SALES OF BOOKS FORECASTING

JIAHONG LIN

Contents

1. Introduction 2
2. Data Analysis 2
3. Feature Extraction 6
4. Model Train 6
5. Conclusions 8

Date: (None).
2020 *Mathematics Subject Classification.* Time series forecasting.

1. Introduction

Sales forecasting using time series is one of the most important applications in business and economics. It helps businesses make production plans, inventory management, sales and marketing strategies, etc. based on the predicted future sales. Time series forecasting can be done using various models and algorithms such as linear regression, exponential smoothing, ARIMA, neural networks, etc.

This sales forecasting task has multiple factors that need to be considered. It is necessary to take into account the impact of multiple factors on the forecasting goal.Consider how to perform data analysis to simplify the model prediction goal.

2. Data Analysis

The data covers six countries: Belgium, France, Germany, Italy, Poland, Spain, two stores: KaggleMart, KaggleRama, and four products.The training set timeline is from 2017 to 2020, and the test set timeline is 2021. Firstly, the monthly sales volume of the overall data is analyzed. This paper finds that there is a cyclical trend in the change of sales.

Figure 1. Monthly sales

Analysis of the store factors will be conducted next. By calculating the daily sales ratio of each store, it is found that KaggleMart's sales ratio remains around 75%, while KaggleRama's sales ratio is 25%. This indicates that the sales ratio of stores is almost fixed and unchanged.

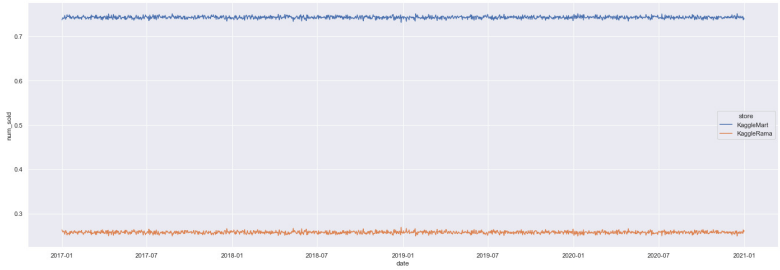

FIGURE 2. Stores ratio

For more intuitive performance, this paper takes the sales volume of KaggleMart as the benchmark, and multiplies the sales volume of KaggleRama by a weight value to remove the impact of the ratio. The observation chart shows that their sales volume change curves almost coincide, indicating that their sales volume is only related to the total sales volume and has nothing to do with the factors of the store itself.

FIGURE 3. Stores ratio trend

We analyzed the national factors in the same way and found similar phenomena. The sales volume of each country is different before 2020, and the proportion is almost fixed, while the proportion of six countries is the same in 2020. It can also exclude the influence of the country itself and predict the sales volume of each country through the total sales volume.

