Comparison of Sales Models in the Post-epidemic Era

Xiaoyu Zhang[1,a,*], Ruibo Chen[2]

[1]*Department of Mathematics, University of Washington, Seattle, United States; Schmitz Hall, Box 355852, 1410 NE Campus Parkway, Seattle, WA, 98195*
[2] *Department of Engineering, University of Washington, Seattle, United States; Schmitz Hall, Box 355852, 1410 NE Campus Parkway, Seattle, WA, 98195*
a. Rayne6537@gmail.com
*corresponding author

Abstract: Compared with the traditional sales forecast based on human experience calculation and intuition prediction, combining existing sales data and model predictions can often accurately determine the market trend and reduce the probability of judgment error. A high degree of precision in market forecasting is essential for business leaders to develop policies and plans, increase store turnover and reduce operating costs. Based on the analysis of the historical sales data of China's liquor retail industry in the post-epidemic era, this study compared two existing forecasting models for forecast accuracy to determine the most suitable model for long-term sales forecasting. In this study, the sales data were collected from Suhe Bar Chain under the company of Alliance Art Group. The raw data set was screened and cleaned to suit each model, and then applied to Linear Regression Model (LRM) and Autoregressive Integrated Moving Average Model (ARIMA) to generate new forecast data. The predicted data were compared with the real sales data, and Root Mean Square Error (RMSE) was used to judge the accuracy of the model prediction. Finally, ARMIA is the better model to predict China's liquor retail in the post-epidemic era. According to the model predictions, reducing inventory, maintaining efficient cash flow, improving the turnover efficiency of goods, and strengthening the ability to adjust market strategy are the more suitable strategies for current liquor sales enterprises.

Keywords: Linear Regression Model, Autoregressive Integrated Moving Average Model, Sales Forecasting, Liquor Retail, Post-pandemic Era

1. Introduction

In the post-epidemic era, it refers to the first half of 2020 to 2022. During this period, the public's attitude toward the virus gradually switched from clearance to coexistence. In this context, the public will gradually restore economic development and lifestyle before the epidemic. However, in the meantime, there are still epidemic prevention and response methods emerging, making it difficult for the public to fully return to the pre-epidemic lifestyle [1]. The relatively slow economic development makes the public tend to be conservative in consumption. Citizens try to reduce economic pressure by reducing expenditures and increasing income. For the retail industry, after a short winter, there is still no sign of recovery. Under such a situation, every business decision should be given extra caution. For retail businesses, the forecast of sales volume and income is indispensable. An accurate prediction

can help enterprises better adapt to the current fast consumption market and improve profit. Compared to other studies, researchers mainly focus on short-term prediction and judge the impact of different subtle impact factors on short-term commodity sales. Jie et al. researched microblogs' sentiment and influence features' affection to short-term sales of movies [2]. They selected both the Linear Regression model and the Support Vector model for the prediction and evaluated the accuracy of the prediction [2]. Daumantas used 16 prediction models to forecast the price of salmon in the short-term [3]. Such a forecast could increase the net profit of a salmon farmer by a noticeable range [3]. This study chooses the liquor retail industry as the research object. But for the retail industry, accurate long-term forecasts and the search for sales trends are of the same commercial significance. However, as commonly used forecasting models are concerned, there are abundant models and the usage tends to be complicated. In addition, the accuracy of research of different models targeting certain industries is limited. Targeting China's liquor retail industry, the study aims to compare the accuracy of long-term liquor sales forecasts made by different data prediction models. Using the real sales data of the Suhe Bar Chain owned by Alliance Art Group, the study is of credential. The feasibility and accuracy of the prediction models are judged by the error analysis of the predicted value and the real value set for contrast. With a feasible and accurate long-term prediction model, the forecast of sales data will be of guiding significance to the entrepreneurs and workers in the Chinese liquor retail industry [4].

2. Data and Models

2.1. Data Description

The raw data of the study comes from the historical sales data of the Suhe Bar Chain under the Alliance Art Group. The data was collected from four stores: Suhe Chongqing store, Suhe Luohu store, Suhe Sanya store, and Suhe Haikou store. Among them, the stores in Chongqing and Haikou are located in second-tier cities. Second-tier cities are characterized by high living costs, high living pressure, and constant fluctuations in business policies due to the pandemic. The stores in Sanya and Luohu are located in third-tier cities, where the pace of life is relatively slow and businesses are less susceptible to business fluctuations caused by the pandemic [5]. In particular, the Suhe Sanya store is located in Sanya City, Hainan Province, which is a typical tourist city in southern China. The turnover will also fluctuate greatly due to the influence of the tourism season [6]. The selection of these four cities can well reflect the overall situation of China's liquor sales industry so that the conclusion based on this study can better predict the development of the industry and provide more accurate suggestions for the industry.

