

Received: April 26, 2023 Accepted: May 17, 2023 Online Published: May 19, 2023

doi:10.22158/asir.v7n2p127 URL: http://doi.org/10.22158/asir.v7n2p127

Abstract

In commodity-based industries, accurate sales forecast is very important for effective inventory management and decision-making. Univariate and multivariate time series forecasting methods have been widely used to predict commodity sales. The purpose of this study is to make a comprehensive comparative analysis of these two methods under the background of commodity sales forecast. Firstly, the concept of time series prediction and its significance in the field of commodity sales are introduced. It emphasizes the challenges related to sales forecast, including demand fluctuation, seasonality and external factors affecting sales model. The forecasting methods of univariate time series, such as ARIMA, are discussed in detail, focusing on their ability to capture the time correlation in a single sales variable. In contrast, the multivariate time series prediction method considers the relationship between multiple variables. Vector autoregressive and multivariate extension of ARIMA. These techniques combine the interaction of various factors, including external influences, to improve the accuracy of prediction. In order to make a comprehensive comparative analysis, a data set including historical sales data of specific commodities is used. Both univariate and multivariate models are suitable for forecasting future sales, and their performance indicators are evaluated by MASE. The results show that although univariate models are easier to implement and explain, they often fail to capture the complex interdependence between different factors that affect sales. On the other hand, the multivariate model shows excellent prediction accuracy by integrating related variables and their dynamic relationships. However, they need more additional data and more complex modeling techniques. Finally, the research gives the practical significance and suggestions of choosing an appropriate forecasting method based on the characteristics of commodity sales data and the specific business environment. It emphasizes the importance of considering the accuracy and interpretability of the prediction model in practical application.

Keywords

time series, machine learning, sales, algorithm, forecast

1. Introduction

Time series forecasting is a popular statistical technique used to predict future values based on past data points. In the context of business, time series forecasting can be used to predict the sales of a product over time. By analyzing past sales data, businesses can make informed decisions about future production and inventory management, which can help them maximize profits and reduce waste. One important application of time series forecasting is in predicting the sales of products in retail stores. Retailers need to have accurate forecasts of their product demand in order to stock the appropriate amount of inventory to meet customer demand. By using time series forecasting, retailers can analyze historical sales data and make predictions about future sales trends, allowing them to make informed decisions about inventory management and production. There are many different techniques that can be used for time series forecasting, including autoregressive integrated moving average (ARIMA), exponential smoothing, and neural networks. These techniques differ in their complexity and the types of data they are best suited to analyze. ARIMA models are commonly used for time series forecasting and are well-suited to analyzing data with a trend and/or seasonal pattern. Exponential smoothing models, on the other hand, are simpler than ARIMA models and are typically used for data with no clear trend or seasonal pattern. Neural networks are more complex than both ARIMA and exponential smoothing models and are often used for data with a high degree of complexity, such as financial time series data. To perform time series forecasting, historical sales data is first collected and analyzed. This data is then used to train a forecasting model, which can be used to make predictions about future sales. The accuracy of the model can be assessed by comparing its predictions to actual sales data. In addition to predicting future sales, time series forecasting can also be used to identify trends and patterns in sales data. For example, a business may use time series forecasting to identify the products that are most likely to sell well during a particular season or to identify the factors that are most strongly correlated with changes in sales. To account for external factors in time series forecasting, a common approach is to incorporate them into the forecasting model. This can be done by using a technique known as regression analysis, where external variables are added to the model to improve its accuracy. For example, if a retailer is trying to predict sales for a particular product and they know that sales are affected by the weather, they may incorporate historical weather data into the forecasting model. They may also consider other external variables such as changes in consumer preferences, economic conditions, or competitor actions. Another approach to incorporating external factors is to use a technique called causal forecasting. This approach involves identifying the causal relationship between external variables and the time series data. By understanding these relationships, it is possible to make more accurate predictions about future trends. When predicting sales or other time series data, external factors can have a significant impact on the accuracy of the forecasts. These external factors may

include changes in the economy, consumer preferences, seasonality, weather patterns, or events like holidays or special promotions. In some cases, it may be difficult to identify or measure the impact of external factors on time series data. In these situations, businesses may use a combination of statistical models and expert judgement to make predictions. It is important to note that incorporating external factors into time series forecasting models can add complexity to the analysis. In some cases, it may also require additional data sources and analytical tools. However, by taking external factors into account, businesses can make more accurate predictions about future trends and adjust their strategies accordingly.

