Comparative study on supply chain demand forecasting based on traditional machine learning models—A Case Study Using the DataCo Dataset

Can Liu*

School of Big Data and Supply Chain, Guangzhou College of Technology and Business, Guangdong, China

Abstract. Good demand forecasting can reduce operating costs and improve service levels. Once the traditional time series model encounters those slightly complex nonlinear laws, it will be more difficult to handle, so machine learning methods are used more and more. In this study, DataCo global supply chain data set was used to compare the three models of logistic regression, decision tree and random forest. First, the data was cleaned, and then the characteristics of time dimension and the statistics of rolling window (including the average value and standard deviation of seven and thirty days) were built. The delay feature was also introduced, and the two indicators of mean absolute error (MAE) and coefficient of determination (R^2) were selected in the evaluation. The model completed the training on 80% of the data, and the remaining 20% was reserved for testing. The experimental results show that the random forest gets the minimum MAE and the maximum R^2, the decision tree is sandwiched in the middle, and logistic regression ranks last because of its linear characteristics. In addition, it can be seen that incorporating lag features and sliding window features benefits tree-based models significantly. Although the data set used is relatively old and there is no hyperparameter tuning, the experimental results show that the random forest has achieved a relatively stable balance between accurate prediction and easy deployment, which can provide a reference for enterprises when selecting demand forecasting tools.

1 Introduction

The development of digital economy has been slowly deepening, slowly turning the modern supply chain into a complex network spanning many links. The rapid change of market demand and the frequent occurrence of uncertain events have brought huge burden to the management work. The advantage is that in the business that runs around every day, a lot of operation data are saved in every link. In fact, there are many signals that can see the change of demand. Demand forecasting is directly related to how to prepare inventory, how to arrange production lines, and how to schedule logistics. The methods used in the past, such as the moving average method or ARIMA model, are often unsatisfactory when they encounter data with high dimensions and easy to jump up and down, so the machine learning method has attracted the attention of researchers and practitioners. Ma and Zhang pointed out in the research report that the demand data of the retail supply chain itself is clustered. If the demand differences between different clusters are ignored, the actual effect of the prediction method will be directly pulled down [1].

Making the demand forecast more reliable is helpful to reduce the operation cost. According to relevant statistics, enterprises that use advanced analysis methods well can reduce the operation cost by about 8% to 12% and improve the service level by 15% to 20%. Logistic regression, decision tree, and random forest are classic machine learning models. The reason is that they are not complex in structure, require low computational power, and the results are easy to understand. When used in the actual production environment, they still have their own places to use. These three models can be put on the same set of data to conduct comparative research one by one. At present, few studies have done so, so the last direct comparative experiment can provide a reference for those enterprises that are planning which forecasting tool to choose.

In recent years, a lot of achievements have been accumulated in thinking about supply chain forecasting along the method of machine learning. Zhang et al. built the CPSO-LSTM model [2] and put it on the order forecast of the packaging industry. Looking back at the demand for 30 days, the average estimate can reach 98.75%. Tan and Qiu used the real delivery flow of manufacturing enterprises in 2025 to compare xgboost, lightgbm and random forest by month, week and day, and found that as the time unit of prediction becomes smaller, the accuracy will also decline [3]. Liu and Wang combined LSTM with random forest in 2025 and applied it to the demand forecasting of e-commerce. The result is that the combined model is better than the single model [4]. Sattar et al used the same set of DataCo data as this article, compared xgboost with the cyclic neural network, and introduced two indicators, cost precision

*LccLxx@outlook.com

efficiency and ESG evaluation. After considering the two dimensions of cost and sustainability, they found that the performance of random forests was very high [5]. Most of the above mentioned focus on a single algorithm or a combined strategyThe three basic and easy-to-use models of logistic regression, decision tree and random forest can be put together. The part for horizontal comparison needs to be further filled in. In addition, Javed and Akhtar put forward a sales demand forecasting framework supported by several machine learning algorithms, and put the decision tree, random forest and xgboost into the retail scene[6], but when they scored, they mainly focused on the prediction error. The complexity of the model itself and the cost of future deployment were not included in the evaluation.