Initial sales data for the four stores are packaged in two documents, Sales_Data_2020_2021.xlsx and Sales_Data_2022.xlsx, respectively. This study will use the sales data of the first half of 2022 as the reference data set, and the sales data of 2020 and 2021 as the analysis data set.

Sales_Data_2020_2021.xlsx has a total of 190,294 data samples (composed of 6 columns * 190,295 rows of data) that contain daily sales information from January 1, 2020, to December 31, 2021, such as brand, sales shops, product name, quantity sold, income, and date. The data from January 24, 2020, to March 24, 2020, are missing. In response to the spread of the virus, China has called on businesses to suspend operations and entertainment venues not to open without permission. The four stores in Suhe are facing closure as a bar type of operation, so no sales data are produced at that time.

Sales_Data_2022.xlsx has a total of 31,498 data samples (composed of 6 columns*31,499 rows of data), which contains daily sales information from January 1, 2022, to June 14, 2022, such as brand, sales shops, product name, quantity sold, income, and date. Considering the creation time of the article, the complete sales data for 2022 is only from January to May, and the sales data for June is

only from the first half of the month, which is difficult to be used as an effective reference. So the control group data only includes data from January 2022 to May 2022.

2.2. Data Analysis

Starting from 2022's sales data as reference, there are seven categories of products sold: appetizers, beer, beverages, cocktails, red wine, imported wine, and services. The sales volume of each category is shown in Figure 1.
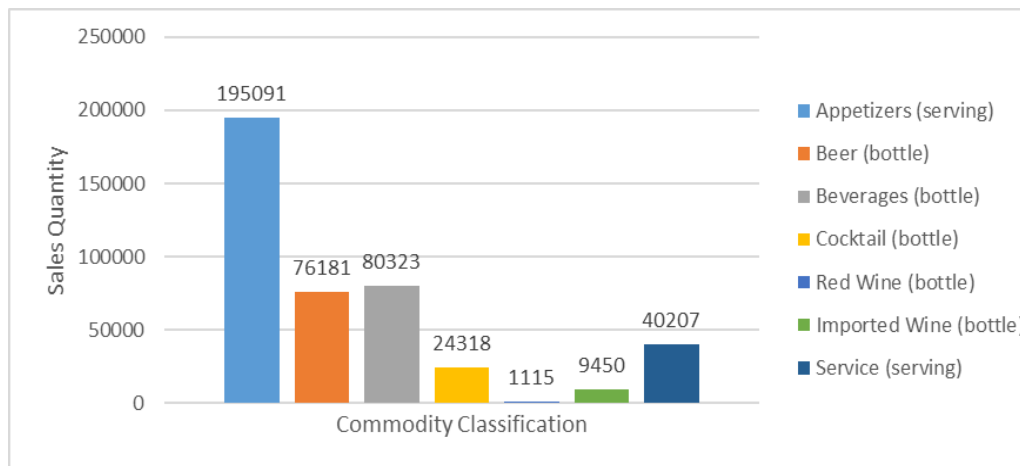

Figure 1: Sales of Suhe Bar Chain from 2020.01.01 to 2022.06.14.

Raw data contained a large portion of sales data for bar by-products, such as appetizers, beverages, and services. These data are not helpful to the research focus of this study. Given such problems, irrelevant data will be removed before analysis. Focusing on the liquor portion, the sales quantity of each category in the first half of 2022 is shown in Figure 2.
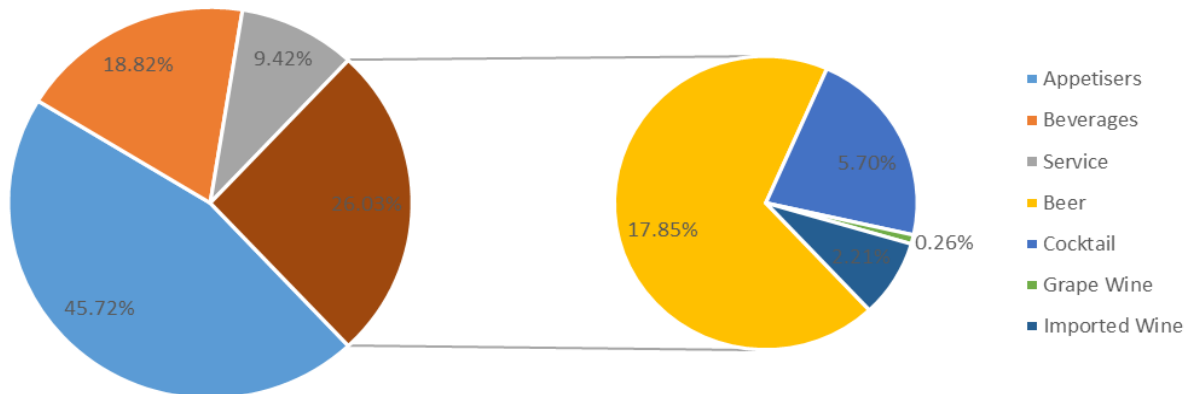

Figure 2: Percentage of Goods Sold by Category of Suhe Bar Chain from 2020.01.01 to 2022.06.14.