In conclusion, time series forecasting is an important tool for businesses that need to predict the sales of their products over time. By analyzing historical sales data and using statistical techniques to make predictions about future sales, businesses can make informed decisions about inventory management and production, helping them to maximize profits and reduce waste.

2. Literature Review

The majority of research studies on sales volume forecasting using time series algorithms rely on a single living time series algorithm. Scholars emphasize the importance of accurately forecasting commodity sales volume for enterprises, as it can impact financial costs and customer satisfaction. Time series algorithms are commonly used for forecasting sales volume, primarily relying on the seasonal fluctuations of sales volume (ramosal et al., 2022). To improve the accuracy of forecasting, time series algorithms can be combined with other algorithms (Rasim et al., 2018). The growing popularity of e-commerce has enabled the use of computer technology for sales volume forecasting, although it is essential to consider multiple influencing factors. The nonlinear graphs generated by time series data can be effectively modeled using the ECS-NARNN model (Li et al., 2018). With the increasing popularity of online transactions in China, sales volume forecasting has become more important, especially in the hotel industry where online sales are the basis of revenue management. Machine learning algorithms are frequently applied in the field of time series prediction, with models ranging from traditional methods to advanced techniques such as long-term and short-term memory networks, closed-loop neural networks, and convolutional neural networks (Duan, 2020). The research results predominantly focus on time series algorithms, and the prediction results can be linear, nonlinear, single-factor, or multi-factor. Improving the accuracy of forecasting involves combining different algorithms based on various dimensions.

3. Method

This study utilized the autoregressive integrated moving average (ARIMA) model from time series analysis to forecast the sales volume of a particular commodity. By utilizing differencing, the ARIMA model is used to model non-stationary time series data, where the resulting stationary differenced time series can be modeled using autoregressive moving average. Unlike other models, the ARIMA model

does not consider specific trends in the historical data of the predicted sequence, but instead uses an iterative approach to identify potential models and validate them based on historical data to ensure an accurate representation of the sequence. The model consists of three parts: autoregressive (AR), differencing (I), and Moving Average (MA) (Li et al., 2018).

To establish the model, the time series must be stationary. If it fails the stationarity test, it must be transformed into a stationary series through differencing, known as the integration order. The differencing process is crucial in the model, and a smooth time series data is necessary for using the ARIMA model for forecasting. Unstable data will not accurately capture the time series model.

Compared to the autoregressive moving average (ARMA) model, the ARIMA model is better suited for non-stationary time series modeling. Using the ARIMA model for forecasting requires only one endogenous variable, making the forecasting process straightforward. However, its drawback is the need to manipulate the data after differencing to achieve stationarity (Bousqaoui et al., 2021).

Overall, this study employed a rigorous approach to analyze and experiment with the original data to obtain the expected results. The application of the ARIMA model in time series analysis is a mature and respected method, providing valuable insights for predicting business decisions.

3.1 AR Model

To describe the relationship between current and past observations in an autoregressive model, it is necessary for the model to demonstrate stability. This is typically achieved by specifying an equation for a pth-order autoregressive process (Yoo & Maddala, 1991).

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \quad (1)$$

In autoregressive modeling, the present value of a time series is represented by y_t, where u denotes the constant term. The order of the model is denoted by p, which signifies the number of past observations incorporated in the model. The autocorrelation coefficient is represented as y_i, while the residual term is referred to as et, and is typically assumed to follow a white noise process.

Expanding an equation:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + u_t \quad (2)$$

If the random disturbance term adheres to a white noise process, the autoregressive (AR) model is known as a pure AR(p) process, which is represented as follows:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \quad (3)$$

The autoregressive (AR) model utilizes past data to make predictions about future values, as represented by the equation where p signifies the number of lagged observations incorporated in the model. However, this technique has certain limitations. Firstly, the AR model is dependent solely on its own data for predictions. Secondly, the time series data must exhibit a certain degree of smoothness. Additionally, the method requires the presence of correlation and is not recommended when the autocorrelation coefficient is less than 0.5. Finally, the AR function is limited to forecasting phenomena that are linked to the time series' prior period (i.e., autocorrelation of the time series).

