Sales Forecast of Retail Commodity on the Basis of LightGBM and Xgboost

Guanru Wang*
*41914062@smail.swufe.edu.cn

Department of Accounting Southwestern University of Finance and Economics Chengdu, China

Abstract—With the development of market economy and economic globalization, enterprises encounter more and more market competition. On this basis, the product selection is crucial for enterprises, which strongly rely on sales prediction. In this paper, the LightGBM and Xgboost are adopted to predict monthly sales of retail products sold in a great number of shops after reasonably handling the abnormal missing data and setting the pre-processing of multiple data sets. By comparing and analyzing the prediction results, LightGBM model performs better than Xgboost model. At the same time, the key factors affecting the monthly sales are found including the number of shops, the types and quantity of items, the items reservation, the price of items, and the past sales. The prediction model and analysis results have a certain reference value for the actual sales prediction in companies' operations. These results shed light on guiding further exploration of sales strategic arrangement and corporate strategy.

Keywords-Sales prediction; LightGBM; Xgboost.

1 INTRODUCTION

Machine learning focuses on the path that machines simulate or realize human learning behaviors to acquire new knowledge or skills and reorganize the existing knowledge structure to continuously improve its own performance. It is the core of artificial intelligence and the fundamental way to make machines acquire intelligence. Specifically, machine learning is learning from data to complete a specific task through various algorithms, a process called training, and then it uses the learned knowledge to make decisions or predictions about events in the real world [1].

Throughout the history of machine learning, the earliest machine learning method can be dated back to the early 20th century and its modern sense is usually related to Frank Rosenblatt, the psychologist from Cornell University [2]. It has been nearly 40 years since machine learning was conducted an independent direction in 1980. In the past 100 years, the efforts of generations contribute to a large number of classic methods, which promote the rapid development and popularization of machine learning.

Companies as well as universities have invested significant resources to develop their knowledge about Machine Learning. Recent development in the field provides reliable results for a wide variety of tasks, which can be comparable to human performance with a high accuracy [3].

In the operation of the company, machine learning methods are also used in many aspects. Among them, sales prediction is one of the practical applications, which refers to the estimation

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328402

of the sales quantity and sales amount of all or some products in a specific time in the future. It is a case that puts forward feasible sales objectives through certain analysis methods based on full consideration of various influencing factors and combination with the actual sales performance of the enterprise. For a company, the quality of sales prediction has essential influences on the length of the company's sustainable development. Sales prediction affects basically all aspects of sales management, e.g., planning, budget and sales determination.

Sales prediction must pay attention to the sales pattern of the target subject, its sales cannot exceed the market potential, which can be achieved by judgment method or advanced measurement method [4]. Generally, its impact on enterprise sales can be analyzed based on PEST analysis model, i.e., from the political, economic, social, scientific and technological aspects, considering the industry environment, business strategy that competitors adopt, internal conditions of the enterprise to carry out qualitative or quantitative forecast of sales [5].

Qualitative analysis method is also called non-quantitative analysis method. It is a kind of forecasting method in which professionals make comprehensive analysis of the characteristics of the predicted objects according to their personal experience and knowledge, and make research and judgment on the future economic status and development trend of enterprises. It refers to the methodology system of using modern mathematical methods to process data on the basis of a complete grasp of various elements related to the predicted object and establish the regular relationship between the variables of the prediction model.

Regression analysis method is a kind of quantitative analysis, which based on the analysis of forecasting objects independent variables (e.g., time) and dependent variable (e.g., sales), to establish a regression model between variables, and is used as an analytical method to predict the future market conditions. The analysis method has been widely utilized in the sales forecast, and it can be divided into time series analysis and causal prediction analysis.

In 1996, Winklhofer et al, have pointed out that enterprise prediction commonly considers three factors: variables of organization and environment that have impact on the prediction, known variables that are specific to the company or environment and connections between different aspects that are easily overlooked. [6]. Reilly, in 2012, has pointed out that in addition to known variables affecting dependent variables, there are also some divorced values to be considered. Specifically, the influence of historical memory values, occasional sales pulses or seasonal sales pulses that are repeated periodically based on certain rules (e. g., by day, week, month, etc.), as well as time trends of variable change levels and slope changes must be taken into account [7]. Therefore, a prediction model generally uses three independent variables: causal variables, memory variables, and dummy variables.

With the advent of the era of big data, big data has gradually become a hot topic in academia and industry, and has been widely used in many technologies and industries, from large-scale database to business intelligence and data mining applications.

