A Method for Sales Prediction Based on Time Series

Chen Chen [1]

chenchen_sppc@163.com

[1] Department of Information and Intelligent Engineering, Shanghai Publishing and Printing College, China

Abstract: This article explores the application of MA, ARMA, and ARIMA models in sales forecasting based on the method of time series analysis. By crawling and cleaning data, the training results of the data model can be better ensured. In sales forecasting, time series analysis can be used to study the patterns and trends of historical sales data, and make predictions for future sales based on this. Time se-ries analysis is a widely used statistical method for data analysis and prediction. By training and testing object data groups, algorithm prediction of sales data is achieved. Experiments have shown that ARIMA has better prediction perfor-mance than MA and ARMA.

Keywords: Sales Prediction; ARIMA; Time Series

1. INTRODUCTION

With the continuous development of internet e-commerce and big data tech-nology, data has become an important resource for enterprise decision-making and operation, and more and more enterprises are paying attention to using data to im-prove sales performance. With the continuous development of internet e-commerce and big data technology, data has become an important resource for enterprise deci-sion-making and operation, and more and more enterprises are paying attention to using data to improve sales performance. Sales forecasting, as an important part of enterprise management, is of great significance for the operation and management of enterprises. A sales forecasting system is a system that predicts future sales by ana-lyzing historical sales data and market trends. It can help enterprises better formulate sales strategies and improve sales performance.

In the current economic environment, the competition in the global market is becoming increasingly fierce, and enterprises need more precise and accurate sales forecasts to cope with changes in market demand and challenges from competitors. Traditional forecasting methods based on experience and intuition can no longer meet the needs of enterprises. Therefore, it is necessary to use big data analysis and artificial intelligence technology to improve the accuracy and reliability of sales forecasting[1].

The sales prediction based on time series analysis designed in this article can not only use historical sales data and market trends to predict future sales, but also com-bine multiple data sources and analysis methods, such as user behavior data, social media data, and competitor data, to improve the accuracy and practicality of the prediction. Compared to traditional experience based sales forecasting, using time series to predict sales data has higher accuracy

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334677

and stronger practicality, which can better help enterprises formulate sales strategies, improve sales performance, and enhance their competitiveness[2].

In terms of research methods, this study will adopt the method of empirical research, and conduct experiments and analysis based on real sales data. First, obtain real and reliable sales data through data collection and cleaning. Then, use time series models such as MA, ARMA and ARIMA to predict the sales data, and calculate the error indicators of the prediction results, such as RMSE, MAE and MAPE, etc. Evaluate the pros and cons and applicability of different models by comparing their predictive performance. In addition, this study will refer to the research results and methods in related fields, continuously improve and optimize the forecasting model, and improve the forecasting accuracy and reliability.

2. DATA COLLECTION AND PREPROCESSING

2.1 Data Collection

2.1.1 API interface

The API (Application Programming Interface) interface is an application programming interface that allows the system to interact with different data sources. Large e-commerce platforms, such as Tmall, JD.com, and Pinduoduo, all provide API interfaces for merchants to extract data. In a sales forecasting system, data from various third-party platforms, systems, or applications can be obtained through API interfaces. The data obtained using API interfaces usually has higher accuracy and real-time performance, but it is necessary to have a deep understanding of the usage rules and limitations of the interface.

2.1.2 Web Crawler.

Web crawler is an automated program that can grab a large amount of data from the Internet. In the sales forecasting system, web crawler can be used to obtain sales and market data from various websites, social media platforms, news media and other channels. The data that can be obtained through crawling includes but is not limited to product information, sales data, competitor information, market trends, etc. The mainstream crawler frameworks include pyspider and scraper, but it is im-portant to ensure that operations are carried out within the legal limits.[3]

2.2 Data preprocessing

2.2.1 Data cleaning

Data cleaning refers to deleting or repairing missing values, outlier, duplicate values and other problems in the dataset. Data cleaning can improve the quality of the da-taset and reduce the impact of errors on the analysis results[4].