This paper intends to take the DataCo global intelligent supply chain dataset as the foundation and build three models of logistic regression, decision tree and random forest to predict the demand.Using MAE and R^2 as evaluation metrics. Through their comparison on the same set of data, where are their strengths and weaknesses when these three types of models encounter the same supply chain data. By this horizontal comparison, the paper can find out the differences in the prediction performance of the three types of models in the same supply chain data, clarify their respective advantages and limitations, so that when enterprises really choose the prediction tools, they can choose the prediction tools that are efficient, deployed and not too difficult to use, and provide a clearer reference basis.

2 Research Methods

2.1 Data Source and Processing

The data used in this article are the DataCo global intelligent supply chain data set [5], which can be downloaded publicly from Kaggle. There are about 180000 order records in total, which lasted from January 2015 to December 2017. The data covers the whole process from purchase, production, sales to distribution, including 53 fields in total. The items closely related to this study include order date, shipment date, product category, sales quantity, sales volume, profit, customer location and transportation method.

The dataset was first cleaned as followsIf more than half of the fields were missing, they were deleted directly; If the value field is not seriously missing, fill it with the median; In the category field, pick the value with the most occurrences and fill it up. Then the paper checked the abnormal situation in the data. The record of negative sales was obviously wrong, and it was directly deleted. The data that deviated from the average by three standard deviations were also removed. The last thing to do is to split the order date field and bring out the year, month, date and day of week from it, thus providing the basis for the later feature construction. Similar methods of date disassembly and time granularity segmentation have also been used in the research of Tan, Qiu [3] and Chai [7].

2.2 Feature Construction

Feature construction is roughly divided into three categories, including time-related features, historical statistical features, and delay features, which are all added to about 20 input variables. The time related feature is the year, month, day, and day of week separated from the date. In addition, a binary feature indicating whether the day is a weekend is added, because the demand for weekdays and rest days is often much different. Zheng et al. verified the effect of time convolution network (TCN) on grasping time-series dependence in the task of e-commerce commodity demand forecasting, but the training process of that method is relatively complex [8]. The method of sliding window is adopted for historical statistical features. The specific method is to take the sales data of seven days and thirty days forward of the current date and calculate an average value and a standard deviation respectively. The average value can reflect the level of recent demand, and the standard deviation can see the degree of fluctuation during this period. These two features are beneficial for the model to grasp the short-term trend. Lag features are constructed by using the sales volume of the previous days as the input of today. Yesterday's sales volume, the sales volume of the previous day and the sales volume of a week ago may be related to today's sales volume. The advantage of this is that it allows the model to capture the general direction of changes in the past few days, rather than only the current data. In addition, there are category features to deal with. Product category, customer location and mode of transportation are all text, not numerical values. They can not be directly thrown into the model, so they are converted into numerical form by using one-hot encoding. All numerical features have been standardized before the training model. The purpose is to unify the dimensions and avoid some particularly large values, such as sales, while ignoring the role of other features.

2.3 Model Selection

This study selected three types of machine learning models that are commonly used for comparison, namely, logistic regression, decision tree and random forest[4]. The purpose of selecting these three models is to observe the performance differences caused by different complexity in the same task. Li used xgboost alone in the cross-border e-commerce inventory forecast and got good results, but the study did not compare xgboost with the traditional linear model and a single decision tree[9]. Zhang and Yang applied logistic regression in supply chain financial credit risk prediction; they then compared XGBoost, random forest, and other models. The results show that under different evaluation dimensions, different models have their own advantages[10]. In this study, logistic regression serves as a benchmark model. Its basic idea is to assign weights to each feature, and then calculate the predicted value by weighted summation. The advantage is that it can be trained quickly and the result is easy to understand. The disadvantage is that it can only grasp the linear

relationship, and it is not feasible when encountering complex data laws. The decision tree is a model built with a tree structure. For example, it is a bit like following the "yes or no" judgment layer by layer. Finally, it comes to a result. It can automatically capture the interaction between features. The whole judgment process is clear. You can see which feature is judged first and which is judged later. However, a single tree has an obvious defect, that is, it has learned too much about the training data, and the performance will fall down with a new set of data. This phenomenon is called overfitting. Random forest is equivalent to upgrading the decision tree. In short, it is to train many decision trees at the same time. The data seen by each tree is different. Finally, the prediction results of all trees are combined to take an average. In this way, even if one tree makes mistakes, other trees may not make mistakes together. On the whole, it is relatively stable, and it is not so easy to fall into the trap of over fitting. Moreover, random forest is not sensitive to noise and outliers in the data, so it is very popular in practical applications.