The design, analysis and engineering of big data algorithms involve many aspects, including massively parallel computing, streaming algorithms, cloud technology etc. Since big data is complex, high-dimensional and changeable, the efficient way to mine the knowledge that people need from the real, messy, and complex database is in urgent demand of more profound machine learning theory to guide. Companies that advance on big data distinguish from conventional analysis in 3 key ways [8]:

- They take note of data flows.
- They no longer rely on traditional data analysts, instead, they turn to cultivate leading data scientists and product and process developers.
- They are shifting analysis from IT functions to core business.

Machine learning enables companies to grasp important data information, which can not only save the cost of manual analysis, but can also provide great business analysis value. Pattern classification and regression is one of the most basic tasks often encountered in the research fields of machine learning and pattern recognition. The main goal is to use the data to build a prediction model with strong generalization ability, so that one can give the most accurate estimation of new unknown objects. Therefore, to improve the generalization ability of learning system is a fundamental problem. Due to the significant advantages of integrated learning in this field and the continuous growth of integrated learning research team, integrated learning technology develop rapidly. Besides, various integrated learning methods are also widely used in various research fields [9]. In this paper, lightGBM and Xgboost will be adopted to predict the future sales.

This paper is basically divided into four parts, the first part is about introduction, which mainly talks about background information of this research; the second part is about methodology, which mainly focuses on the methods that are used to predict sales in this research; the third part is about results, which describes the outcome and gives some explanations and limitations; the final part is about conclusion.

2 METHODOLOGY

2.1 Data

The data comes from the daily historical data of different stores from January 2013 to October 2015. In the actual prediction process, the research frequency is month, that is, after processing the data, the monthly sales of products in each store in November 2015 is predicted by using a total of 33 months from January 2013 to October 2015. The original data set contains 22,170 items, 84 commodity categories and 60 stores. The test set contains 45 stores, which are the samples to be predicted in this paper. A total of 214,199 samples in the test set need to be predicted.

2.2 Model

The models used in this research are lightGBM and Xgboost, the main ideas of which both come from GBDT, and one of the most important basic theoretical frameworks of GBDT is gradient boosting.

2.2.1 Gradient boosting

Gradient boosting subordinates to boosting, and is one of the largest classes of boosting. The core idea of gradient boosting is constructed and developed based on gradient descent method. The basic principle of it is to train the newly added weak classifier according to the negative

gradient information of the current model loss function, and then combine the newly initiated trained weak classifier with the current model in the form of accumulation.

GBDT, known as gradient boosting decision tree, is formed by gradient boosting method using the decision tree as weak classifier. Gradient boosting refers to the core idea of gradient descent to search the optimal function. In the gradient descent, one sees that the final optimal solution θ^* is obtained after T times iterations of the initial value. Below are the specific algorithm steps. First of all, one needs to set

$$\Theta_0 = -\frac{\delta L(\theta)}{\delta \theta_0} \tag{1}$$

Subsequently, the definition is applied as:

$$\theta^* = \sum_{t=0}^{T} a_t * \left[-\frac{\delta L(\theta)}{\delta \theta}\right] \tag{2}$$

In this formula, one determines $\theta = \theta_{t-1}$, which requires to calculate the first derivative of the Taylor expansion of $-\delta L(\theta)/\delta(\theta)$. Based on algorithm above, the gradient boosting defines the loss function of the model as $L\big(y, F(x)\big)$. The mathematics goal is to reach the optimal function $F^*(x)$. The specific algorithm steps are as follows:

- Setting the initial value as: $F_0(x) = f_0(x)$
- Taking the function $F(x)$ as a whole, the updating process of it is concurred with that of gradient decent.
- The optimal function $F^*(x)$ obtained after T times iterations is:

$$F^*_{(x)} = \sum_{t=0}^{T} f_t(x) \tag{3}$$

where,

$$f_t(x) = -a_t g_t(x) = -a_t * \left[\frac{\delta L\big(y, F(x)\big)}{\delta F(x)}\right], F(x) = F_{t-1}(x) \tag{4}$$

It can be concluded the difference between these 2 methods is that, the gradient variable in gradient boosting is a function and is solved in the function space, while in gradient descent it is a parameter. Moreover, the gradient descent is in the negative gradient direction in the multi-dimensional parameter space. The update function in gradient boosting modifies the model through the negative gradient direction of the current function to make the model better to reach an approximately optimal method.