Focusing on 26.03% of the liquor sales category, there are five categories of liquor sold in Suhe stores. Considering that it is almost impossible for consumers to change their consumption habits in the short term, the study will assume that customers will constantly shop for their frequently purchased products. To ensure that there is data available for items selected across all three years, the study will focus on liquor products of the top 3 sales quantities of each category. For the imported wine category, the study chose Remy Martin, Remy Martin VSOP, and Dewar's 12-Year-Old. For the red wine category, the study chose Chateau-Margaux, Cabernet Sauvignon, Pago de cirsus vendimia seleccionada. For the beer wine category, the study chose Budweiser Beer, Corona Extra,

and Budweiser Anheuser-Busch. For the cocktail category, the study chose Power Station, Long Island Iced Tea, and See You Tomorrow (All names of products are listed in the order of sales quantity). Finally, 19219 data units from 2020 and 2021 were selected for the prediction of the model.

2.3. Data Preparation

Since the final recommendations of this study are for Suhe brands and the overall liquor retail industry in China, store perceptions will be ignored when applied to the model predictions. For each product, Excel was used to generate the corresponding. CSV file for the two prediction models. Because LRM uses the linear regression method for prediction, the data used for model learning can only include the data from 2020 and 2021. The ARIMA model requires continuous learning to correct the forecast. To ensure the number of outputs is equal, the data used will be enclosed data from January 2022 to May 2022 as well.

2.4. Linear Regression Model (LRM)

The prediction model of LRM is based on the linear change between the dependent variable and independent variable, which is often not the case in practical applications [7]. This, in turn, will lead to inaccurate predictions. In the linear regression model, the dependent variable will change with the change of the independent variable. As shown in Equation 1, k and b values will be calculated during the early learning stage. The k-value is the regression coefficient of the model, and the b-value is a constant. In this forecast, Y is the forecast value of goods sales, and X is the time value. The complete formula of LRM is given below:

$$Y = K \cdot X + b \tag{1}$$

When using the LRM model for prediction, the actual data is first imported into the LRM model. Data from January 2020 to May 2022 are drawn in the generated graph for comparison in the blue line. The LRM model was then used to derive the projected values for January to May 2022, which are plotted in the same graph in the red line. Table1 contains the LRM forecast values for the twelve products from January 2022 to May 2022. Figure 3 contains the corresponding graphs, where units for Sale Quantity are all bottles, and units for Date are the number of months after January 2020.

Table 1: Predicted values for twelve products using LRM (Unit: Bottles).