3.2 MA Model

The sliding average model is a crucial element of the autoregressive model, which emphasizes the accumulation of error terms. When the random disturbance term (u_t) in the model deviates from a white noise process, the resulting series (y) is typically considered a moving average of order q.

$$u_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \quad (4)$$

The white noise sequence represents a random process ε_t with uncorrelated values.

The moving average (MA) model is obtained when the current values of a time series are determined solely by a linear combination of past white noise values, with no correlation between present and past observations. In contrast, the autoregressive (AR) model captures the impact of past white noise on present forecast values indirectly, by affecting the series values over time. The equation for the MA model is specified as follows:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \quad (5)$$

Using a moving average for forecasting can effectively eliminate random fluctuations in the time series (Slutzky, 1937).

3.3 ARMA Model

The combination of autoregressive model of order p (AR(p)) and moving average model of order q (MA(q)) results in the autoregressive moving average model (ARMA (p, q)) (Naher et al., 2022).

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \quad (6)$$

3.4 ARIMA Model

The ARIMA model incorporates both the AR and MA models, in addition to the differencing method (denoted by "d") to ensure data stationarity. The resulting model is denoted as ARIMA (p, d, q), where "p" and "q" represent the orders of the AR and MA models, respectively. By combining these components, a differential autoregressive moving average model is constructed for predicting future values of a time series. This process involves using the differencing method to transform non-stationary data into a stationary form, which is then used to estimate the parameters of the ARMA model. The resulting ARMA model can be utilized for making predictions about future values of the time series (Batool & Tian, 2021).

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots \beta_q \varepsilon_{t-q} \quad (7)$$

3.5 Auto ARIMA

An algorithm used for automatic selection of parameters for an ARIMA model in time series forecasting. It is a classical time series model composed of autoregressive (AR) component, integrated (I) component, and moving average (MA) component.

The purpose of Auto ARIMA is to automatically select the optimal parameters for the ARIMA model based on the input time series data, enabling accurate time series forecasting. It evaluates the performance of multiple candidate models and chooses the model with the minimum information criterion (such as AIC, BIC) or minimum residual sum of squares (RSS) as the best model.

The specific steps of the Auto ARIMA algorithm are as follows:

For a given time series data, it first applies differencing (if needed) to make it a stationary time series.

It tries different combinations of ARIMA model parameters and selects the best model based on model evaluation criteria (such as AIC, BIC, or RSS).

Once the best model is determined, the algorithm trains and forecasts the time series based on the optimal model parameters.

Auto ARIMA does not have a specific formula but rather involves evaluating and selecting models based on different parameter combinations. The selection of the best model is typically based on minimizing the values of evaluation criteria such as AIC, BIC, or RSS, which are calculated based on the model's goodness of fit and complexity.

Auto ARIMA algorithm simplifies the process of model selection in time series forecasting by automatically choosing the optimal ARIMA model parameters and provides accurate prediction results. It is a convenient and powerful tool, particularly suitable for cases where there is no clear basis for model selection or when automated model selection is needed.

3.6 MASE

MASE (Mean Absolute Scaled Error) is a commonly used metric for evaluating the accuracy of predictions, particularly in the context of time series data. Its purpose is to measure the relative magnitude of forecast errors compared to a benchmark forecast, providing a means to assess the relative performance of forecasting models (Bean, 2023). The formula for MASE is as follows:

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - \hat{y}_i - 1|}\right) \quad (8)$$

- y_i represents the actual observed values (the ground truth time series).
- $\hat{y}_i$ represents the forecasted values.
- $|y_i - \hat{y}_i|$ denotes the absolute error between the actual observed values and the forecasted values.
- $|y_i - \hat{y}_i - 1|$ represents the absolute error between consecutive actual observed values.
- n is the number of data points in the time series.

A smaller value of MASE indicates a higher accuracy of the forecasting model. The key advantage of MASE is its ability to compare the predictive performance of different time series by considering relative errors, thereby mitigating the impact of data scale. Furthermore, MASE facilitates comparisons and interpretations of forecasting results with other time series forecasting metrics, making it easier to understand and contrast the outcomes of different forecasting models.