2.4 Experimental Setup and Evaluation Method

The experimental environment uses Python 3.13, uses scikit-learn to build the model, pandas and numpy are responsible for data processing, and Matplotlib is used for visualization. In terms of data division, the first 80% of the data is allocated to the training set for the model to learn, and the last 20% is reserved for the test set to test the model. The training set used by the three models is identical to the test set, which is relatively fair.

Two indicators are used to evaluate the prediction effect. The average absolute error (MAE) depends on the average difference between the predicted value and the real value, the smaller the better; The determination coefficient (R^2) is between 0 and 1. The closer to 1, the better the fitting of the model to the data law.

3 Research Results

On the DataCo dataset, the prediction performance of the three models of logistic regression, decision tree and random forest are compared together. Two indicators are selected for the evaluation, one is the average absolute error (MAE), the other is the coefficient of determination (R^2). The results of the three models running on the test set are listed in Figure 1.

From the results of the operation, the random forest achieved the best results in both indicators, with the lowest MAE and the highest R^2. The MAE and R^2 of the decision tree are sandwiched between the random forest and the logistic regression. It can be seen that the generalization ability of a single decision tree is indeed poor compared to ensemble methods. The maximum MAE and the minimum R^2 of the logistic regression show that the linear model is hard to deal with when it encounters those fluctuations without clear linear laws in the supply chain demand data. The performance gap between the three models can be seen in Figure 1.

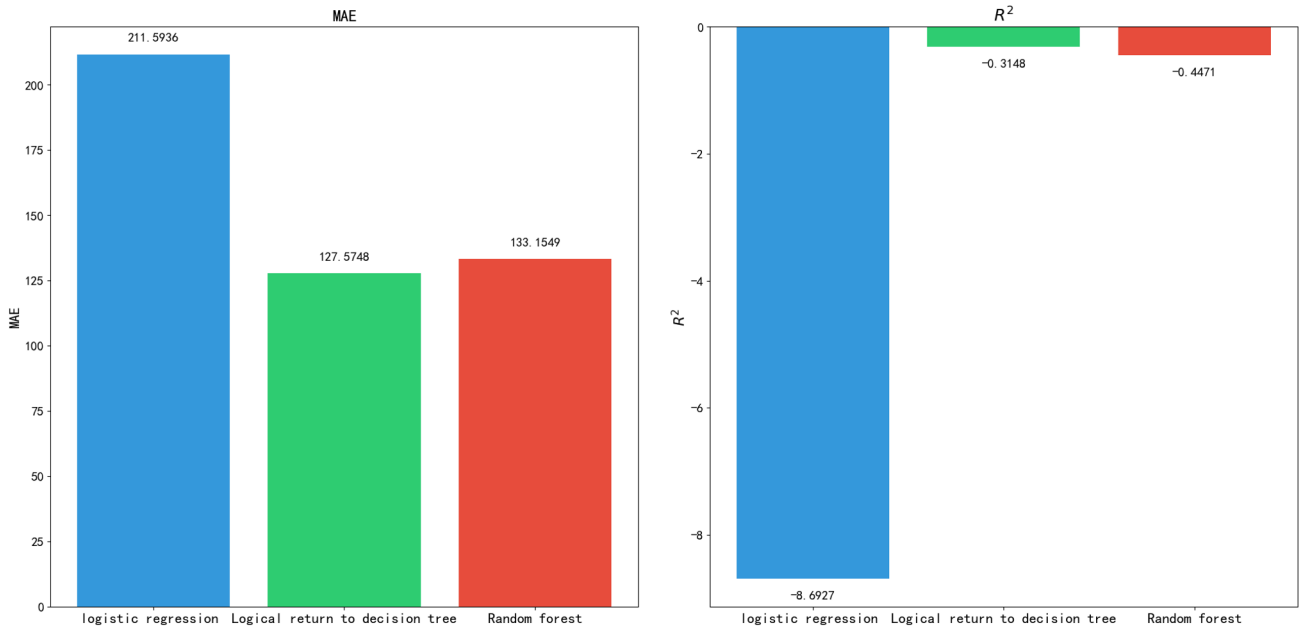

Figure 1: Performance comparison of three models on the test set (Picture credit: Original)

The performance difference between the three models is fundamentally due to the different modeling capabilities. Logistic regression can only grasp the linear relationship, and it is a little difficult to cope with the complex demand change laws. Decision trees can learn the nonlinear characteristics interactively, but a single tree is too true to the noise in the training data, and it is easy to fail to keep up with the performance in the test set. The random forest relies on training multiple decision trees at one go, and then adding the sample disturbance and characteristic disturbance to reduce the risk of over fitting. The prediction accuracy and stability are better than the other two models.