2.2.2 Data Integration

Data integration is the process of aggregating data from different sources into a unified dataset. Data integration can expand the scope of the dataset and improve the accuracy of data analysis.

2.2.3 Data Transformation

Data transformation can transform raw data into a form suitable for analysis. For example, data can be normalized to convert different metrics into the same standard, enabling better comparison and analysis of the data.

2.2.4 Data Protocol

Data protocols reduce storage and computing costs by compressing data. For ex-ample, you can reduce the amount of data by sampling or clustering, but retain the basic characteristics of the dataset.

2.2.5 Data Integration

Data integration is the process of aggregating data from different sources into a unified dataset. Data integration can expand the scope of the dataset and improve the accuracy of data analysis.

These methods can be implemented through data mining tools, such as pandas, numpy, scipy and other libraries in Python, or commercial software, such as RapidMiner, SPSS and other tools. Choosing appropriate data preprocessing methods can improve the effectiveness of analysis and yield more accurate sales forecast re-sults.

3. DATA MODEL ESTABLISHMENT METHOD

3.1 Moving Average Method

The moving average method is a prediction method based on time series data, which calculates the average of historical data to predict future data for a period of time.

In sales forecasting, the moving average method can be used to predict future sales volume. Assuming we have a time series of sales data, we can use the moving average method to predict future sales volume.

The basic idea of the moving average method is to average the sales volume over a period of time and obtain an average value, which is used to represent the sales volume for a future period of time. In the moving average method, window size is an important parameter that determines the length of time period to consider when calculating the average.

The advantage of the moving average method is that it is easy to understand, implement, and explain. However, the moving average method also has some limitations. Firstly, it assumes that future sales have similar trends and periodicity to past sales, but in reality, market conditions and competitive environments may change, which can affect the accuracy of the prediction results. Secondly, the moving average method cannot consider the impact of other factors on sales volume, such as advertising investment and marketing activities.

There are two types of moving average methods: simple moving average method and weighted moving average method:

(1) Simple moving average method: Simple moving average method refers to predicting the average sales volume over a period of time. For example, if we use the sales volume of the

past 7 days for prediction, the predicted value is equal to the average sales volume of the past 7 days. When determining the window size for moving average, the selection can be based on the periodicity and trend of historical data.

(2) Weighted moving average method: Weighted moving average method refers to predicting the weighted average of sales volume over a period of time. For example, if we use the sales volume of the past 7 days for prediction, the predicted value is equal to the weighted average of the sales volume of the past 7 days, where the weight of the latest sales volume is greater and the weight of the previous sales volume gradually decreases. When determining the weight of the weighted moving average, it can be selected based on the importance of historical data.

Model fitting and prediction: Based on the selected moving average method and window size, we can fit the model and predict future sales. When making predictions, it is necessary to select an appropriate window size based on the periodicity and trend of historical data, and adjust the predicted values based on business experience.

3.2 Autoregressive Moving Average Model

ARMA model is a prediction method based on time series data, which combines the characteristics of autoregressive model (AR) and moving average model (MA), and can effectively solve the problems of autocorrelation and randomness in time series data. The following is an example of a simple ARMA model for sales forecasting:

(1) Firstly, we need to collect historical sales data, assuming we have already collected sales data for N months.

(2) For the ARMA model, we need to determine two parameters p and q, which represent the order of the autoregressive term and the moving average term, respectively. These two parameters can be determined through the autocorrelation function ACF and the partial autocorrelation function PACF.

(3) According to the principles of the ARMA model, we can obtain the expression of the ARMA (p, q) model is shown as Eq.1.

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (1)$$

y_t represents the sales volume in month t, and c is a constant term, $\phi_1 \sim \phi_p$ is the autoregressive coefficient, ε_t is the Error term of white noise, $\theta_1 \sim \theta_q$ is the moving average coefficient.

