Sales Prediction Based on Machine Learning Approach

Yifan Sun*

Department of Engineering, Virginia Tech, Virginia VA 24060, USA

*syifan@vt.edu

Abstract. To help brick-and-mortar merchants set reasonable sales goals, this study proposes and implements a retail sales forecast method based on machine learning theory. Specifically, this paper used the XGBoost model, LightGBM tree structure model, long and short-term memory network (LSTM) model and model fusion method, took the sales data of 1115 physical stores of Rossmann of Germany as the research object, used three single models and three fusion models to predict sales. First, the three single models were trained and verified through feature engineering and parameter tuning; then the three single models were fused via three weighted average methods with different weights, and the fusion model was optimized and verified. Finally, two evaluation indexes, MAPE and RMSPE, were implemented to evaluate the model, and the MAPE and RMSPE values of several models were compared. Experimental results indicated that the MAPE and RMSPE values of the single model were above 0.049 and 0.065, respectively, while the MAPE and RMSPE values of the fusion model were below 0.047 and 0.062, respectively. It showed that although the single model method was effective and feasible, the fusion method effectively improved the prediction accuracy and generalization ability of the model, and obtained better performance than the single model.

Keywords: Sales Forecast, XGBoost, LightGBM, LSTM, Model Fusion

1 INTRODUCTION

With the popularity of the Internet, the rise of technologies such as artificial intelligence, big data and cloud computing, retail enterprises are facing unprecedented opportunities and challenges. The massive sales data of retail enterprises contains huge business opportunities, and the demand for predictive analysis of enterprise business departments shows a trend of "explosive growth" [1]. At present, the important way to realize the value of data is to forecast the data through machine learning. Sales forecast will affect the retail enterprise's sales plan, inventory stability, etc., thus affecting sales stability. The success or failure of an enterprise depends on whether it can predict the future according to the existing data and formulate a suitable sales strategy. Only by making accurate sales forecast, can it timely update marketing decisions, reasonably set goals, improve management efficiency and operating efficiency, so as to maintain the stable development of the enterprise and obtain greater benefits.

R. Magdalena et al. (eds.), *Proceedings of the 2024 9th International Conference on Social Sciences and Economic Development (ICSSED 2024)*, Advances in Economics, Business and Management Research 289,
https://doi.org/10.2991/978-94-6463-459-4_113

Due to the influence of many factors, the sales forecasting of enterprises is a very complex nonlinear forecasting problem. However, with the continuous development of machine learning, this kind of prediction problem has been better solved. Luo analyzed the influencing factors of clothing sales, and used genetic algorithm to establish a back propagation neural network sales forecasting model [2]. H.Chan using several time-shifted weather features and machine-learning techniques to quantifying the impact of weather on sales of individual products.While G. Verstraete focus on the value of weather information for the forecasting to total sales or category sales[3,4].Florian Haselbeck find that the performance of state-of-the-art machine learning is better than classical forecasting algorithms for horticultural sales predictions,especially for data-sets with multiple seasons[5].Yu combined the extreme learning model and neural network model to predict the sales of clothing [6]. Through comparative analysis with BP neural network, the former prediction effect is better. Sun studied RNN network, LSTM network and BP neural network to predict the stock price trend of the US stock market, and then analyzed the characteristics and accuracy of the three networks [7]. In terms of analyzing time series and combination forecasting model, Dang only used the time series of historical monthly sales with timestamp to study the seasonal variation of supermarket sales by Autoregressive Moving Average (ARMA) model, and achieved good forecasting results in short-term forecasting [8]. Zhou et al. proposed a prediction model based on K-means clustering and machine learning regression algorithm to solve the sales prediction problem of multiple products in the retail industry [9]. The prediction effect of the model is significantly better than that of the machine learning model without clustering.

By referring to some new research and prior knowledge and aiming at the problems existing forecasting methods, this paper uses a combination of machine learning and deep learning to forecast retail sales, mainly studying the application of XGBoost and LightGBM tree structure model and LSTM model in sales forecasting. The purpose is to obtain a forecasting model with high prediction accuracy, strong learning ability, strong generalization ability and wide applicability, so as to guide merchants to make reasonable sales decisions and obtain greater profits. The contributions of this paper are divided into three main points as follows:

- Analyze and process sales data and store data. The data information is large and complex, and the focus is on processing of the data to find the law of commodity sales.
- Feature engineering construction. Based on the original data, it generates, extracts, deletes or modifies the features in the data set, obtains the relevant features required for building the model, and establishes the feature engineering.
- Construction and analysis of sales forecasting model. Firstly, the establishment of three single models and three fusion models is introduced, and then six prediction models are verified and analyzed on the verification set, all of which achieve good prediction results, and also prove that the performance of the fusion model is better than that of the single model.

