Sales Prediction Optimization via Gradient Boosting Decision Tree

Yutian Shi*

Sun Yat-sun University Guangzhou, China

* Corresponding Author Email: shiyt25@mail2.sysu.edu.cn

Abstract. Sales prediction is always a top priority for brands, effectively reducing unnecessary waste under the essential circumstance that all customers' needs have been met. With an efficient prediction model, brands can make accurate inventory decisions and get a significant increase in profit. This paper provides an efficient sales prediction model for cross-border e-commerce companies with Gradient Boosting Decision Tree. We use the sales volume in the past months, their statistical characteristics, difference characteristics and month on month characteristics as the characteristics we choose. We use the Mean Absolute Percentage Error perimeter to compare the accuracy of our method with the Moving Average Method to assess the performance of our model. The results show that our iterative model has higher accuracy compared to the Moving Average Method. It also has a significant edge in both automation and analyzing big-scale data sets, which means that it solves the low feasibility of existing methods. On this basis, we finally conclude that our machine learning model can make great contributions to sales prediction.

Keywords: Moving Average Method, Gradient Boosting Decision Tree, Machine Learning, Sales prediction, Mean Absolute Percentage Error.

1. Introduction

Sales prediction plays an outstanding role in the sales business. At its best, current stock should meet sales demand to the maximum, and only minimum stock will waste. Thus if the model for sales prediction is accurate enough, factories can get guidance in producing an appropriate amount of products, which effectively reduces unnecessary waste under the essential circumstance that all customers' needs have been met.

In the past two years, cross-border e-commerce companies have flourished in China, especially in Pearl River Delta. This kind of cross-border e-commerce company generally make its brand, focuses on a specific category, has its own production factories and establishes its own stores on E-commerce platforms such as Amazon, Ali-Express and Wal-Mart. Their customers are mainly in the US and Canada markets. Such products need to be transported to the local warehouse in the United States by sea, thus meeting the timely purchase needs of users. [2] Therefore, the challenge faced by such companies is that the cycle of the supply chain is incredibly long, the reaction to the market is relatively lagging, and the production is easy to be surplus or insufficient.

Especially since 2020 when COVID-19 breaks out, many products of such cross-border e-commerce brands have become more and more popular, and their saleshave gone up. Under this circumstance, production capacity has become a bottleneck. [1] However, most companies haven't paid enough attention to this, which eventually led to unnecessary waste and losses of orders. Therefore, the problem was born---- brands need a more scientific sales forecasting model to obtain a more accurate sales forecasting model, which can be used to guide the production capacity ranking of the factory, to meet the market demand to the greatest extent, thus maximizing profits, reducing inventory and improving turnover [3].

In this paper, we put forward a sales prediction optimization based on the Gradient Boosting Decision Tree. For some special goods, such as those with seasonality, strong weather correlation, or other particular timeliness considerations, we don't consider them. On the base of the selected data set, we use the moving average method (MAM) and gradient boosting decision tree(GBDT) to obtain the smoothing coefficient and corresponding weight and get the sales prediction results of the two

ways. Firstly, we calculate the most suitable parameters with the principle of MAM and the actual data. Then we use this parameter to build the MAM prediction model and make predictions. Then we utilize the lightGBM framework to realize the algorithm of GBDT, thus selecting the feature model and predicting the next sales. The features we selected include statistical features (sum, average, maximum and minimum, sum and average of the first quarter in the past six months, sum and average of the second quarter in the past six months), neighbourhood differences for months and quarters, month on month characteristics and so on. We used the feature importance in lightGBM to help us determine the final selected features and complete the GBDT sales prediction model.

To verify the effectiveness of the methods, we use the Mean Absolute Percentage Error (MAPE) perimeter to compare the results of GBDT and MAM, against actual sales. The experimental results show that both GBDT and MAM work well in predicting, but GBDT has more minor relative error and stronger operability. GBDT is also easy to iterate, which dramatically enhances its competitiveness.

2. Method

This section describes the process of establishing a predicting model using the moving average method and gradient lifting decision tree. We first select the more standard product data as our data set and simply process it (Sec.A). Next, we introduce the index MAPE, which is frequently used in the research process (Sec.B). Then we use MAPE and other parameters to get the critical parameters needed for the prediction of the moving average method(Sec.C). Finally, we complete the feature selection of the GBDT algorithm and establish the optimized prediction model(Sec.D).[4]

2.1.Data sets and sources

2.1.1. Preliminary selection

We assume brand A as a standard brand, which means that a few of its products will bring most of the profits, and this part of the products is needed by most families throughout the year. It represents that there is a specific need for rigidity and periodic replacement. In other words, this kind of standard products will have an inevitable demand cycle, and 20% of the products meet 80% of the users' needs. We select the commodity Stock Keeping Unit (SKU) that plays a significant contribution to the sales of brand A as our data sets. The batch of SKUs contributed more than 8 layers of total sales. After preliminary screening, there are about 900 SKUs.