FIGURE 4. Countries ratio

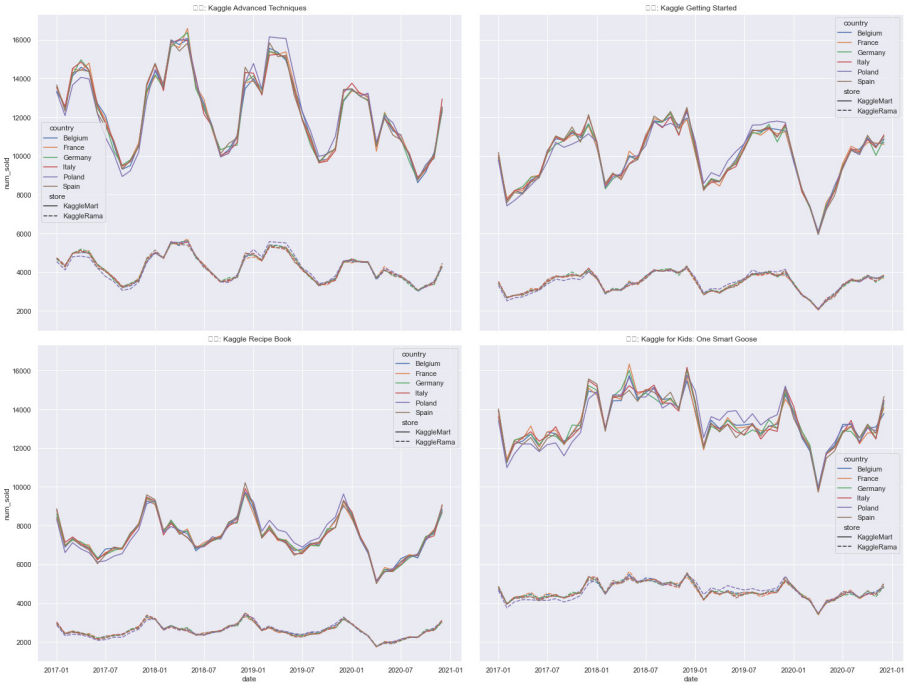

FIGURE 5. Countries ratio trend

Finally, we will analyze the store and the country together, and use the KaggleMart of Belgium as the benchmark to draw the broken line chart of the flower removal rate. It was found that the sales curve almost coincided.

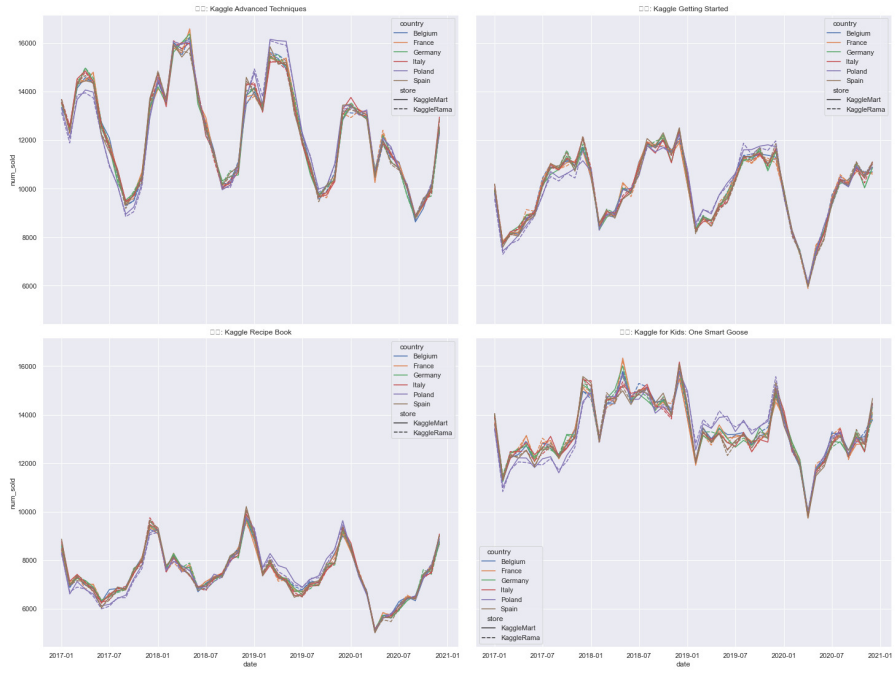

FIGURE 6. Countries and Store trend

After that, we analyzed the change of the proportion of the four commodities and found that there was a two-year cycle of change.

FIGURE 7. Product ratio trend

Therefore, we can only predict the total sales volume of each day by aggregating the time series, and then predict the sales volume of various products in stores in various countries by the total sales volume.