2 DATA AND METHOD

2.1 Data Sources and Feature Engineering

The experimental data in this paper came from the sales prediction competition of the kaggle platform, using the historical data of 1,115 chain Rossmann physical stores in Germany to predict the sales in the future period. The whole original data set is divided into three parts: store information data set, training data set and test data set.

After obtaining the data, further feature engineering on the dataset is required. Feature engineering refers to the extraction, generation, modification and deletion of data features in the data set through the knowledge in the field or automated methods, so as to extract relevant features from the original data to learn and train the model to the maximum extent [10]. Feature engineering mainly includes three parts: data preprocessing, feature extraction and feature selection. The main purpose of data preprocessing is to clean the data, that is, to deal with missing values, outliers and duplicate data in the data, including data type conversion, text encoding and data segmentation. Feature extraction refers to the processing of original data to generate features meaningful to the target, and the transformation, conversion, combination and normalization of discrete features, time series features, cross features and text features to generate features conducive to machine learning. Feature selection refers to using a certain method to select features with higher relevance to the target to achieve feature dimension reduction, so as to select a more meaningful feature set input model for training.

Table 1. Parameter Initial value setting

Parameter type	Parameter	Default value
Learning task parameter	objective	reg:linear
	seed	0
Basic parameter	booster	gbtree
	nthread	Maximum number of available threads
	silent	0
Lifting parameter	eta	0.3
	max_depth	6
	gamma	0
	subsample	1
	lambda	1
	min_child_weight	1
	alpha	1
	colsample_bytree	1

2.2 Method

Based on XGBoost [11], LightGBM [12], Long Short Term Memory network (LSTM) [13] and model fusion method, three single models and three fusion models are used to forecast sales in this paper. XGBoost stands for extreme gradient lifting, which belongs to one of the algorithms in Boosting, an integrated learning algorithm. It is a large-scale

end-to-end, scalable and improved GBDT algorithm based on weak learner proposed by Chen [14], aiming to achieve fast and high-performance computing. Moreover, XGBoost adds the regular term to the loss function, avoids overfitting, improves the generalization ability of the model, reduces the noise of training data and largely makes the algorithm more robust. The core idea of XGBoost algorithm is to add a tree in each iteration, which is used to fit the residual difference between the predicted value and the real value in the last iteration, and constantly approximate the actual value. For data set D with m dimension and n samples, the formula of XGBoost prediction model is shown in Eq. (1):

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1,2,\cdots,n) \tag{1}$$

where f(x) is the regression tree (CART), K represents the number of regression trees, and F represents the set space of the regression tree. Setting initial values for the XGBoost prediction model based on experience is shown in Table 1. Light Gradient Boosting Machine (LightGBM), proposed by Microsoft in 2017, is a framework based on GBDT algorithm, which builds a decision tree according to the leaf growth strategy and limits its maximum depth. Prevent overfitting while ensuring efficiency. It can be used in a variety of machine learning problems such as regression, sorting, prediction and classification, and can achieve efficient parallel training. Compared with GBDT and XGBoost, it has the advantages of open source, fast training speed, low memory usage, high accuracy, strong interpretability, strong generalization ability, etc., while supporting distributed training and processing large-scale data. The algorithm principle of LightGBM and XGBoost is the same, the difference is that LightGBM is optimized on the framework, mainly the optimization of the training speed, and the optimized LightGBM speed is 10 times that of GBDT and 3 times that of XGBoost. Since the parameters of LightGBM model are relatively complex, and the more complex the parameters, the greater the time complexity of parameter adjustment, the optimization of some important parameters is given priority. In the process of parameter optimization, GridSearchCV, a grid search algorithm on sklearn, is used, as shown in Table 2.

Table 2. LightGBM model grid search parameter table

Parameter	Range of values
max_depth	3,4,5,6,7,8,9,10
max_bin	62,128,200,256,512
feature_fraction	0.5,0.6,0.7,0.8,0.9
num_leaves	60,70,80,90,100

Long Short Term Memory was proposed by Sepp Hochreiter and Jurgen Schmidhuber in 1997, and has been improved and implemented by Alex Graves recently. Based on the improvement of recurrent neural network, the LSTM neural network algorithm is obtained. It adds the function of short-duration memory on the basis of RNN. While maintaining the persistence of RNN network, the model can be dependent for a long time, and can remember longer information and forget the information that does not need to be saved. As shown in Fig. 1, the internal structure of LSTM contains many

memory modules, and each memory module is mainly composed of cell, input gate, forget gate and output gate. Multiply input gate by cell input to control the input of information, forget gate to control the forgetting and retention of historical cell status information, output gate to control the output of information, f is the activation function of the three gates. Generally, sigmoid function is used, which controls the information valve to filter the input data and determine which input data enter the model, that is, to determine which information to remember or forget. Table 3 shows the value range of each parameter of the LSTM model.