| Date	Budweiser Anheuser-Buschc	Budweiser Beer	Cabernet Sauvignon	Chateau-Margaux	Corona Extra	Dewars 12 Year Old
202201	456.59	11015.46	73.92	157.73	752.53	286.32
202202	456.02	11046.72	71.76	155.90	726.73	269.72
202203	455.46	11077.98	69.60	154.07	700.94	253.12
202204	454.90	11109.24	67.44	152.24	675.15	236.52
202205	454.34	11140.50	65.28	150.41	649.36	219.92
Date	Long Island Iced Tea	Pago de cirsus vendimia seleccionada	Power Station	Remy Martin VSOP	Remy Martin	See You Tomorrow
202201	1217.51	14.41	2846.02	336.42	675.65	230.00
202202	1184.86	14.40	2860.33	321.93	665.20	224.20
202203	1152.21	14.39	2874.64	307.45	654.74	218.39
202204	1119.56	14.37	2888.94	292.96	644.28	212.59
202205	1086.91	14.36	2903.25	278.47	633.82	206.78
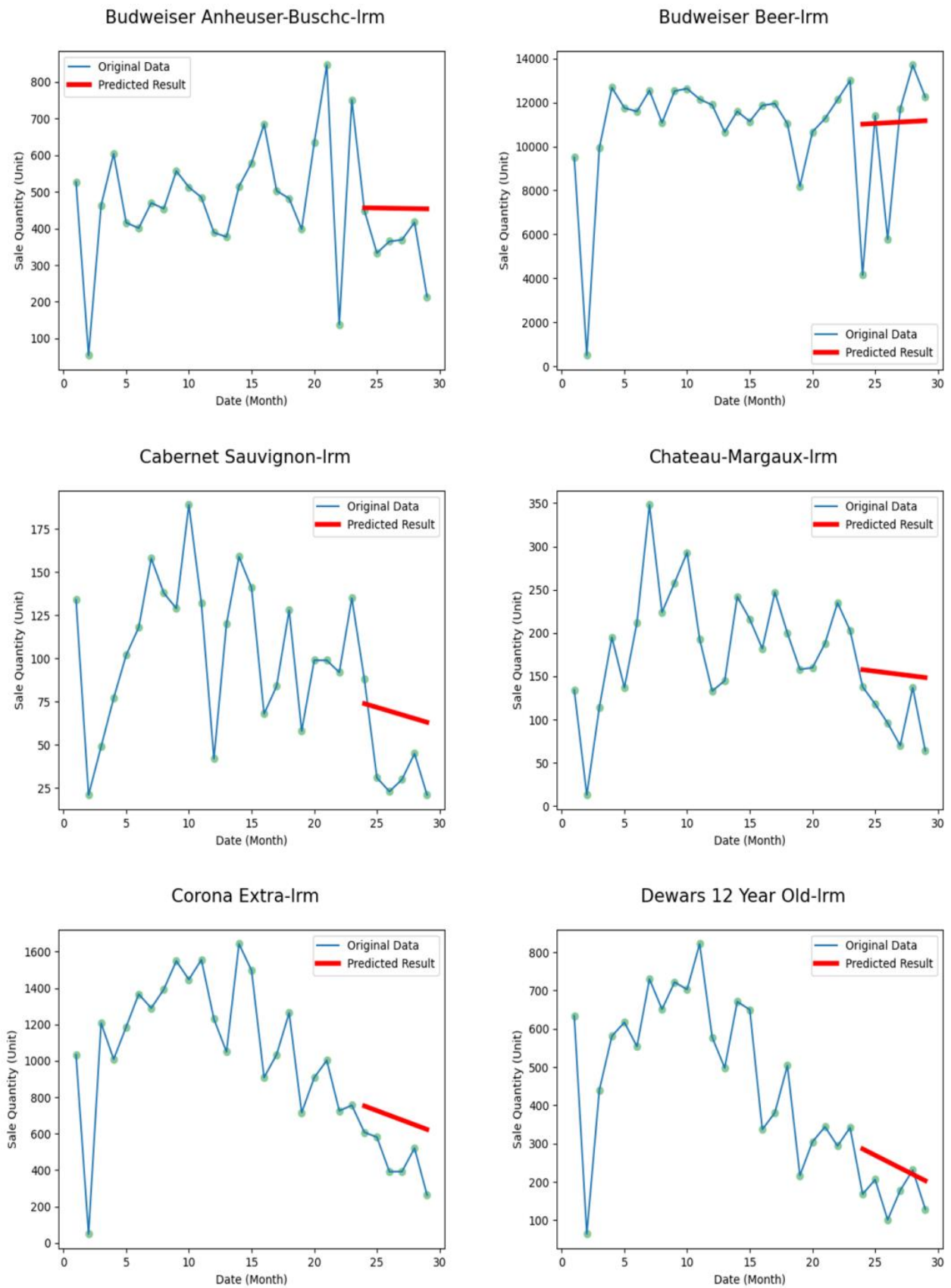
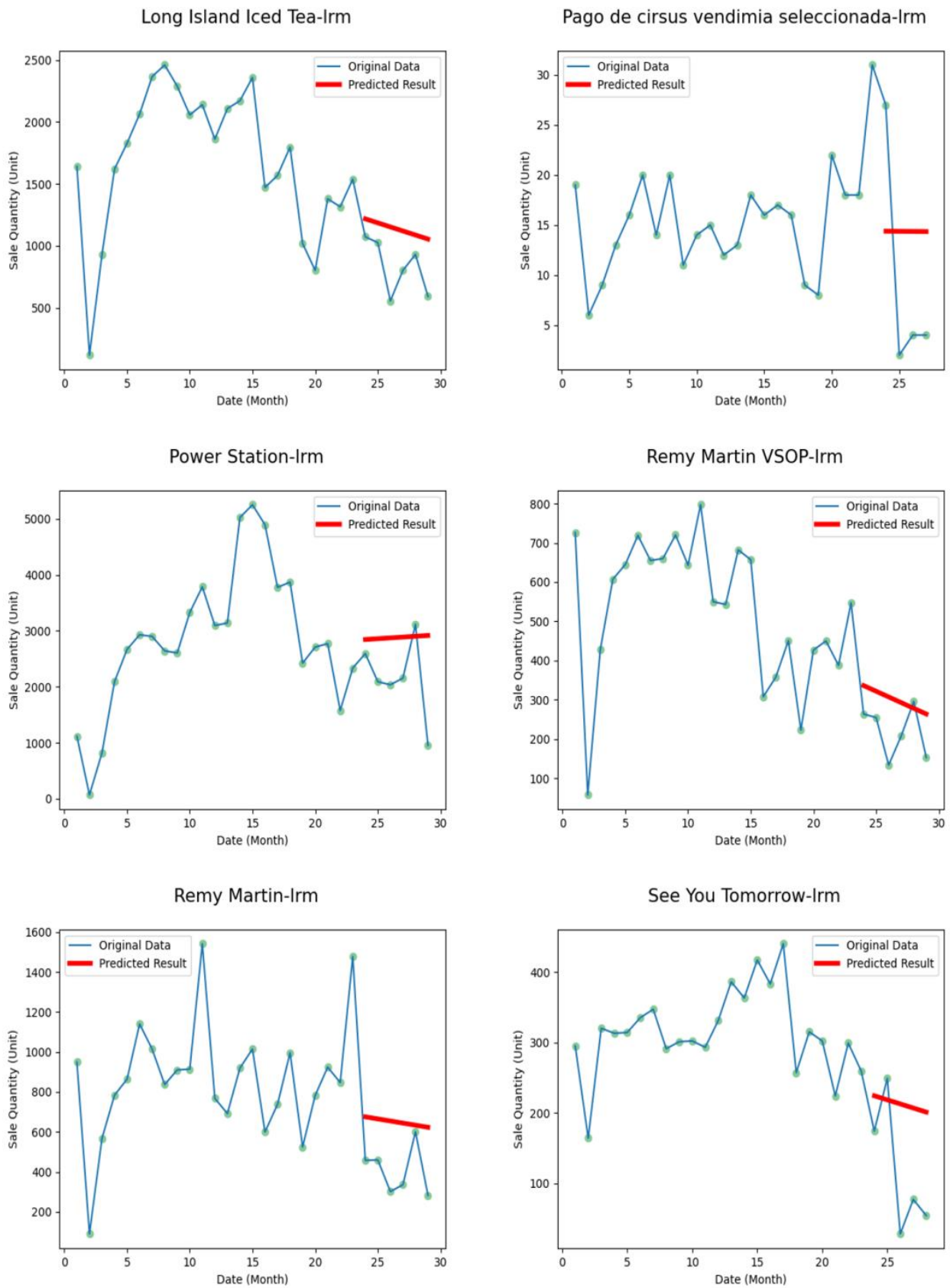Figure 3 (a): Predicted graphs for twelve products using LRM.