(4) Next, we need to perform parameter estimation on the model. Maximum likelihood estimation or least squares method can be used to estimate model parameters.

(5) By estimating the ARMA model, future sales can be predicted. For the prediction of time series data, recursive methods or Kalman filters can be used to calculate the predicted values at future times.

3.3 Autoregressive Integral Moving Average Model

The ARIMA model is a time series based prediction model that can be used to predict future time point values. The full name of ARIMA model is "Autoregressive Integrated Moving Average", which is a combination of autoregressive model (AR model), difference (I model), and moving average model (MA model)[5].

The basic principle of ARIMA model is to analyze and model time series data, and predict future data by fitting the pattern of historical data. The core idea of the ARIMA model is that there is a certain correlation between past and future time points, so past data can be used to predict future data.

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

d is the order of difference, y'_t represents the sales volume in month t after d-order difference processing, and c is a constant term, ε_{t-1} is the noise term (random error), ϕ_1~ϕ_p is the autoregressive coefficient of the AR model, θ_1~θ_q is the moving average coefficient of the MA model, and p and q are the orders of the AR and MA models.

Before establishing and predicting the ARIMA model, we need to perform stationarity testing and model order selection on time series data. The basic establishment steps are as follows:

(1) Stability test: Firstly, we need to perform a stability test on sales data, as the ARIMA model requires data that is a stationary time series. The stationarity test can be conducted through ADF (Augmented Dickey Fuller) test or KPSS (Kwiatkowski Phillips Schmidt Shin) test. If the data is not stable, we need to perform differential processing on it until we obtain stable time series data.

(2) Model order selection: We need to select the difference order, autoregressive terms, and moving average terms in the ARIMA model. The selection of model order can be done using autocorrelation function (ACF) and partial autocorrelation function (PACF). We can choose the appropriate order based on ACF and PACF graphs.

(3) Model fitting and diagnosis: By selecting the appropriate order, we can fit and diagnose the ARIMA model. Diagnosis can include testing the autocorrelation and normality of residuals. If the residual has autocorrelation or does not conform to the normal distribution, we need to adjust the model.

(4) Model prediction: Based on the established ARIMA model, we can predict future sales. The accuracy of the prediction can be determined by the confidence interval of the model prediction.

3.4 Model Selection

Choosing a suitable model requires comprehensive consideration based on the nature of time series data, prediction needs, and actual situations. Therefore, it is necessary to compare three methods:

(1) MA model: The moving average model uses the moving average value in the time series data to predict, and its basic idea is to build the prediction value of the future time on the linear combination of historical observations and white noise Error term. MA model only

considers the impact of white noise error at the current time on the future, so it is applicable to stable time series data, but not to non-stationary data.

(2) ARMA model: The autoregressive moving average model combines the char-acteristics of the autoregressive model and the moving average model. ARMA model is based on the autocorrelation and partial autocorrelation functions of time series data, and predicts future values by fitting the linear combination of historical obser-vations and white noise Error term. The ARMA model can be applied to stationary or weakly non-stationary time series data.

(3) ARIMA model: The autoregressive integral moving average model is an exten-sion of the ARMA model, incorporating the processing of non-stationary data. The ARIMA model performs differential processing on time series data, converts non-stationary data into stationary data, and then performs fitting and prediction on the ARMA model. The ARIMA model is suitable for strongly non-stationary time series data.

In summary, the MA model is suitable for predicting stationary time series data, the ARMA model is suitable for predicting stationary or weakly non-stationary time series data, and the ARIMA model is suitable for predicting strongly non-stationary time series data.

4. EXPERIMENTS

In this experiment, the data is sourced from sales data of a certain e-commerce platform store and has been desensitized. Train and predict through MA, ARMA, and ARIMA models, and compare the models.