In order to improve the forecasting accuracy of the model, the model fusion strategy is used to forecast sales [15]. Considering the large differences between XGBoost, LightGBM and LSTM models, the weighted average method fusion strategy is adopted for fusion. For the weighted average method, different weights can be set. In general, the weight of the single model with good prediction effect is set larger, and the weight of the single model with poor prediction effect is set smaller. From experience, the weight value set manually can achieve good prediction effect. Therefore, when the three single models of XGBoost, LightGBM and LSTM are used for weighted fusion, the weights of the three single models (1,1,2), (1,2,1) and (2,1,1) are manually set respectively to obtain three fusion models of M1, M2 and M3.

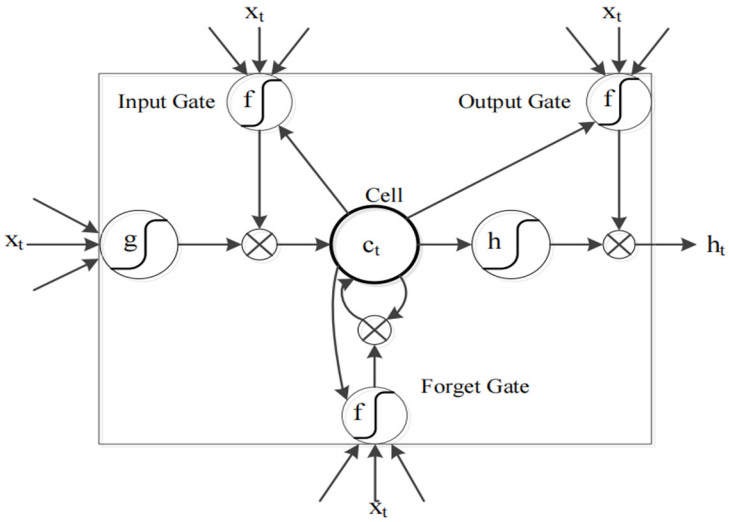

Fig. 1. LSTM structure diagram (Photo/Picture credit: Original).

Table 3. LSTM model grid search parameter list

Parameter	Range of values
time_step	1,2,3
epochs	600,700,800,900,1000,2000
batch_size	10,20,30,40,50,60
Number of neurons n	10,20,30,40,50,60

3 RESULTS AND DISCUSSION

The experimental operating system is Windows10 X64, the experimental platform is Anaconda, the programming language is python, the deep learning framework is TensorFlow, and the CPU model is Intel(R) Core(TM) i5-12400H. The memory and hard disk sizes are 16GB and 1TB respectively. Experimental evaluation indexes MAPE(mean absolute percentage error) and RMSPE(Root Mean Square Percentage Error), which are commonly used in regression prediction, were selected to objectively evaluate the prediction results of the model. The smaller their values are, the better the model predicts. The final experimental results are shown in Table 4.

Table 4. Experimental results of each model

Model	MAPE	RMSPE
XGBoost	0.056174	0.076943
LightGBM	0.053723	0.072415
LSTM	0.049036	0.065460
M1	0.046002	0.060880
M2	0.046168	0.060935
M3	0.046508	0.061034

By comparing the prediction results of XGBoost, LightGBM and LSTM models, the three single models have achieved good fitting effect, and the fitting effect of LSTM model is better than that of XGBoost and LightGBM model. The fitting effect of LightGBM model is better than that of XGBoost model. Compared with the evaluation index values of the three single models, the three models all achieve ideal prediction results, but the MAPE value and RMSPE value of XGBoost model are the largest. The MAPE value and RMSPE value of LSTM model are the smallest, which indicates that the three single models are feasible and effective in sales forecasting. The LSTM model works best because it solves long-term dependency problems well. Its internal structure contains many memory modules, and the input gate, forgetting gate and output gate of each memory module control different information transmission respectively, using the sigmoid function, which can ensure the best parameters after the training of the three gates. On the other hand, the MAPE value and RMSPE value of the single model are higher than those of the fusion model, which shows that the model fusion strategy is effective, and further indicates that the prediction effect of the fusion model is better than that of the single model. This is because this paper studies a regression prediction problem of continuous target variables. Although most models can effectively solve this kind of problem, the model principle is different, and the predicted results are very different. The prediction ability of the three single models proposed in this paper is relatively strong, but the accuracy of the prediction needs to be improved. The model fusion can make full use of the advantages of each single model, improve the prediction accuracy and stability of the model, and avoid the bias of the single model. XGBoost and LightGBM models are tree models, while LSTM models are neural network models. As the principles of these two models are quite different, the results produced are

not highly correlated. Therefore, adopting model fusion strategy can effectively improve the accuracy of prediction results.