4. Result

4.1 Date Set

A sales dataset from a company includes seven data fields: Sale_Time, Season, Payment Method, Satisfaction, Weather, Discount, Price and Quantity. The dataset consists of a total of 730 records, and 700 of the data will be used for training the algorithm, while the remaining 30 will be used for testing. The goal is to forecast Quantity, which is the target field. Table 1 provides a sample of the data.

Table 1. A sales Dataset from a Company

Sale_Time	Season	Payment Method	Satisfaction	Weather	Discount	Price ($)	Quantity
2021/1/1	1	1	3	2	2	20.5	11
2021/1/2	1	1	3	2	1	20.5	9
…	…	…	…	…	…	…	…
2022/12/31	4	3	3	4	5	19	26

4.2 Exploratory Analysis

In Figure 1, the data in the data set can be detected. If there is no breakpoint in the graph, it means that the whole data is complete and there are no missing values. If there is a breakpoint, you need to delete or fill in the data of the breakpoint.

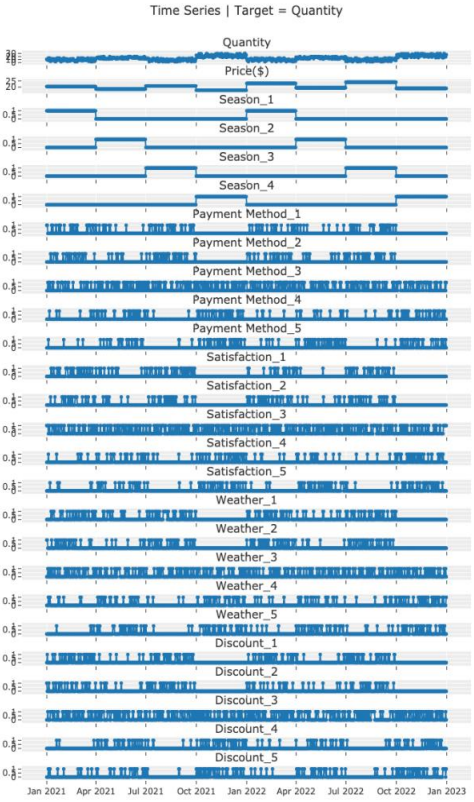

Figure 1. Check the Integrity of the Data

Through the analysis of Figure 2, it can be seen that the exogenous variables such as season _ 2, payment method _ 3, satisfaction _ 3, weather _ 3 and discount _ 3 have no strong correlation with labels. It can be removed when modeling.

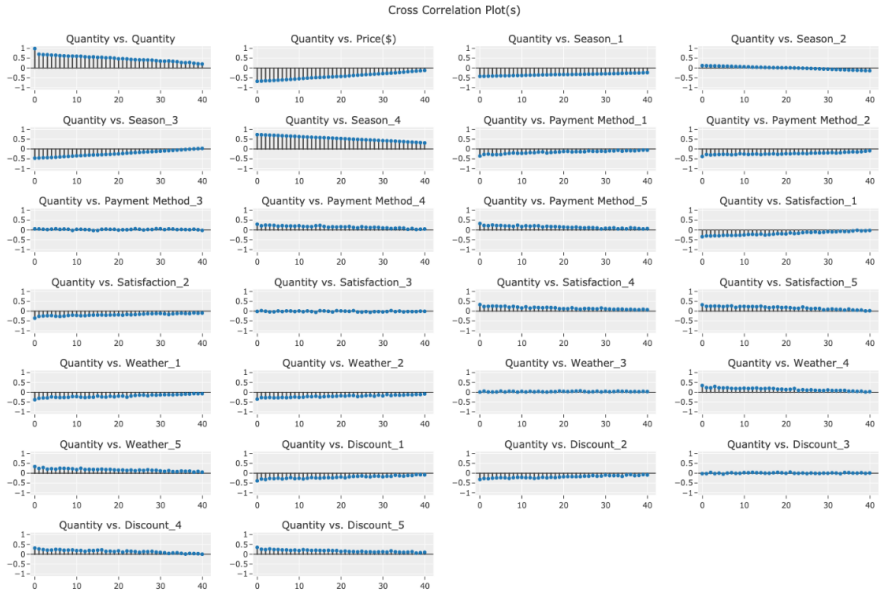

Figure 2. Correlation Test between Features and Labels

4.3 Compare Models

4.3.1 Univariate Forecasting without Exogenous Variables

The experimental steps first remove all exogenous variables, and then test the performance of the algorithm. Results It can be seen from Table 2 that the performance of dt_cds_dt algorithm is the best, and MASE is equal to 1.1182, which is greater than 1.