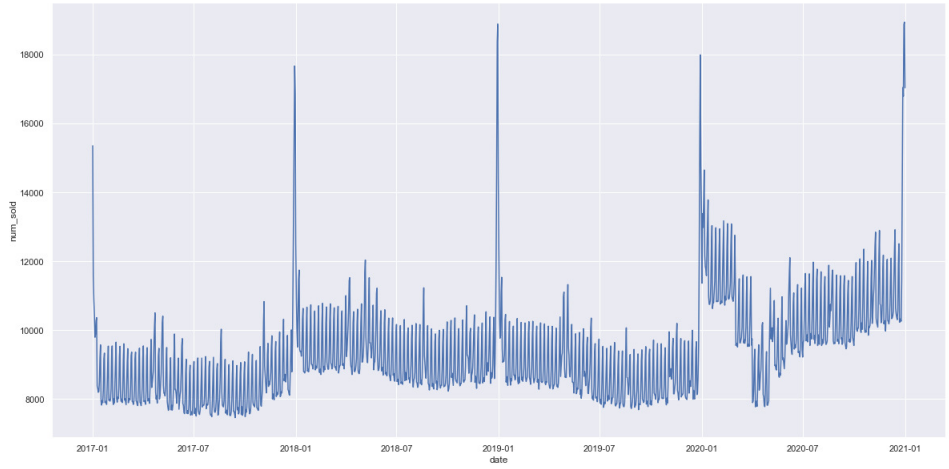

FIGURE 8. Aggregated time series

3. Feature Extraction

Through the above data analysis, we found that we can predict all results as long as we predict the total daily sales and the proportion of product sales. Therefore, we need to extract the time feature of the training set.

By analyzing the sales curve of a week, we find that the sales volume from Sunday to Wednesday is almost the same, while the sales volume from Thursday to Saturday is different every day, so we can extract the time feature of a week.

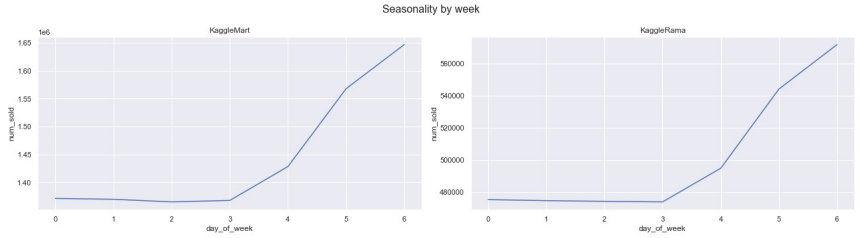

FIGURE 9. weekday feature

At the same time, in order to make Fourier transform understand the monthly data, we remove the sine value and cosine value of the month. At the same time, we also considered the feature of important festivals in the year. Finally, we get that the features of the training set include 23 features.

4. Model Train

We use Ridge regression from the "linear˙model" module to correlate the relationship betweensales and time features , and predict on the test set.Train the model with time features as X and sales as y.The sales of the test set is predicted according to the time features of the test set.Finally, we obtained the total sales prediction curve.

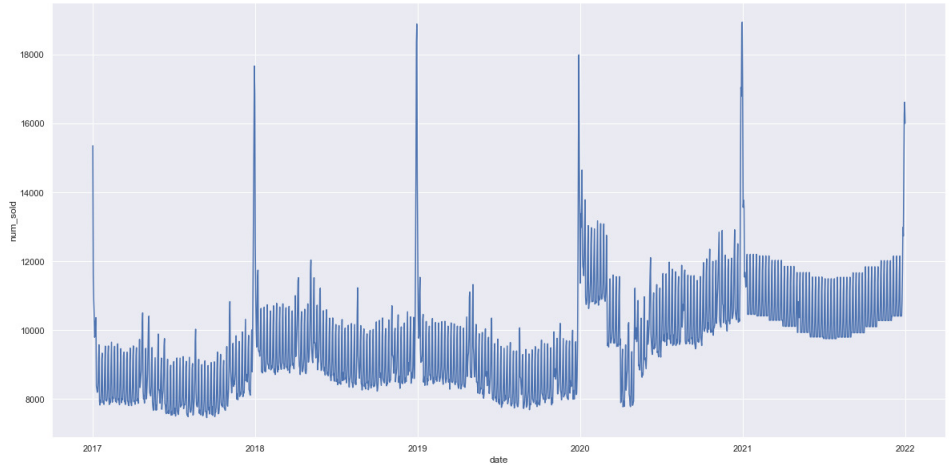

FIGURE 10. total sales forecasting

We found that the proportion of sales for a product has a cyclical variation with a period of two years. Therefore, we assign the daily proportion of sales for each product in 2019 to 2021.