4 CONCLUSION

To sum up, this study takes the sales data and store data of 1115 chain physical stores of Rossmann Company in Germany as the research object to forecast its sales. Firstly, the source and specific information of the data are introduced, and the feature engineering of the data is carried out. Then, the sales prediction of XGBoost model, LightGBM model, LSTM model and model fusion are compared. The experiment proves the validity and feasibility of each model, and analyzes the problems existing in the model. Finally, the experiment shows that the fusion model obtains better prediction effect, which is better than the single model in terms of model accuracy and evaluation criteria.

Although the model constructed in this paper has achieved good results in the prediction effect, there are still some areas that need to be improved, as follows. For complex time series regression forecasting problems, other models can also be tried to predict. In addition, the model fusion strategy proposed in this paper is relatively simple. Besides weighted average, other fusion strategies can be used to improve the prediction accuracy of the model. In the process of model training, mesh search algorithm is used to optimize the parameters in this paper, but only the optimal parameters are used in the end. As a result, there is still room for optimization in parameter selection of the model. This study mainly analyzes and forecasts the characteristics of the store itself. Besides the factors of the store itself, there are other non-self-factors that affect the sales forecast. In the later stage, the influence of non-self-factors on the sales of the store can also be studied. The experiment of this study preliminarily realizes the sales forecasting model based on machine learning theory, and shows the feasibility, effectiveness, and universality of the sales forecasting model through theoretical research and experimental results. However, it is still necessary to further study and explore new machine learning algorithms and establish a more perfect sales forecasting model to further improve the performance and make it more widely used.

REFERENCES

1. Zhou, Y., Duan Y.: Retail sales forecasting based on clustering and machine learning. Application of Computer System, 30(11), 188-194 (2021).
2. Luo, R., Liu, S., Su, C.: BP neural network clothing sales forecasting method based on genetic Algorithm. Journal of Beijing University of Posts and Telecommunications, 37(4), 39-43 (2014).
3. G. Verstraete, E.-H. Aghezzaf, B. Desmet:A data-driven framework for predicting weather impact on high-volume low-margin retail products. J. Retail. Consum. Serv. 48, 169–177(2019).
4. H. Chan, M.I.M. Wahab: A machine learning framework for predicting weather impact on retail sales,Supply Chain Analytics,5,2024.

5. Haselbeck, F., Killinger, J., Menrad, K., Hannus, T., & Grimm, D. G.: Machine learning outperforms classical forecasting on horticultural sales predictions. Machine Learning with Applications, 7, 2022
6. Yu, M.: Research on Mid-term Sales Forecasting of Clothing Based on Machine Learning Theory. Zhejiang University of Science and Technology, 2014.
7. Sun, R.: Research on Prediction model of US Stock Index Price Trend Based on LSTM neural Network. Beijing: Capital University of Trade and Economics, 2015.
8. Dang, J.: Application of ARMA Time Series Model in Sales Forecasting. Computer & Telecommunication, 4, 55-57 (2010).
9. Zhou, Y., Duan, Y.: Retail Sales Forecasting Based on Clustering and Machine Learning. Application of Computer System, 30(11), 188-194 (2021).
10. Xie, B., Lin, S., Lin, Z. et al.: Research on Feature engineering algorithm based on Reinforcement learning. Application of Electronic Technique, 47(07), 29-32+43 (2021).
11. Guo, L., Wang, Y.: Research on prediction of grain storage temperature based on XGBoost optimization algorithm. Grain & Oils, 35(11), 78-82 (2022).
12. Dai, J., Zhong, X., Wang, M.: Used Car Valuation Based on LightGBM and Random Forest Algorithm. Journal of Normal University Science and Technology, 42(12), 15-22 (2022).
13. Lu, Z., Zheng, J., He, L.: Research on sales Forecasting based on improved SSA-LSTM. Computer Times, 10, 50-53+58 (2023).
14. Liu, Q., Zhang, B.: Residual Useful Life Prediction of lithium battery based on GBDT Algorithm. Journal of Electronic Measurement and Instrument, 36(10), 166-172 (2022).
15. Ma, W., Chen, S.: Research on e-commerce customer churn prediction model based on multi-model fusion. Wireless Internet Technology, 19(17), 143-145 (2022).