Figure 3 (b): Predicted graphs for twelve products using LRM.

According to the predicted results of the twelve products, the current sales trend of the Suhe Bar Chain is generally declining. Only a few products can maintain the same status or show a slight growth trend. This phenomenon is also in line with reality. Given that China's current policy in response to the pandemic is still "dynamic zero," it is common in many parts of the country to ask all profitable businesses except supermarkets and pharmacies to suspend operations. To ensure the implementation of government policies, bars, as a kind of entertainment industry, are often faced with the situation of closure, which leads to unstable turnover and a continuous downward trend.

2.5. Autoregressive Integrated Moving Average Model (ARIMA)

Autoregressive Integrated Moving Average (ARIMA) is a generalization of the simpler autoregressive moving average and adds the concept of integral. ARIMA is one of the most widely used time series prediction models. Seasonal and trend forecasts are also included in ARIMA. Considering that the tourist season and holidays of each year are relatively fixed, the effect of such cyclical rules on the turnover of bars will be similar every year. Therefore, periodic forecasting will be adopted in this study. Like in LRM, the dependent variable in the AMRIA forecast model will be time, and the independent variable will still be the sales quantity. In the AMRIA model, the p-value is the number of lag observations, the d-value is the number of times differenced, and the q-value moving average window size.

Compared with other time series models, ARIMA is often more accurate in prediction. However, this model is in a linear structure, but time series are rarely linear [8]. Because of different management methods among different stores and the unique noise of bars, the data generated might be insufficient. In particular, it is difficult to analyze the relationship between customers' economic strength and sales due to the excessive liquidity of bars. Similar to LRM, when using the ARIMA model for prediction, the actual data is first imported into the ARIMA model. ARIMA will take data from January 2020 to December 2021 as training data for the model. All actual sales data from January 2020 to May 2022 are drawn in the generated graph for comparison in the blue line. The ARIMA model was then used to derive the projected values for January to May 2022, which are plotted in the same graph in the red line. Table 2 contains the ARIMA forecast values for the twelve products from January 2022 to May 2022. Figure 4 contains the corresponding graphs, where units for Sale Quantity are all bottles, and units for Date are the number of months after January 2020.