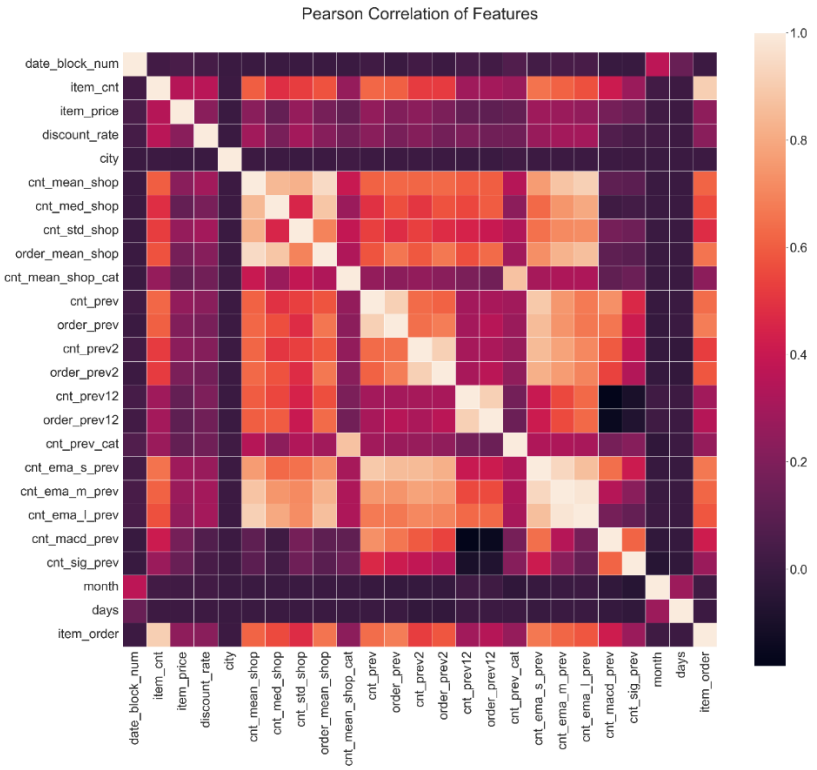

Figure 1. Heatmap of the features correlation coefficients.

2.2.2 Xgboost

Xgboost is a Boosting library developed by Tianqi Chen of The University of Washington in 2016, which has both linear scale solver and tree learning algorithm [10] and is an extension of gradient boosting. Boosting classifier is an ensemble learning model, whose basic idea is to combine hundreds of tree models with low classification accuracy into one that with high classification accuracy.

Compared with GBDT, the Xbgoost made a lot improvements, which are more efficient. The traditional GBDT method only uses the first-order derivative information, while Xgboost has both GB-tree and linear classifier to get linear regression or logistic regression with L1+L2 penalty and it also does the second-order Taylor expansion of the loss function. Moreover, it adds the regular term in addition to the objective function to obtain the overall optimal solution, which is used to weigh the decline of the objective function and the complexity of the model to avoid overfitting and improve the solving efficiency of the model.

2.2.3 LightGBM

Proposed by Microsoft in 2015 [11], LightGBM is a new Boosting framework model which introduces two new technologies on the basis of traditional GBDT: one is Gradient-basic one-side sampling technology and the other is Exclusive Feature Bundling technology. Gradient-basic

one-side sampling (GOSS) can eliminate a large part of data with a small Gradient and only use the remaining data to estimate information gain, thus avoiding the influence of the long tail part of low Gradient [12]. As the data with large gradient is more important for information gain, GOSS technique can still obtain quite accurate estimates under the premise of much less data than traditional GBDT.

Exclusive Feature Bundling (EFB) is able to reduce the number of features, which decrease the risk of getting over fitting in an efficient way. In addition, in LightGBM, the histogram algorithm is used to replace the traditional pre-sorted feature points to reduce memory consumption, which also simplifies and optimizes the most time-consuming steps taken in GBDT.

2.3 Metrics

This research uses RMSE, MAE and R^2 to measure the accuracy of the prediction models. The smaller the RMSE and MAE result, the more accurate the prediction. The closer the R^2 gets to 1, the more accurate the prediction is.

$$RMSE = \sqrt{\frac{1}{N}\left(-\sum_{i=1}^{n} x_{obs,i} - x_{mod\,el,i}\right)^2} \tag{5}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_{-}^{(i)}\hat{y}^{(i)}\right| \tag{6}$$

$$R^2 = 1 - \frac{ss_r}{ss_t} \tag{7}$$

Here, SS_r means residual sum of squares, and SS_t means total sum of squares.

3 RESULTS & DISCUSSION

3.1 Calculation process

Future sales prediction uses Windows11 operating system and python3.7 (64-bit) to build models. The important packages and libraries include pandas, Numpy, Matplotlib, Seaborn, Sklearn, Xgboost, and LightGBM. As LightGBM and Xgboost are used for analysis, the setting of parameters has a great influence on the prediction results, so some parameters need to be tuned. For The LightGBM model, this research mainly optimized 3 main parameters: learning rate, the number of iterations and the number of leaves. For Xgboost model, this research mainly tuned 4 main parameters: learning rate, the number of iterations, the depth of the largest tree, the sample weight of each leaf node. The specific steps are as follow:

- First adjust the learning rate parameter. Set learning_rate equal 0.05, use default value for others, and use GridSearchCV function to determine the appropriate number of iterations.