Firstly, preprocess the data, including data cleaning and stabilization processing. Next, the dataset is divided into a training set and a testing set in a 7:3 ratio, where the training set is used to train the model and the testing set is used to evaluate model performance. The specific method is to use Python's Statsmodes package for model training.

Then, use MA, ARMA, and ARIMA models for training and predict the test set. The model was evaluated using three indicators: root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The smaller the values of the three indicators, the smaller the error.

For the same set of sales data, experiments using MA, ARMA, and ARIMA models for prediction yielded the following results.

Fig. 1 shows the root mean square error (RMSE) index values of each model: the result of the MA model is 107.02, the result of the ARMA model is 95.87, and the result of the ARIMA model is 89.71.

Fig. 2 shows the average absolute error (MAE) index value of each model: the result of MA model is 80.85, that of ARMA model is 73.78, and that of ARIMA model is 68.32.

Fig. 3 shows the average absolute percentage error (MAPE) index values of each model: the result of the MA model is 14.33%, the result of the ARMA model is 12.45%, and the result of the ARIMA model is 11.2%.

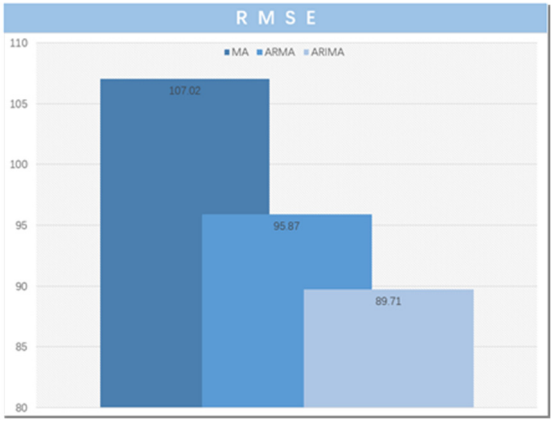

Fig. 1. - Comparison of RMSE Test Data

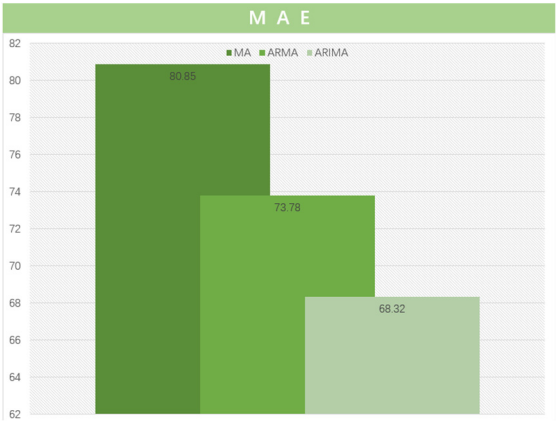

Fig. 2. - Comparison of MEA Test Data

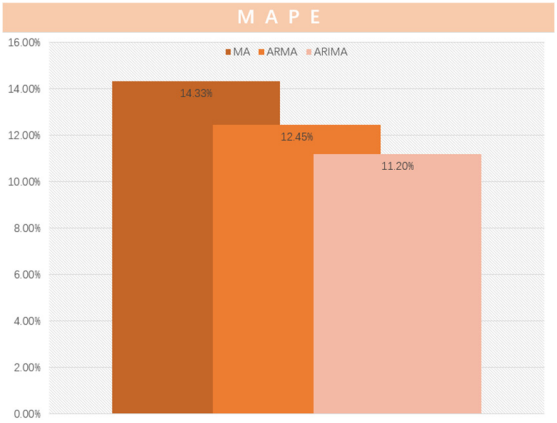

Fig. 3. - Comparison of MAPE test data

From the experimental results, it can be seen that the ARIMA model has the best prediction performance with the smallest error, followed by the ARMA model, which has the worst prediction performance. The RMSE, MAE, and MAPE indicators all reflect the conclusion that the ARIMA model has the highest prediction accuracy and the MA model has the lowest prediction accuracy. In practical applications, the ARIMA model can be considered for sales forecasting.