2.1.2. Further selection

Considering the influence of seasonality, timeliness and other factors on the continuity of standard products, we eliminate more than 200 SKUs. In addition, since we predict the monthly sales volume of three months in the first quarter of 2020 based on the data for the whole 2019, if some SKUs change significantly in the next month due to special reasons such as production, supply or market, it is inappropriate and unnecessary to predict such SKUs, so they are also eliminated.

In the final analysis, only less than 500 SKUs remain. In fact, 75% of the sales performance of brand A in the whole year of 19 years came from these less than 500 SKUs. Therefore, even if half of the SKUs are excluded, the analysis based on the remaining less than 500 SKUs as the data set plays a certain role in the sales forecast of brand A.

2.2. Parameters

Mean absolute percentage error(MAPE) is the evaluation index we use to determine some parameters and evaluate prediction models:

$$\text{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \quad (1)$$

It ranges from 0 to unity. A better model should have a lower MAPE error. When MAPE error comes to 0%, the model can be esteemed as a perfect one.

2.3. Moving average method

Typically, brand companies use the moving average method to forecast need and sales. It predicts the demand and capacity of the companies' products in one or more future periods with a set of recent actual data values. The moving average method is suitable for spot forecasting. In the paper, we use the n-Period moving average [5].

2.3.1. Principle

The N-Period moving average is a static sales prediction method. Assume that the observed value of the last period is:

$$\tilde{x}_t=\frac{1}{n}(x_t+x_{t-1}+\ldots+x_{t-n+1}) \tag{2}$$

Then, we use the N-Period moving average method to predict the periods forward, and the predicted value of each period should be:

$$\hat{x}_t(1) = \frac{x_t+x_{t-1}+\cdots+x_{t-n+1}}{n}$$
$$\hat{x}_t(2) = \frac{\hat{x}_t(1)+x_t+\cdots+x_{t-n+2}}{n}$$
$$\vdots$$
$$\hat{x}_t(\mathrm{l}) = \frac{\hat{x}_t(l-1)+\cdots+\hat{x}_t(1)+x_t+\cdots+x_{t-n+l}}{n}$$

2.3.2. Selecting parameter

For each SKU, 24 months of sales data in 19 and 20 years are used. We use the exhaustive method to list some values of n, including 3, 4, 5 and 6, which represent the sales data of these periods, as our parameter n of the Moving Average Method. When training parameters, data in 2019 is used to predict the sales of each month of 2020. As mentioned earlier, from factory production to overseas sales, the supply chain has many links and takes a long time, so it needs to be predicted several months in advance. Due to this, the forecast values are used to predict the next month.

For example, when n =3, only the actual sales of 2019 October, November and December are used to predict 2020 January. When it comes to predicting 2020 Febuarary, we use the actual sale in 2019November and 2019December years and the predicted sale in 2020January. In this way, we can get the predicted sales for each month in 2020. When the value of n changes, the number of previous months which used to predict also changes. Table I shows our prediction results for different parameters n for each month. The error here uses the MAPE index mentioned above.

Table 1. Prediction error of different n

MAPE	MONTH					
	Jan.	Feb.	Mar.	Apr.	May.	Jun.
n=1	0.76	0.79	1.90	2.43	1.80	1.76
n=2	0.80	0.82	1.87	2.46	1.76	1.67
n=3	0.78	0.81	1.87	2.54	1.97	1.90
n=4	0.76	0.80	1.75	2.12	1.67	1.50
n=5	0.81	0.84	1.63	2.08	1.87	1.45
n=6	0.79	0.83	1.56	1.98	1.76	1.34
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
n=1	2.12	1.87	1.35	1.19	0.82	0.78
n=2	2.09	1.76	1.32	1.07	0.79	0.75
n=3	1.99	1.45	1.26	0.94	0.77	0.69
n=4	1.67	1.65	1.11	0.89	0.81	0.77
n=5	1.55	1.34	0.95	1.01	0.67	0.79
n=6	1.44	1.31	0.91	0.83	0.65	0.71

From Table I, we can find that the error from January to April showed an upward trend. It gradually decreased for two months, then increased, and then gradually decreased from August to December.