Table 2. Comparison of Algorithm Model Results (without Exogenous Variables)

Model	MAE	RMSE	MAPE	SMAPE	MASE	R2
dt_cds_dt	4.9779	6.3052	0.5500	0.4118	1.1182	-0.8319
grand_means	6.3411	7.3962	0.7392	0.4940	1.4244	-1.5244
…	…	…	…	…	…	…
croston	7.2890	8.1572	0.7331	0.5649	1.6376	-2.1758
…	…	…	…	…	…	…

Note. dt_cds_dt=Decision Tree w/ Cond. Deseasonalize & Detrending

grand_means=Grand Means Forecaster

It can be seen from Figure 3 that the blue line has surpassed the black line, indicating that the prediction deviation is relatively large.

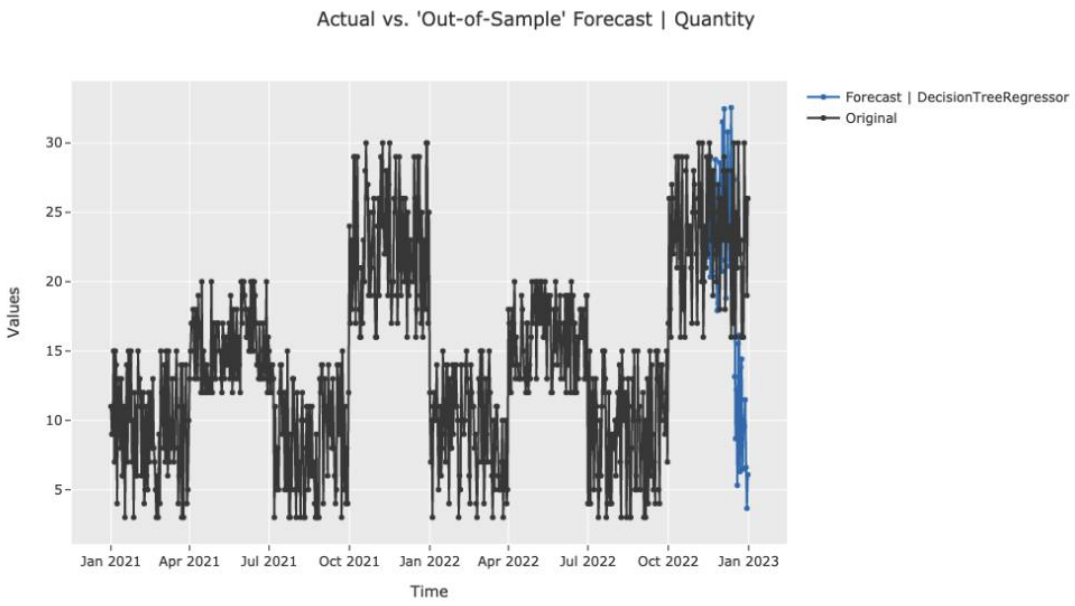

Figure 3. Univariate Forecast Trend

4.3.2 Univariate Forecasting with Exogenous Variables

This time, all exogenous variables are reserved, and then the performance of the algorithm is tested. Results it can be seen from Table 3 that the label is predicted after all exogenous variables are reserved, and it is found that the performance of auto_arima algorithm is the best, with MASE equal to 0.6231, which is less than 1.

Table 3. Comparison of Algorithm Model Results (with Exogenous Variables)

Model	MAE	RMSE	MAPE	SMAPE	MASE	R2
auto_arima	3.0183	3.5842	0.3568	0.2848	0.6231	0.4022
arima	3.7170	4.4749	0.4675	0.3501	0.7779	0.0353

From Figure 4, we can see that the blue line and the black line are basically consistent in the same area, which shows that the prediction result is more accurate.

