Journal of Applied Mathematics and Computation, 2024, 8(1), 83-87
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

Product Demand Forecasting Based on LGBM Deep Learning Models

Yanting Liao, Chufeng Yang, Sisi Zheng*

School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong, China.

How to cite this paper: Yanting Liao, Chufeng Yang, Sisi Zheng. (2024) Product Demand Forecasting Based on LGBM Deep Learning Models. *Journal of Applied Mathematics and Computation*, 8(1), 83-87.
DOI: 10.26855/jamc.2024.03.010

Received: February 28, 2024
Accepted: March 27, 2024
Published: April 26, 2024

***Corresponding author:** Sisi Zheng, School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong, China.

Abstract

As the first line of defense of the enterprise supply chain, product demand forecast plays an important role in enterprises in different industries, so it is of research value to forecast it accurately. This paper preprocesses the product demand data set of a large domestic manufacturing enterprise, and through the exploration of data characteristics, it is concluded that seven different variables have a certain influence on product demand. Using feature engineering, a new column is constructed, and it is coded by single heat and map mapping. By using lag features and window statistics, 49 data features are constructed and screened. Four machine learning models, such as XGBoost, LightGBM, decision tree regression, and random forest, are constructed respectively, and the parameters are compared vertically with the grid parameters. Three indexes, mean square error (MSE), root mean square error (RMSE) and determinable coefficient (R2) are selected to evaluate the model. Accurate estimation can reduce the inventory cost of enterprises and make a pricing scheme with higher information content.

Keywords

Characteristic engineering, LightGBM model, Net style ginseng, Product demand, Predict

1. Introduction

In recent years, the external environment of enterprise management has been changeable, and the upstream and downstream supply chains and markets of all walks of life are facing great challenges [1]. With the rapid development of the market, customer demand under customized production mode is often dynamic, multifaceted, and uncertain. In order to cope with this fast-changing market and improve the overall production level of enterprises [2]. Enterprises need to predict customer demand in advance, understand the key economic situation, and make key supply decisions. Product demand forecast is not only the first line of defense of enterprise supply chain, but also a theoretical conclusion based on historical data and future prediction. Accurate product demand forecast is an important support for enterprises to expand market share and improve profits [3]. This paper is based on the background of Question B of the 11th "Teddy Cup" data mining challenge in 2023 (this paper won the third prize in China and the second prize in Guangdong Division, data source[1] [4]. From the perspective of machine learning, four different product demand prediction models, namely LightGBM, XGBoost, Random Forest and Decision Tree, are established to predict the product demand, and the parameters are optimized by using the method of grid parameters. Through horizontal and vertical comparison, the best model is obtained, which has certain practical significance for forecasting the product demand of enterprises in different industries.

[1] https://pan.baidu.com/s/1_E1GcUbIa0SoIg17BxcSxA? Pwd=tbl0 extraction code: tbl0).

2. Data Cleaning

Because the analysis and prediction of product demand rely on historical data, duplicates and outliers exist in these data, which may have an impact on the results. Therefore, data cleansing is required to better predict product demand. Data cleaning is the process of rechecking and verifying data in order to remove duplicate information, correct errors, and ensure data consistency. It includes the processing of duplicate values, missing values, and outlier values.

During data missing value processing, each variable's null value is identified and filled, including statistical and rule filling. The test results indicate that the data set contains no missing values. After analyzing the missing value in the dataset, repeat value processing is performed. The dataset is then processed using the duplicated () method of the dataframe class in the Python package, removing 312 duplicate value data. Among them, the duplicate value refers to data where each column and row are identical for more than two rows. To improve the accuracy of the data set, outlier processing is also performed. Common techniques for detecting outliers include simple statistical analysis, the 3δ principle, boxplots, clustering algorithms, etc. Here, a boxplot is selected for outlier detection in order to display the distribution information of the population sample from a macro perspective and detect outliers simultaneously. By observing and analyzing the price data of products with product codes 301 and 303, it is determined that 303's price is high, 301's price is low, and 301 has abnormal data greater than 250,000, which is contrary to common sense. Therefore, the outlier should be eliminated following careful consideration. In order to make the data set more visual, the order_date format was converted into three columns: year, month, and day.

3. Characteristic Engineering

3.1 Construction and Coding

In order to better predict product demand, we need to explore factors such as product price, product location, and different sales methods in the data set. In addition, it is also necessary to consider the impact of different time periods (beginning, middle, and end of the month), seasons, holidays, promotion days, and other factors not reflected in the data set on product demand, so as to determine the variables related to product demand and help us build a better forecasting model.

(1) The product price was divided into three levels: low, medium, and high, and the groups' function was used to carry out statistical processing on the data set of low, medium, and high grouping, and then the plt.ylabel() function was used to carry out statistical analysis on the product demand, and the conclusion was drawn that the product demand was the highest when the product price level was low.

(2) According to the grouped statistics of different sales area codes, the analysis shows that the average demand for product code 104 is the highest, while the total demand is the lowest. This may be due to factors such as product prices and promotions.