The reason why there is such a significant error in the data for March, April and May in 20 years is that the global epidemic broke out, and the US and Canadian markets were affected by the epidemic. The local people were forced to stop going out and the offline market sales were poor. This phenomenon leads to more time spent at home, which makes online sales increase rapidly. Therefore, using the moving average method to predict the months affected by the epidemic in 2020 based on the data of 2019 is not a good choice. Most of the MAPE error values are more significant than 1, and some exceed 2, indicating that the model is not good.

Since the outbreak of the epidemic directly led to the doubling of sales over the same period in previous years, the errors in prediction results are larger. But specific rules can still be seen from the curve of MAPE.

The closer the month is to the actual month, the smaller the MAPE is. The inaccuracy of MAPE will be superimposed with the previous prediction result data as the primary prediction data, resulting in the lower and lower accuracy of prediction.

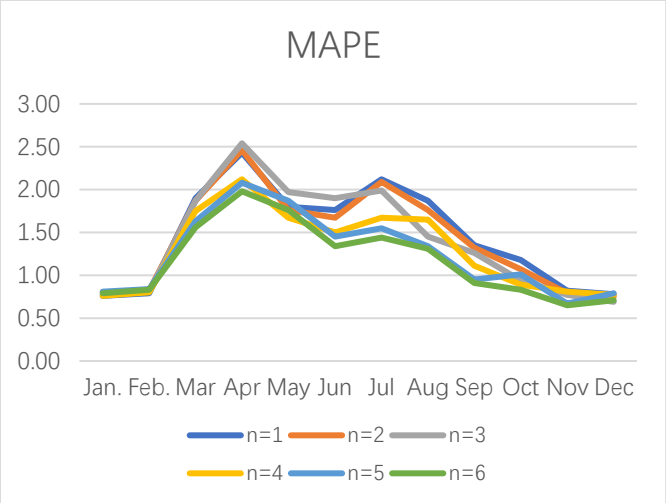


Fig.1 Prediction error of n for each month under different values

Fig.1 shows the line chart of the prediction error of n for each month under different values. From Fig.1, it can be seen that trends of different parameters are generally similar. There is no significantly superior parameter. When parameter $n = 6$, the accuracy is the highest.

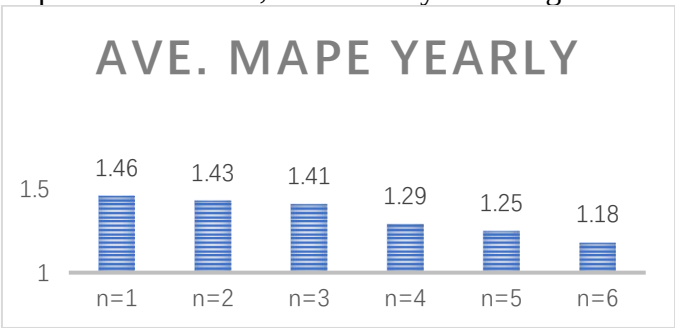


Fig.2 Annual average MAPE error of parameter n under different values

Fig.2 shows the annual average MAPE error of parameter n under different values. From the perspective of annual error, the accuracy of the prediction is still the best when $n = 6$. However, at the practical level, there are other influencing factors, such as the frequency of new product launches and the cost of measurement. Therefore, the brands can have their own choice when choosing the value of parameter n to guide the production trend.

We set $n = 6$ since we only need to predict the sales of SKU in this paper. So far, the moving average model has been established.

2.4. Gradient boosting decision tree

The Gradient Boosting Decision Tree is an integration algorithm based on Decision Tree. It gets new training models through a gradient descent algorithm, which makes the model more exact. In this paper, we utilize the light GBM framework to select a feature model and predict the following sales.[6]

2.4.1. Preliminary feature selection

The core of the decision tree algorithm is to classify the data set according to its characteristics. In our initial feature selection, the following types of data are considered: we consider raw data as sales in the past six months, statistical data as sum, average, maximum and minimum, sum and average of the first quarter in the past 6 months, sum and average of the second quarter in the past 6 months. Gap value data includes both differences between adjacent months and difference between season1 and season2. Month on month data is about the quotient of sales volume in adjacent months and their differences. And we choose some particular month characteristics for they influence the sales a lot, such as whether there is a 'prime day' in this month, whether there is a 'Black Friday' in this month.[7]

2.4.2. Further feature selection

Light GBM is a gradient framework which can learn in parallel and deal with a large-scale data. It enables GBDT to process more data at a higher speed [8].

The 'feature_importance' method in light GBM can show the relativity between characteristics and results suitable for filtering features. We use 'feature_importance' to visualize the correlation between the aforementioned characteristics and the final sales [9]. Fig.3 shows how many times those features are selected when making predictions. The more times the model is selected, the more influential the feature is valued by the model.