- Then, use the GridSearchCV function again to search the other two main parameters of the model automatically.

- Find the appropriate learning rate so that the number of iterations can reach the least when the error is at the minimum.

Moreover, for LightGBM and Xgboost, the model ‘plot _importance’ is used to extract the importance of features to the model.

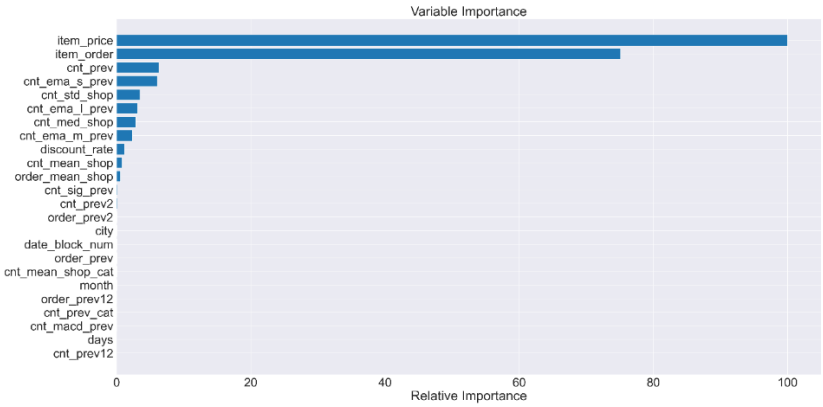

Figure 2. Feature importances of Xgboost.

3.2 Empirical analysis

The correlation information between important features can be obtained from Figure 1. In this figure, the darker the color, the weaker the correlation between two features, and the lighter the color, the stronger the correlation between two features.

For instance, it can be inferred from the dark color that the cnt_mean_shop and the order_mean_shop have strong shop correlation, which indicates that the average order quantity of stores in this data set is affected by the average number of stores. In addition, it can be inferred from the light color that the correlation between order_mean_shop and city is weak, indicating that the average order quantity of stores will not be affected by the region, and the two are highly independent. After visualizing the correlation information of the data, one can have a further understanding of the data, facilitating the data processing before prediction.

Table 1 Metrics of different models

	LightGBM	Xgboost
RMSE	0.193	0.234
MAE	0.019	0.024
R^2	0.969	0.955

According to Table. 1, the following conclusions can be drawn: LightGBM is superior to Xgboost. Based on RMSE, the accuracy of LightGBM is almost up to 0.19, while Xgboost optimal accuracy is up to 0.23. Moreover, lightGBM also performs better on the results of MAE and R^2, which also illustrates that LightGBM has a more accurate prediction than Xgboost in this prediction.

Seen from Figure. 2 and Figure. 3, it can be easily concluded that LightGBM provides a more specific and accurate result of feature importances, while Xgboost performs well only on few features. According to Figure. 2, combined with the ranking of feature importance of all data sets, it can be concluded that the key factors affecting sales are:

- The number of stores. Since this research forecasts the monthly sales of products in each store, the number of stores is in great significance in this prediction.

- Types and quantities of products supplied by the store. The importance of this indicator highlights that sales prediction depends on the varieties of product to a certain extent. The number of product types will affect the result of model prediction to in a degree.

- Product reservations. This feature importance shows that in the actual forecasting process, the orders of a product play an important indicator role in predicting future sales. Meanwhile, it can be inferred that the reservations of a product in the past also play reference role, as combining the past and present commodities' orders, the changing trend of commodities demand can be presented more macroscopically, so as to make a more accurate prediction.

- Product price. This indicator indicates that whether the product is expensive or not has an impact on the forecast of future sales to a certain extent, and this impact is also reflected in the discounts of the product. If one product has a recent sales discount, it will have a positive impact on the forecast of sales.

- Past sales of commodities. The past sales of a commodity can reflect the market demand for the commodity in a degree, so as to provide valuable reference experience for future commodities forecasting according to the past sales history.

- The forecast frequency also has an impact on sales prediction.

According to the feature importance in Figure. 3, shop count is the key indicator of impact, followed by item order, then item price and item sales. Features that show significant importance in Figure. 3 consist of these four indicators.