(3) Group statistics, effective values, maximum and minimum values, mean value, and standard deviation of the data are processed. Regardless of online or offline sales, with the increasing demand for product orders, the number of nuclear densities generally increases first and then decreases. When the order demand is 0-100, the kernel density number of offline sales is higher than that of online sales, indicating that offline sales have advantages over online sales when the order demand is small.

(4) The product demand corresponding to the product category code and the product category code is screened out on the data set, and the product demand is divided according to the month. When the product category code is 308, the product demand fluctuates the most in December 2016. When the product category code changes with the number of days, the product demand fluctuates little. The average product demand corresponding to different product subcategories is also changes with the change of months, while the product subcategories are coded as 401 and 412, and the changing trend is slower than that of other products.

(5) Organize the data set to obtain the data on holidays and promotion days, among which the peak value is mostly holidays and promotion days. It is not difficult to conclude that holidays and promotion days have a significant impact on the average product demand.

(6) The year, month, and day were extracted and the line chart was drawn. It was found that the average product demand was also changing with the continuous change of time and month. On the 25th day, the average demand reached the lowest value, close to 65.0; On the 30th day, the highest peak is reached, that is, about 81.0.

(7) By integrating the influence of all dates and corresponding seasonal factors on the order demand and conducting correlation coefficient analysis, it can be seen that the influence of spring and winter festivals on the product demand is greater than that of summer and autumn seasons.

Based on the analysis presented above, it can be concluded that product price, product location, different time periods, holidays, promotion days, seasonal factors, and other variables affect product demand. Therefore, these factors must be

fully considered when developing machine learning models.

3.2 Lag Characteristics and Sliding Window Statistics

3.2.1 Construct new columns based on the data feature exploration results

(1) Promotion day column: Select the three most classic promotion days “6-18”, “11-11”, and "12-12", and use the double branch function to judge whether the date of each row is a promotion day.
(2) Date column: Construct columns D, W, and M, corresponding to "day", "week", and "month", respectively, that is, day, week, and month as granularity.
(3) Season column: The four seasons are spring (March to May), summer (June to August), fall (September to November), and winter (December to February).

3.2.2 Encode and process the influencing factors

This paper employs unique thermal coding and map mapping coding, i.e., mapping keys to objects of value, when analyzing the results of variable selection. A mapping cannot contain duplicate keys, and each key can be mapped to a maximum of one value [5]. The map mapping function map(dict_) is primarily used to encode "promotion day", "holiday", "sales mode", "month time", and "season time".

3.2.3 Hysteresis characteristic and Sliding window statistics

The lag feature, as a specialized sliding window, is a conventional method for transforming the time series prediction problem into the supervised learning problem. As shown in

If the product demand on December 20, 2018, needs to be predicted, and the length of the time window is assumed to be 3, then [106,187,205] can be used for prediction. If y is moved down a time window in order to construct multiple time features, the blue box in the column below can be transformed into a horizontal blue box. At this time, if the product demand on December 21, 2018, is to be predicted, it must be [205,187,106].

Introducing a lag for the target variable of sales, the maximum delay used in this model is sixty days. Structure s = df. Groupby ([' sales area code ', 'product coding, coding product categories,' product fine class code]). Lags of need_lag_ lag data are constructed using lags = [1,2,3,6,12,24,36,48,60].

Sliding window statistics, also known as rolling statistics and running statistics, compute a moving weekly sales average. More features, such as rolling minimums, maximums, or summations can also be calculated. Likewise, the same function can be used to calculate requirements.

Definition window=7, That is, the window size is 7, and a subset of the DataFrame is processed in batches using the rolling() function. In the case of a window that scrolls forward in time, the window's size remains constant. Therefore, only recent values are considered, while past values are disregarded. In this paper, the expanded mean of the demand will be calculated.

In this paper, the construction and selection of features are carried out, and 49 data features are created and finally screened out from the original data [6].

4. Model Construction and Solution

4.1 Model Evaluation Strategy

There are many evaluation indexes of machine learning models. In this paper, the mean square error (MSE), root mean square error (RMSE), and determinable coefficient (R) are selected to evaluate the training set and test set of each model. Through the evaluation indexes of cross-validation set, the hyperparameters can be continuously adjusted to obtain a reliable and stable model.

4.2 Model Parameter Adjustment Strategy

In this paper, the grid search method is mainly used to find the optimal parameter combination, and a subset in the hyperparametric space is randomly specified for traversal search to achieve the purpose of parameter optimization. Grid search exhausts all combinations of given parameters, trying to find an optimal set of hyperparameters.

4.3 Model Result Analysis

The data set filtered by feature engineering is used to establish a single model prediction [7-10]. Aiming at the further parameter optimization of the model, the mean square error evaluation index is used to evaluate the model before and after optimization, and it is found that the prediction effect of the model is further improved after parameter optimization. The

model evaluation before and after parameter optimization is as follows.