In addition, it is also noted that adding the delay feature and sliding window statistical feature to the input scheme can help these three models. The most beneficial one is the random forest. This situation can be seen. Clearly making the dependency of the sequence in the

time series into a structured feature is of great help to those tree models that do not have time series modeling capability.

4 Discussion

This study has achieved relatively clear model comparison results on the DataCo dataset, but there are still several limitations. First of all, the data set only covers the end of 2017. It has been a long time since then. The timeliness of the data may affect the applicability of the conclusion in the current market environment. Secondly, this experiment only sets default parameters or basic settings for each model, without systematic hyperparameter tuning. There is room for improvement in the ability of the model. Third, the feature is mainly around the time dimension and historical statistics dimension. External factors such as promotion activities, holiday effects and macroeconomic indicators have not been added yet, but these things often bring considerable fluctuations in the actual supply chain demand.

Future research can be carried out in the following directions. One is to get some updated data sets, and then integrate more external features to test whether the model is stable under different market conditions. The second is to take out the random forest and decision tree, and systematically conduct a round of hyperparameter search, such as the depth of the tree, the number of trees, and the splitting conditions, so as to fully tap the potential performance of the model. Third, the interpretability analysis is also added to the prediction model, for example, to calculate the importance score of each feature, and reveal the key reasons behind the demand fluctuations in a more direct way, so as to bring some better references to the actual decision-making of the supply chain.

5 Conclusion

Based on the dataset of DataCo's global intelligent supply chain, this paper constructs three traditional machine learning models: logistic regression, decision tree and random forest, which are used for demand forecasting. The average absolute error and the coefficient of determination are used for comparison in the evaluation. From the results of the experimental operation, the performance of random forest in MAE and R^2 is better than that of logistic regression and decision tree, which also verifies that the integrated method is indeed effective in the task of supply chain demand forecasting. Logistic regression is limited by the ability of the model itself, and it can not keep up with the unstable demand data. A single decision tree is easy to fall into the trap of overfitting, and the generalization ability is obviously weaker than that of random forest. When several models are put together, the random forest achieves a good balance among the prediction accuracy, generalization ability and the difficulty of getting started. Among the three models, it is the most suitable for the supply chain demand forecasting task. The results of the analysis and operation of this research can provide a reference for enterprises to select prediction tools and deploy them in real use scenarios, which is both efficient and easy to implement.

References

1. Ma, J. and Zhang, W. (2025) Simulation of retail supply chain demand forecasting algorithm based on data field clustering. Computer Simulation, 8.
2. Zhang, X. (2025) Establishment and practice of order forecasting model for packaging industry based on customer demand. Packaging Engineering, S1.
3. Tan, J. and Qiu, Y. (2025) Research on multi-time granularity demand forecasting based on machine learning: Taking the supply chain management of a manufacturing enterprise as an example. Statistics and Applications, 14, 213-224.
4. Liu, X. and Wang, J. (2025) Product demand forecasting of e-commerce platform based on LSTM-random forest combination model. E-commerce Review, 14, 1023-1034.
5. Sattar, M.U., Ahmed, N., Khan, M.A., et al. (2025) Enhancing supply chain management: A comparative study of machine learning techniques with cost-accuracy and ESG-based evaluation for forecasting and risk mitigation. Sustainability, 17, 5772.
6. Javed, S. and Akhtar, R. (2024) Data driven approaches for demand forecasting in supply chain for business decisions. In: 5th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, February 19-20, 2024, pp. 1-7. IEEE.
7. Chai, H., Zheng, J. and He, L. (2023) FMCG demand forecasting model based on self-attention mechanism and SARIMA-LSTM algorithm. Computer Age, 8.
8. Zheng, L., Wu, X., Li, J., Mao, Z. and Chen, W. (2023) Demand forecasting of e-commerce commodities based on TCN. Digital Technology and Applications, 1.
9. Li, R. (2024) Research on cross-border e-commerce inventory forecasting based on XGBoost algorithm. Journal of Taiyuan City Polytechnic, 1.
10. Zhang, D. and Yang, C. (2024) The role of supply chain partner information in the prediction of supply chain financial credit risk: A comparative analysis based on multi-machine learning model. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 44, 56-60.