5. CONCLUSIONS

This experiment aims to compare the effects of using MA, ARMA and ARIMA three time series models for sales data forecasting, and analyze their advantages and disadvantages. During the experiment, the sales data was preprocessed firstly, including data cleaning and smoothing. Next, the dataset is divided into training set and test set for model training and evaluation. Finally, forecast sales data using MA, ARMA and ARIMA models and compare the forecasting performance of the models.

According to the experimental results, we draw the following conclusions:

First of all, the MA model is a simple moving average model, which is suitable for the prediction of stationary time series. However, in the prediction of sales data, the performance of the MA model is not ideal, and its prediction error is large, probably because the model cannot take into account the trend and seasonality of the time series.

Secondly, the ARMA model is an autoregressive moving average model, which is suitable for the prediction of non-stationary time series. In the prediction of sales data, the ARMA model has better predictive performance than the MA model, and can better capture the trend and correlation of time series. However, the ARMA model still has certain limitations, especially when faced with data with complex seasonality and trends, its forecasting accuracy may be limited.

Finally, the ARIMA model is based on the ARMA model and adds a difference operation to deal with non-stationary time series. The experimental results show that the ARIMA model performs well in sales data forecasting. Compared with the MA and ARMA models, its forecasting error is smaller and the forecasting accuracy is higher. Through the difference operation, the ARIMA model can transform the non-stationary time series into a stationary series, thus improving the predictive ability of the model.

The sales prediction method based on time series is a commonly used prediction method, which is suitable for situations where there are time series patterns in the data. This article introduces three basic time series models, MA, ARMA, and ARIMA, as well as their application methods in sales forecasting. The MA model is suitable for predicting stationary time series, the ARMA model is suitable for predicting non-stationary time series, and the ARIMA model combines the advantages of the two. It can differentiate non-stationary time series into stationary time series before making predictions. In the experiment, the ARIMA model had better predictive performance than the MA model and ARMA model.

It should be noted that the selection and application of time series prediction models need to consider multiple factors, including data characteristics, model applicability, prediction

effectiveness, etc. In practical applications, it is necessary to select the most suitable model based on the specific situation, and adjust and optimize the parameters of the model to obtain more accurate prediction results.

References

[1] Zhao, J., Xiong, F.W., Jin, P.Q., (2022) Enhancing Short-Term Sales Prediction with Microblogs: A Case Study of the Movie Box Office. Future Internet., 14(5):141. https://doi.org/10.3390/fi14050141

[2] Zhou, Y.H., Hu, H.P., Liu, Y.C., (2021) Analyzing User Behavior Patterns in Casual Games Using Time Series Clustering. In: Proceedings of the 2nd International Conference on Computing and Data Science. San Francisco. pp.387-397. https://kns.cnki.net/kcms2/article/abstract?v=YhL_Bl4XtC7yyLQqjQmWvQGFaHRks9Y7bg1BupiimUItghm96ASpca9rKclOYUYn_cnLOwfGGDIW2iZOUKf0hJF4JfWaubCmDhPWB-hK4nQ%3d

[3] Lin, G., Liang, Y.C., Tavares, A. (2023) Design of an Energy Supply and Demand Forecasting System Based on Web Crawler and a Grey Dynamic Model. Energies., 16(3):1431-1431. https://www.mdpi.com/1996-1073/16/3/1431

[4] Zhao, Y.F., Bian, L., Cong, X., (2017) Review of Data cleansing methods. Software Guide., 16(12):3. https://www.nstl.gov.cn/paper_detail.html?id=56b559e31eed0cef217692fd9cea9744

[5] Khulood, A., Bassam, Z., Ahmed, M. (2023) Time Series Forecasting using LSTM and ARIMA. International Journal of Advanced Computer Science and Applications, 14(1): 1. https://doi.org/10.14569/IJACSA.2023.0140133.