From Fig.3, we find that SKU (the number of SKU), last_1_sale (the sale in the last month), 2_4_diff(the difference between the sale in the month before last month and the sale in the fourth month before), 1_3_diff(the difference between the sale in the last month and the sale in the third month) show a higher relevance compared to other features. Thus they are chosen as the final features. Besides, some of the month and month on data and month characteristics are also of some import, so they are selected too.

At this point, we have completed the most critical step of building the GBDT prediction model - feature selection. Then we use these features to write code and do some machine learning to get the final predictions.

3. Results and Conclusion

In this section, we use the sales data of the brand in 2020 to predict the monthly sales volume in the first quarter of 2021. The prediction using MAM is analyzed in Sec.A, and the GBDT model is in Sec.B. We compare the average MAPE error of MAM and GBDT models (Sec.C).

3.1.Moving Average Method

We use the sales data of the brand in 2020 to predict the monthly sales volume in the first quarter of 2021 in the actual verification. We use MAPE to compare the predicted value with the true sales value, in which the total error is expressed as the error of each month, and then calculate the average value again. The final results are shown in Table II.

Table 2. Error from Jan to Mar 2021(MAM)

Jan. MAPE	Feb. MAPE	Mar. MAPE	Apr. MAPE
0.28	0.41	0.39	0.36

From Table II, we can see that adopting the historical data of the same period combined with the moving average method for sales forecast has guiding solid significance for the sales forecast in the

next month, with an accuracy rate of 97%, MAPE accuracy reaches 0.28. However, for months after the next month, MAM doesn't work well, with an average MAPE accuracy of 0.386. These months use the previously predicted data for secondary prediction. Since the Moving Average Method model has only one parameter n, it will have significant problems in error amplification when predicting sales in further future.

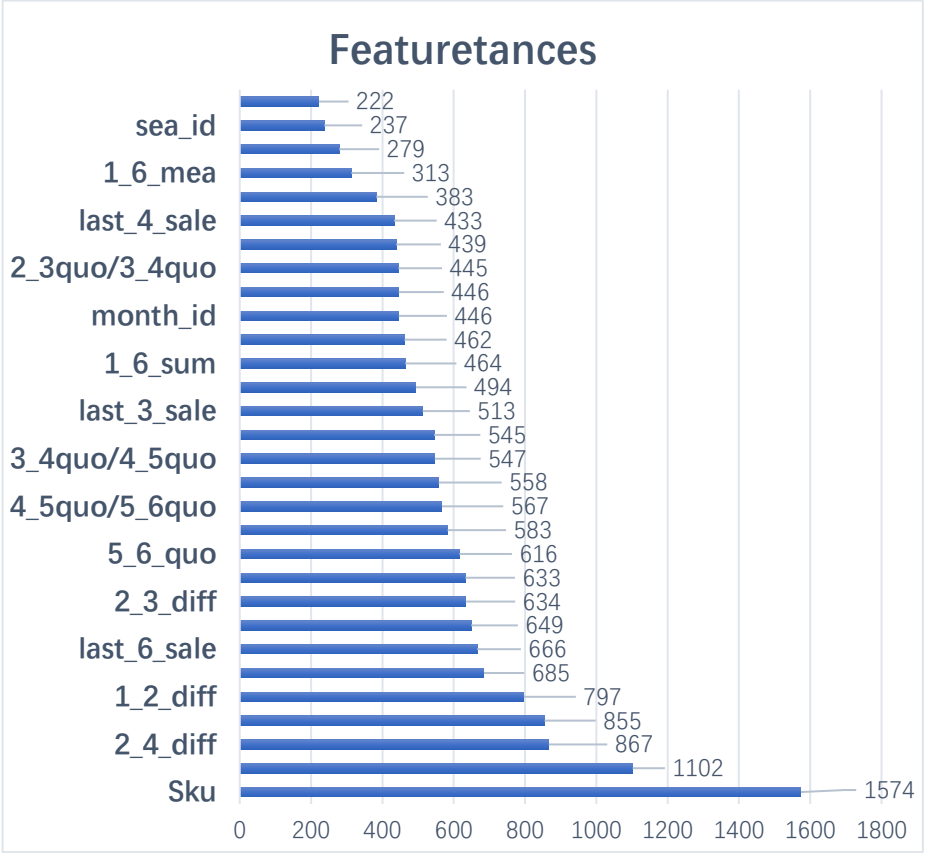


Fig.3 The number of times the feature is selected in the model

Besides, the iterative ability of the MAM model is poor. The parameter used is obtained from part of the data. After a specific time, for example, 6 months later, the old model will be abandoned and a new model needs to be rebuilt from scratch, according to sales in these new 6 months.