Table 2: Predicted values for twelve products using ARMIA (unit: bottle).

| Date	Budweiser Anheuser-Buschc	Budweiser Beer	Cabernet Sauvignon	Chateau-Margaux	Corona Extra	Dewars 12 Year Old
202201	383.37	9515.89	81.29	175.16	703.42	250.20
202202	612.87	7761.26	76.84	157.94	666.96	271.30
202203	395.22	8927.56	55.28	111.31	539.45	146.87
202204	349.85	7877.34	32.34	86.88	471.29	178.73
202205	374.47	11805.45	37.80	133.28	428.67	163.57
Date	Long Island Iced Tea	Pago de cirsus vendimia seleccionada	Power Station	Remy Martin VSOP	Remy Martin	See You Tomorrow
202201	1167.59	19.19	2462.39	385.61	985.24	275.71
202202	1175.96	24.28	2037.15	402.97	983.86	238.49
202203	713.55	24.10	2137.00	212.02	395.07	219.14
202204	824.08	15.50	2167.76	207.44	362.46	180.81
202205	832.50	18.26	3105.08	213.36	433.58	63.18
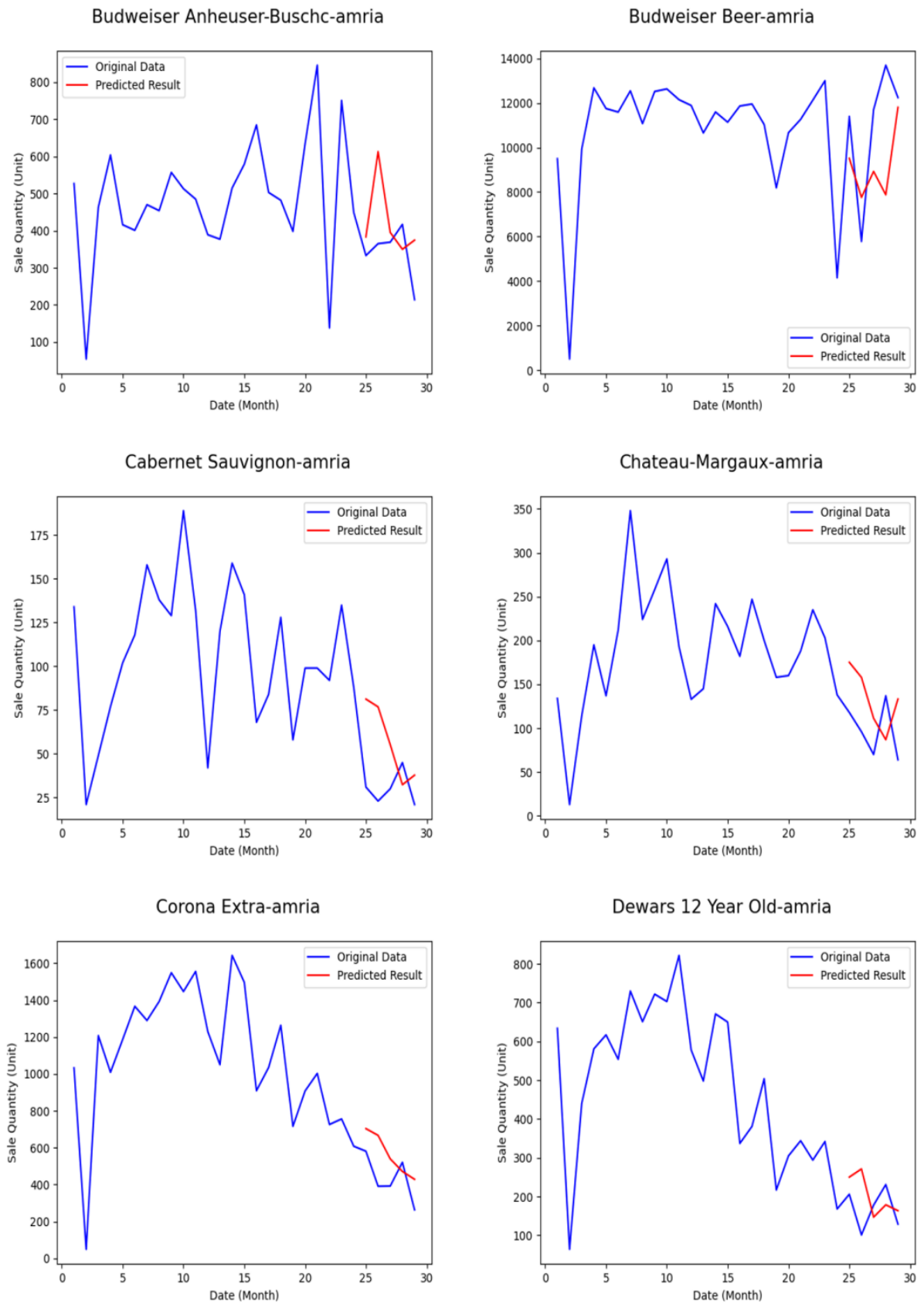
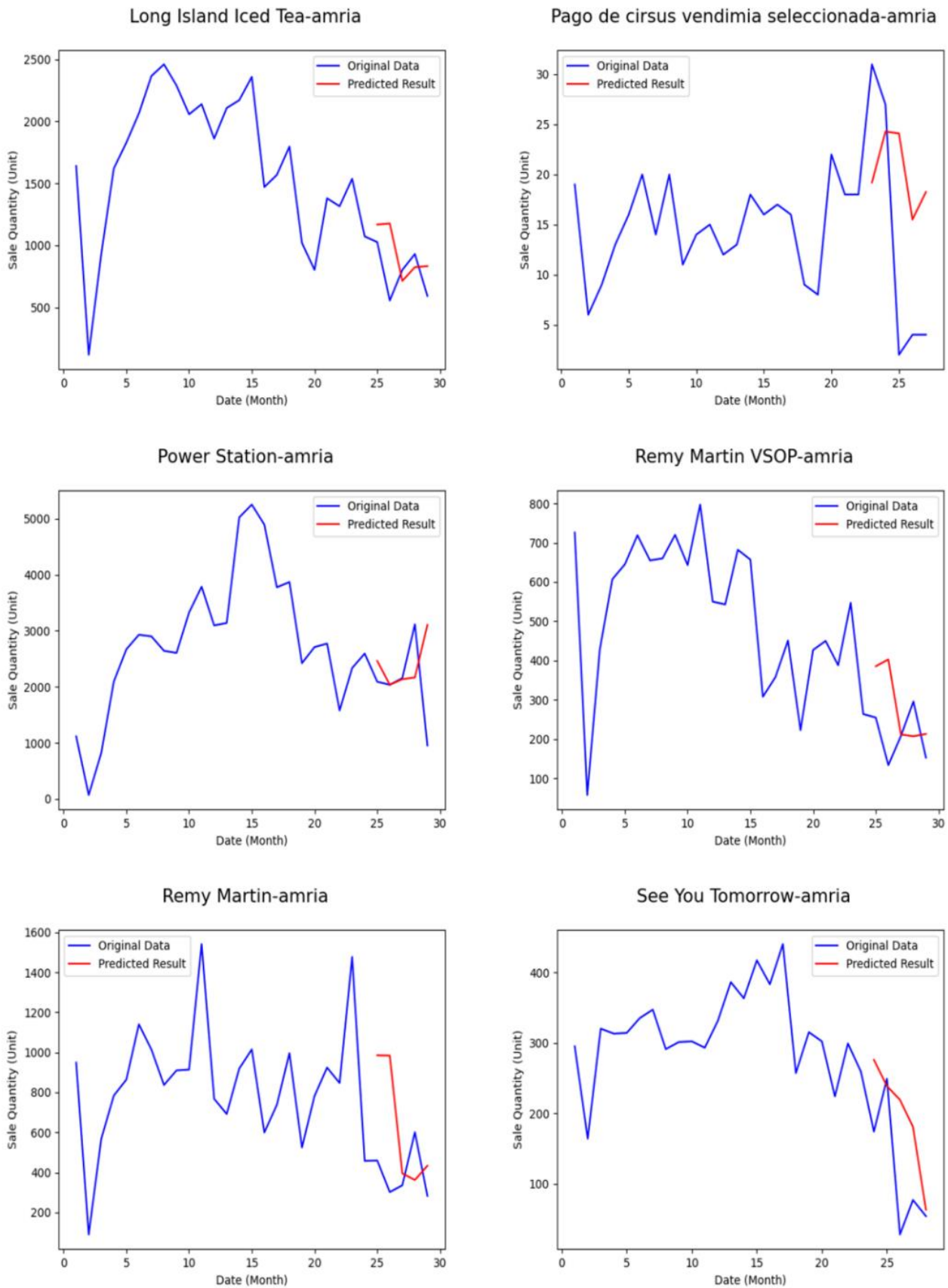Figure 4 (a): Predicted graphs for twelve products using ARMIA.