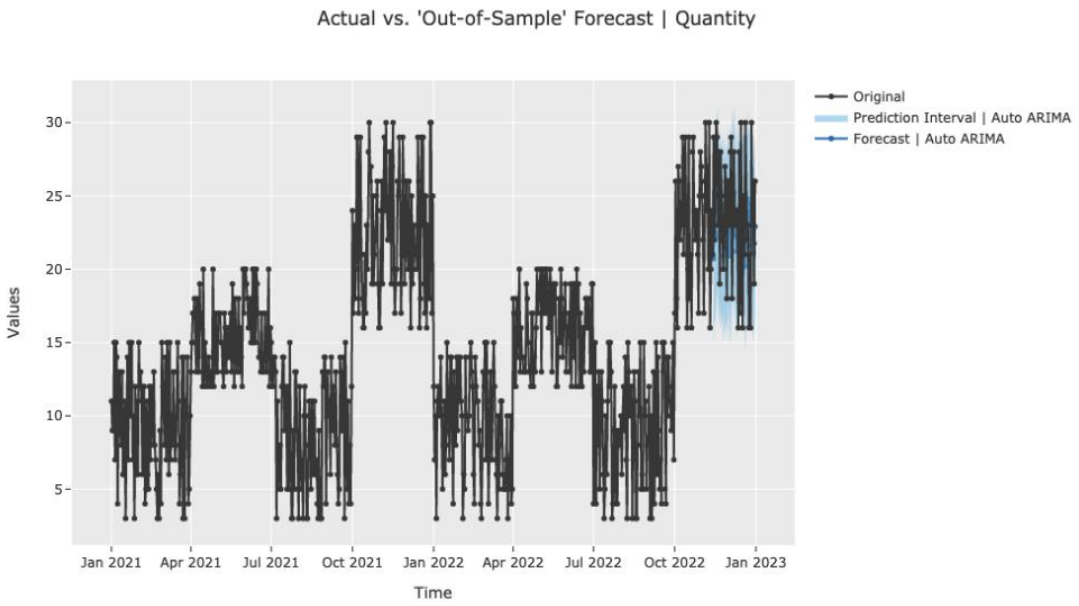

Figure 4. Multi-Exogenous Variables Forecast Trend

4.4 Make Future Predictions for the Target Variable

Comparing the above two situations, we choose the method of all exogenous variables to forecast the sales volume of goods, and forecast the data from January 1, 2023 to January 5, 2023, and get Table 4.

Table 4. Display the Sales Forecast Results from January 1, 2023 to January 5, 2023.

Sale_Time	Season	Payment Method	Satisfaction	Weather	Discount	Price ($)	Quantity
2023/1/1	1	2	3	2	2	24.5	13
2023/1/2	1	1	3	2	1	24.5	12
2023/1/3	1	2	4	3	4	24.5	10
2023/1/4	1	2	4	4	2	24.5	9
2023/1/5	1	1	3	3	2	24.5	14

Figure 5 is the predicted data trend.