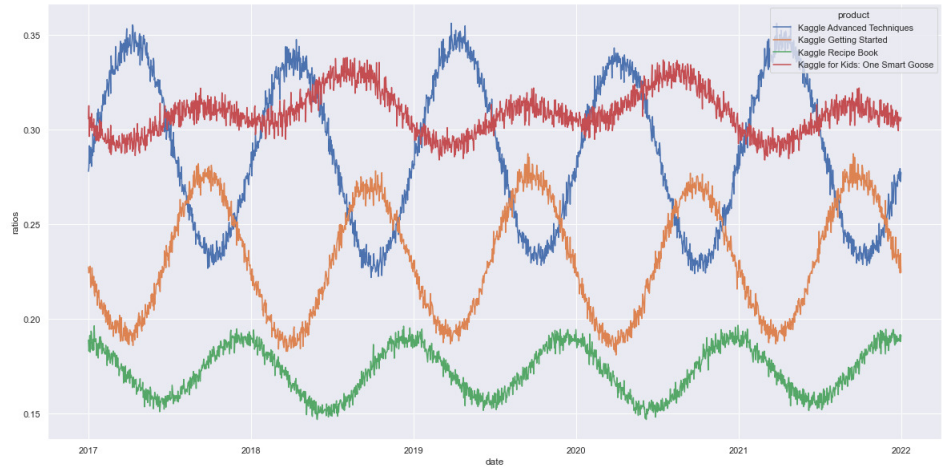

FIGURE 11. Product Ratio Forecast

We assume that the proportion of sales in countries in 2021 is the same as in 2020. In 2020, the proportion of sales in countries accounts for 1/6.The proportion of different stores remains fixed, KaggleMart accounts for 75%, and KaggleRama accounts for 25%.

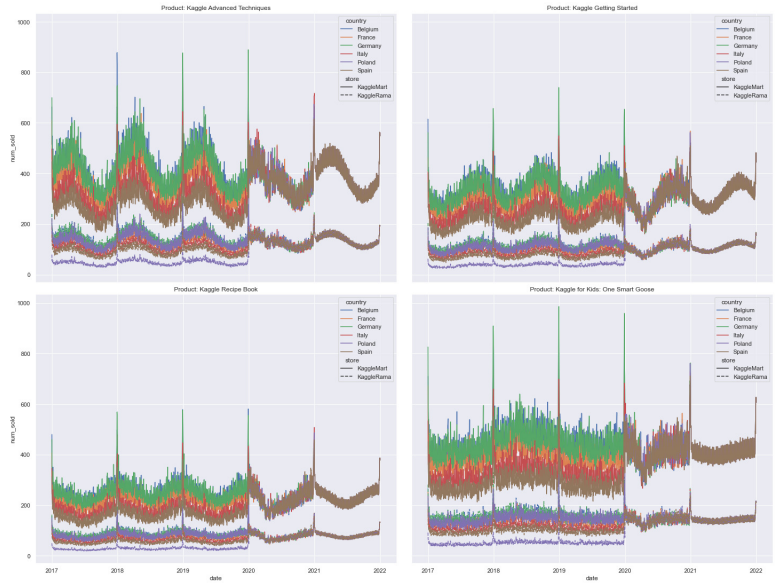

FIGURE 12. Final Forecasting

5. CONCLUSIONS

This is a time series forecasting problem that includes complex elements.Simplifying the effects of complex factors through analyzing patterns discovered through single factor analysis.Use linear regression method to predict the relationship between sales volume and time characteristics. The quick brown fox jumps over the lazy dog. Jackdaws love my big Sphinx of Quartz. Pack my box with five dozen liquor jugs. The five boxing wizards jump quickly. Sympathizing would fix Quaker objectives.

(A. 1) School of Economics and Management,, Nanjing University of Science and Technology, China