Figure 4 (b): Predicted graphs for twelve products using ARMIA.

In the figure obtained, most of the results predicted by the ARIMA model show a downward trend. But sales of several items, led by Budweiser Beer and Power Station, demonstrate the sign of rebounding. Although the forecast sales after the rebound are still difficult to keep up with the historical sales data, it still reflects the trend of gradual recovery. Along with China's current policy, the liquor retail industry is expected to start recovering from the impact of the pandemic. But given the volatility of the data, it is not certain that the industry is out of the recession yet. Although the overall situation is still on a downward trend, there is a possibility of industry expansion.

3.   Result Comparison

To compare the accuracy of the prediction generated by the two models, the Root Mean Square Error (RMSE) will be calculated for each model. RMSE is a measure of how the residuals spread out, and residuals are a measure of how far from the regression line data points are [9]. To derive the RMSE value, both the predicted values and the actual data will be considered. The RMSE numerical calculation formula is shown as Equation 2:

$$\text{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}} \tag{2}$$

As for variables in the equation, y_pred is the predicted value, which in this case is the predicted value of the sales quantity generated by the two prediction models, and y_ref is the actual data, which is the actual sales quantity corresponding to the predicted values. In terms of accuracy based on RMSE, a lower RMSE value indicates that the prediction model is more accurate [10]. Considering that the numerical calculation of RMSE requires multiple pairs of predictive values and actual values, this study will calculate RMSE values based on each product. Thus, twelve RMSE values were calculated for each model. By averaging the RMSE values, one final RMSE value can be obtained for each model. The RMSE calculation results are shown in Table 3.

Table 3: RMSE result for LRM and ARIMA.

Models	Budweiser Anheuser-Buschc	Budweiser Beer	Cabernet Sauvignon	Chateau-Margaux	Corona Extra	Dewars 12 Year Old	
LRM	80.19	4055.13	34.19	49.28	216.72	94.60	
ARIMA	137.78	3139.49	36.08	56.78	169.00	84.68	
Models	Long Island Iced Tea	Pago de cirsus vendimia seleccionada	Power Station	Remy Martin VSOP	Remy Martin	See You Tomorrow	Average
LRM	325.13	11.90	667.73	97.30	249.00	105.08	498.85
ARIMA	310.65	13.93	1064.39	142.06	405.93	107.56	472.36

The final RMSE value of the LRM is 498.85, and the final RMSE value of ARIMA is 472.36. Because the RMSE value of ARIMA is smaller, the study concludes that the prediction of ARIMA is more accurate for the sales forecast of China's liquor retail industry.

4.   Conclusion

No matter which forecasting model is adopted, the sales data for twelve products show a downward trend. It can be inferred that under the influence of the current pandemic and combined with China's policies, it is difficult for China's liquor retail industry to recover its sales status back to the pre-pandemic level in a short period. Thus, the industry is still in the "cold winter". In this case, the company can reduce the stock of goods, and maintain a healthy cash flow state. The superiority in the scale of funds could help companies quickly recover the turnover after the "cold winter". In the

ARIMA model prediction, the latter trend line is waved. This also corresponds to reality. When the policy allows business operation, the company needs to quickly grasp the opportunity, actively cooperate with the upstream industry, and do rapid product mobilization. If the company can respond to inventory more quickly, it should be able to gain an advantage in the industry. Considering the improvements for later study, first of all, LRM could take segmented data. Based on quarters, data can be turned into quarterly data of each year connected. Such segmented data could be used to predict the target year's quarterly sales data. Considering the similarity of sales patterns in the same quarter, such segmented forecasting can increase the forecast accuracy of LRM. Secondly, more models can be considered for comparison. By comparing more models, it is likely to find a more accurate model to provide more accurate suggestions for the development of the industry.

Acknowledgements

References

[1] D. Liu, W. Sun, and X. Zhang. "Is the Chinese economy well positioned to fight the COVID-19 pandemic? The Financial Cycle Perspective," Emerging Markets Finance and Trade, vol. 56, no. 10, pp. 2259–2276. 2020.

[2] J. Zhao, F. Xiong, and P. Jin. "Enhancing short-term sales prediction with microblogs: A case study of the movie Box Office," Future Internet, vol. 14, no. 5, p. 141. 2022

[3] D. Bloznelis. "Short-term Salmon Price forecasting," Journal of Forecasting, vol. 37, no. 2, pp. 151–169. 2017.

[4] J. A. Hoyle, R. Dingus, and J. H. Wilson. "An exploration of sales forecasting: Sales manager and salesperson perspectives," Journal of Marketing Analytics, vol. 8, no. 3, pp. 127–136. 2020.

[5] D. Wong. "China's city-tier classification: How does it work?," China Briefing News. [Online]. https://www.china-briefing.com/news/chinas-city-tier-classification-defined/. 19-Apr-2021

[6] T. L. Nguyen, N. T. Nguyen, and V. C. Nguyen. "Identifying factors influencing on the profitability of Tourist Enterprises: Evidence from Vietnam," Management Science Letters, pp. 1933–194. 2019

[7] Satyavishnumolakala. "Linear regression -Pros & Cons," Medium [Online]. Available: https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0. 12-Jun-2020

[8] S. Barak and S. S. Sadegh. "Forecasting energy consumption using ensemble Arima–ANFIS hybrid algorithm," International Journal of Electrical Power & Energy Systems, vol. 82, pp. 92–104. Nov. 2016

[9] "RMSE: Root mean square error," Statistics How To [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/. 31-May-2021.

[10] J. Moody, "What does RMSE really mean? - towards data science," Towards Data Science. [Online]. Available: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e. 05-Sep-2019.