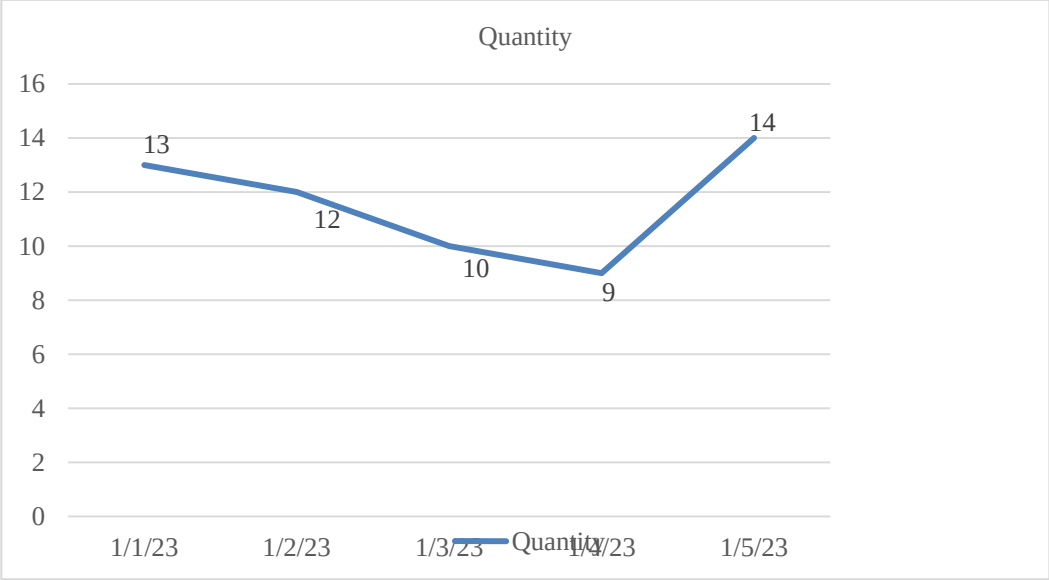

Figure 5. Trend Chart of Sales Forecast Results

5. Discussion

Single-variable and multivariate time series forecasting methods have their own advantages in commodity sales forecasting, and choose the appropriate method according to the specific situation.

- Univariate time series prediction;

Applicability: Single-variable time series forecasting is applicable to the situation that only the historical data of sales volume itself is considered, and the influence of other related factors or external variables on sales volume is not considered.

Advantages: The univariate time series model is simple and easy to implement, without additional data input or feature engineering.

Usage scenario: When the historical data of commodity sales have certain laws and trends, and external factors have little influence on sales, the single variable time series model can provide more accurate prediction results.

- Multivariable time series prediction;

Applicability: Multivariable time series forecasting is suitable for considering the influence of multiple related factors or external variables on sales volume, such as promotional activities, weather, competitor sales data, etc.

Advantages: Multivariable time series model can capture more influencing factors and predict sales more accurately. It can use the information of external variables to improve the prediction performance.

Usage scenario: When the sales volume of commodities is affected by many factors, which can be used as additional input data, the multivariate time series model can provide more accurate sales forecast results.

To sum up, univariate time series forecasting is suitable for simple sales forecasting scenarios, while

multivariate time series forecasting is suitable for complex sales forecasting scenarios, in which the influence of multiple related factors is considered. Choosing an appropriate method needs to consider factors such as data availability, prediction accuracy requirements and interpretability requirements.

References

Batool, H., & Tian, L. (2021). Correlation determination between covid-19 and weather parameters using time series forecasting: a case study in Pakistan. *Mathematical Problems in Engineering*, *2021*, 1-9. https://doi.org/10.1155/2021/9953283

Bean, R. (2023). Forecasting the monash microgrid for the ieee-cis technical challenge. *Energies*, *16*(3), 1050. https://doi.org/10.3390/en16031050

Bousqaoui, H., Slimani, I., & Achchab, S. (2021). Comparative analysis of short-term demand predicting models using arima and deep learning. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(4), 3319. https://doi.org/10.11591/ijece.v11i4.pp3319-3328

Duan, D. (2020). Research on hotel online sales forecast model based on improved wavenet. *Journal of Physics: Conference Series*, *1544*, 012067. https://doi.org/10.1088/1742-6596/1544/1/012067

Li, M., Ji, S., & Liu, G. (2018). Forecasting of Chinese e-commerce sales: an empirical comparison of arima, nonlinear autoregressive neural network, and a combined arima-narnn model. *Mathematical Problems in Engineering*, *2018*, 1-12. https://doi.org/10.1155/2018/6924960

Naher, S., Rabbi, F., Hossain, Md. M., Banik, R., Pervez, S., & Boitchi, A. B. (2022). Forecasting the incidence of dengue in bangladesh—Application of time series model. *Health Science Reports*, *5*(4). https://doi.org/10.1002/hsr2.666

Ramosal, A., Nicolas, R., & Marino, W. (2022). Improving sales forecasting by combining key account managers' inputs and models such as sarima, lstm, and facebook prophet. *Journal of Applied Business and Economics*, *24*(6). https://doi.org/10.33423/jabe.v24i6.5715

Rasim, Junaeti, E., & Wirantika, R. (2018). Implementation of automatic clustering algorithm and fuzzy time series in motorcycle sales forecasting. *IOP Conference Series: Materials Science and Engineering*, *288*, 012126. https://doi.org/10.1088/1757-899x/288/1/012126

Slutzky, E. (1937). The summation of random causes as the source of cyclic processes. *Econometrica*, *5*(2), 105. https://doi.org/10.2307/1907241

Yoo, J., & Maddala, G. S. (1991). Risk premia and price volatility in futures markets. *Journal of Futures Markets*, *11*(2), 165-177. https://doi.org/10.1002/fut.3990110204