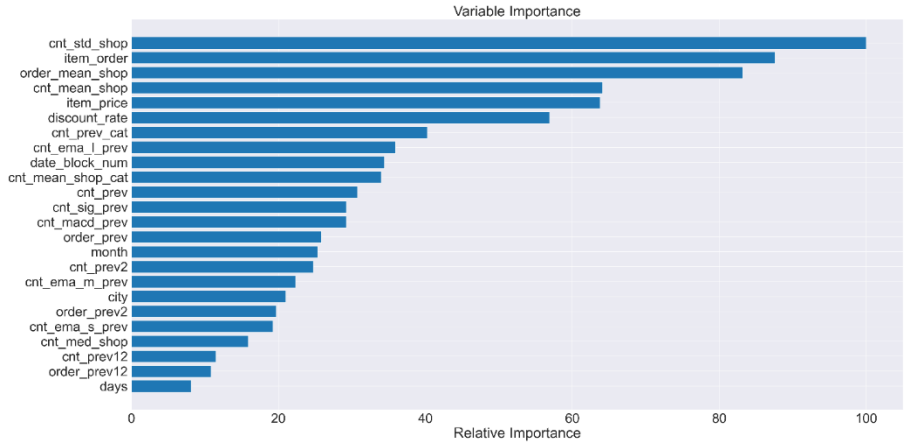

Figure 3. Feature importances of LightGBM.

3.3 Limitations

First of all, the deficiency of this research is that although LightGBM can get a good prediction accuracy, both Xgboost and LightGBM are based on heuristic algorithms, so the solutions sought are local optimal rather than global optimal. Subsequently, in terms of sample selection, because the sample includes the prediction of different products of different companies, the sample range is so wide that the prediction object is not targeted precisely.

4 CONCLUSION

In summary, lightGBM and Xgboost machine learning methods are used for commodities monthly sales prediction. LightGBM has better forecasting ability and accuracy compared with Xgboost. According to the prediction results, in the actual operation of the company, one needs to focus on the product category, product price, product reservation and other indicators. Meanwhile, the corresponding summary of past experience can also provide meaningful reference value.

In future research, methods related to feature selection, e.g., the recursive feature elimination with cross validation (RFECV) and the tree-based feature selection can be appropriately introduced to prediction, which can eliminate the correlation of features so as to reduce the risk of getting over fitting. Combining different kinds of methods can provide more accurate and comprehensive prediction. In addition, prediction can be made on different data sets and test whether this method can still play a good prediction role in an encounter with other situations, in order to expand the scope of application of the study. A good sales prediction can provide necessary guarantee for the sound operation and development of the company. Further investigation of this field is still necessary to develop and improve. Overall, sales prediction is an important cornerstone for the formulation of corporate strategy and these results offer a guideline for sales strategic arrangement.

REFERENCES

[1] S. Zhang, et al. "Big data and big data machine learning." Journal of Hebei University (NATURAL SCIENCE EDITION), vol. 38.3, 2018, 299

[2] A. L. Fradkov. "Early history of machine learning." IFAC-PapersOnLine vol. 53.2, 2020, pp. 1385-1390.

[3] M. I. Jordan, and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science, vol. 349.6245, 2015, pp. 255-260.

[4] D. W. Cravens, and N. Piercy. Strategic marketing. Vol. 6. New York: McGraw-Hill, 2006.

[5] N. Teng, et al. "Application of regression analysis in sales forecasting." Accounting of Chinese township enterprises, vol. 12, 2019, pp. 107-109.

[6] A. D. H. Winklhofer, and S. F. Witt. "Forecasting practice: A review of the empirical literature and an agenda for future research." International Journal of forecasting, vol. 12.2, 1996, pp. 193-221.

[7] N. R. Potharaju, and D. P. Reilly. "Reducing case fatality rate of acute encephalitis syndrome in developing countries." Journal of Pediatric Neurology, vol. 10.4, 2012, pp. 257-268.

[8] T. H. Davenport, et al. "How'big data'is different." MIL Sloan Managemnet review, vol 54. 1, 2012, pp. 22-24.

[9] Q. He, et al. "Overview of machine learning algorithms under big data." Pattern recognition and artificial intelligence, vol. 27.4, 2014, pp. 327-336.

[10] W. J. Rapaport. "Searle's experiments with thought." Philosophy of Science, vol. 53.2, 1986, pp. 271-279.

[11] R. I. Damper. "The logic of Searle’s Chinese room argument." Minds and Machines, vol. 16.2, 2006, pp. 163-183.

[12] Y. Tang. "Research on loan default prediction model based on Xgboost algorithm and lightGBM algorithm." Modern computer, vol. 27.32, 2021, pp. 33-37