Table 1. Evaluation of each model

	Model evaluation before optimization			Optimized model evaluation		
Model name	MSE	RMSE	R^2	MSE	RMSE	R^2
XGBoost	9033.27	95.04	0.76	8968.79	94.70	0.76
LightGBM	7689.64	87.69	0.797	7213.21	84.94	0.81
Decision tree regression	12618.32	112.33	0.657	11012.97	104.94	0.69
Random forest	10215.29	101.07	0.674	10188.45	100.94	0.682

By optimizing the parameters of a single model and cross-validation, the mean square error of each model is reduced, and the small difference of mean square error of these models indicates that the model effect is good, among which LightGBM has the smallest mean square error and the best model performance. The fitting diagram of the output true value and the predicted value is as shown in the figure.

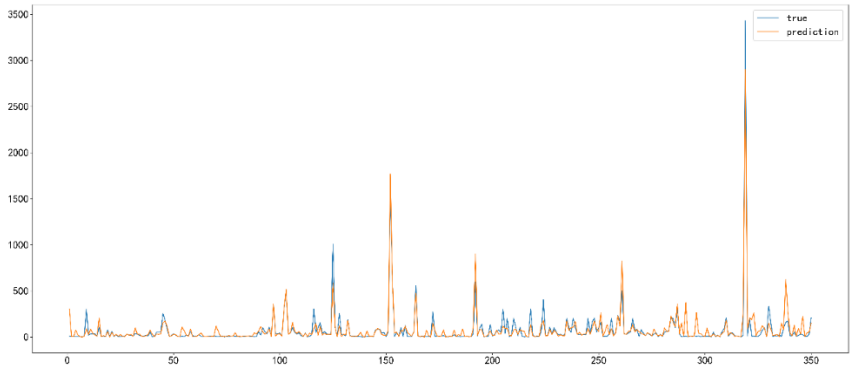

Figure 1. Fitting diagram of true value and predicted value.

5. Conclusion

It is inevitable to process data when building a machine learning model, and better prediction results can be achieved by properly handling missing values and abnormal values. In this paper, 49 features are finally obtained by constructing new columns, coding, constructing, and screening new features by using lag features and window statistics. Four different machine learning models are established, the parameters are optimized, the optimal parameter combination is found, and the influence of different parameters on the model results is analyzed. Using MSE, RMSE, and R2 to evaluate, the forecast result of LightGBM model is better than that of XGBoost, decision tree, and random forest model, so this model can be used to forecast the product demand of enterprises. An accurate prediction can reduce the inventory cost of enterprises, better confirm the development direction and specific goals of enterprises, and make pricing with higher information content, which has far-reaching significance for major enterprises. In this paper, a number of machine learning models are constructed to predict product demand, but there are still some shortcomings, such as the product data set of a large domestic manufacturing enterprise always has the problem of high dimensions and small samples, and different machine learning algorithms have different sensitivity to the distribution of data samples, so it is necessary to choose appropriate methods to achieve the ideal effect. The prediction of product demand is still the difficulty and focus of future research, and various algorithms need to be tried in practice to better introduce machine learning and other technologies into the quality prediction of intelligent products.

Funding

This paper is supported by the Guangdong Basic and Applied Basic Research Fund of Provincial and Municipal Joint Fund Project (2022A1515110437); Guangdong University Key Platform and Research Project Young Innovative Talents Project (2018KQNCX251); Teaching Quality and Teaching Reform Project of Huizhou University (X-JYJG2021047).

References

[1] Zhou Huiting. Study on Optimization of pigment supply Chain Demand Management of B Company [D]. Shanghai International Studies University, 2022.

[2] Wei Wei, He Leiyong, Li Chuihui. Demand forecasting method of electric locomotive products based on genetic algorithm and grey neural network [J]. Packaging Engineering, 2022, 43(12):37-44.

[3] Wu Gengqi, Niu Dongxiao, Geng Shiping, et al. Product Demand Forecasting of Manufacturing Enterprises Based on Deep Learning Algorithm from Multi-value chain perspective [J]. Science Technology and Engineering, 2021, 21(31):13413-13420.

[4] Rui Si. The 11th "Teddy Cup" Data Mining Challenge [EB/OL]. https://www.tipdm.org:10010 /#/competition/1620719578957127680/question. [2023-05-20].

[5] m0_59138290. Map mapping. https://blog.csdn.net/m0_59138290/article/details/127562412 [2023.05.21].

[6] Xu Jiang, Zhang Hongyu, Li Junhuai, et al. Parallel processing of streaming data based on Sliding window [J]. Heavy Machinery, 2021, (01):29-36.

[7] Zhu Fuxi. Artificial Intelligence [M]. 3rd edition. Beijing: Tsinghua University Press, 2012.

[8] Cai Jingbo, Cai Zhijie. Prediction of high turnover in the A-share market based on Machine learning Model [J]. Mathematical Modeling and Application, 2020, 9(04):74-84.

[9] Lu Weimin, Sun Chenfeng, Ren Likun, et al. An aero-engine Gas-path Fault Diagnosis Method Based on CN-LGBM [J]. Journal of Armaments, 2023: 1-11.

[10] Yu Yuan. Decision Tree-Regression: https://zhuanlan.zhihu.com/p/42505644. [2023.05.21]