3.2.Gradient Boosting Decision Tree

As how we deal with the result for MAM, we calculate the average MAPE value for each month. The errors of each month predicted by the GBDT model are shown in Table III.

Table 3. Error from Jan to Mar 2021(GBDT)

Jan. MAPE	Feb. MAPE	Mar. MAPE	Apr. MAPE
0.32	0.35	0.39	0.27

From Table III, we know that the GBDT model doesn't show a significant differences when predicting the nearest month, compared with further months. Since the result of the GBDT model is related to more than 10 different features, the predicted data doesn't play a highly weighted role. Therefore, the GBDT model has an advantage in predicting sales for further months. Table III even shows a more minor MAPE error in April than in January.

Moreover, the GBDT model has a good ability for iteration. After obtaining the actual monthly sales volume, the new data can be added to the training model. With the passage of time and the expansion of the data set, the precision of the GBDT model will rise. This means that this model can be used and optimized for a long time, which is very important in practical applications.

3.3.Comparison

Table 4. Average MAPE of MAM and GBDT

	Jan. MAPE	Feb. MAPE	Mar. MAPE	Ave.	Result
MAM	0.28	0.41	0.39	0.36	
GBDT	0.32	0.35	0.39	0.27	√
Delta(%)	-14	15	0	25	

From the table, we can see that besides the MAPE error in January, the GBDT method works better in each of the rest months. With an average MAPE accuracy of 0.27, compared to 0.36 by the MAM model, we can say that the GBDT model shows a better overall accuracy. Though the delta is not significant in this study, it will be more evident after the accumulation of data. In other words, in the actual operation of a large amount of data, the GBDT may show a more tremendous advantage in accuracy compared to the result in Table IV.

In the case of actual long-term use, the GBDT model should still perform better. First, it is automatic in precise prediction. Next, it iterates and optimizes itself based on the more extensive dataset. With more training data, the precision of the model can be higher. And GBDT has more features than the only one parameter for MAM, giving better performance in avoiding error amplification.

4. Conclusion

First and foremost, the GBDT model has certain advantages in accuracy. And because of its own iteration and simplicity, the GBDT model enjoys a significant edge in automation and in the following optimization. However, machine algorithms have high requirements for Feature Engineering and data processing, so they have specific conditions for the magnitude of data. According to the ability of enterprise IT facilities construction, the moving average method can also be used as a reference for brands in the case of insufficient computing power.

In the future, some hybrid models should be proposed. Further prediction and optimization attempts based on the GBDT model should aim at the actual complex rules and SKU characteristics, such as the propaganda slogan of ‘out of stock’ [10].

References

[1] Yuqin, Gong, “Analysis on Financial Performance of LightInTheBoX Cross-border E-commerce Platform in the Epidemic Background [C]”, 3rd Guangzhou International Forum on Finance (GZIFF 2020), November 2020, pp.65-70.

[2] Hou, F.. "The Chance of China's Cross Border E-commerce in Coronavirus Pneumonia Epidemic." Region - Educational Research and Reviews 2.2, 2020, pp.26.

[3] Liu X, “Influence and Response of China's Cross-Border E-Commerce Exports under PHEIC -Take COVID-19 as an Example[C]”, 2020 International Conference on E-Commerce and Internet Technology (ECIT), April 2020, pp.58-61.

[4] Liu, E., et al. "A hybrid model for predicting spatial distribution of soil organic matter in a bamboo forest based on general regression neural network and interative algorithm.", National Key Research and Development Projects of P.R.China, 2020, pp.1673-1680.

[5] Wei L Y, Cheng C H, Wu H H. “A hybrid ANFIS based on n-period moving average model to forecast TAIEX stock[J]”. Applied Soft Computing Journal, 2014, pp.86-92.

[6] Kim, J. Y., H. S. Lee, and J. S. Oh. "Study on prediction of ship's power using light GBM and XGBoost." 한국마린엔지니어링학회지 44, April 2020.

[7] Menghao, et al. "Study on the Driving Factors of Cross-Border Online Consumer Purchase Behaviour.", International Conference on Innovative Development of E-commerce and Logistics, 2018, pp.477-488.

[8] H Wang, Yang C, H Tang. “Research on the Short-Term Load Forecasting Using Improved GBDT Based on Light GBM [J]”. Process Automation Instrumentation, 2018, pp.76-82.

[9] Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery [J]. Bioinformatics, July 2012, pp.1766-1774.

[10] Boppana, P. . "Method and system for modeling sales processes." US, US20040102995 A1. 2004.