UDC 004.89:005.8
DOI: 10.15587/2706-5448.2024.315495

Pavlo Teslenko,
Serhii Barskyi

ANALYSIS OF MACHINE LEARNING MODELS FOR FORECASTING RETAIL RESOURCES

The object of research is the process of forecasting loosely structured data of retail artifacts by means of machine learning.

The paper analyzes data and models for forecasting retail resources. The analysis is carried out for a specific business situation and task, when a large corporation needs a fuller loading of its own warehouses with goods and resources that will be used in future periods for sale or in projects. The task is to reduce overall corporate costs by purchasing the necessary goods/resources in advance. The data required for forecasting, their sources and properties are defined. It is shown that the data will come from different sources, will have a different time interval, categorical component and logistic reference. RNN, LSTM, Random Forest, Gradient Boosting, XGBoost models and forecasting methods were chosen for such data. They were analyzed according to the criteria of data source, time interval, categorization of data, availability of a logistic component, flexibility of tools in working with heterogeneous data, requirements of tools for computing resources, interpretability of modeling results.

Data sources explain where the data for analysis comes from. Usually it is: stores, warehouses, logistics companies, projects and strategic plans of the corporation. The time interval characterizes the frequency and regularity of receiving data for analysis. The criterion "data categorization" characterizes how this type of data affects the quality of the analysis. The logistic parameters of the data also characterize the impact on the analysis. "Flexibility in working with heterogeneous data" shows the ability of the model to effectively work with data of different formats and sources. Requirements for computing resources determine their necessary power for training and operation of the model. Interpretability of a model characterizes its ability to explain how and why it makes specific decisions or predictions based on input data. The more complex the model, the more difficult it is to interpret. In the retail business, interpretability is important for explaining demand forecasts.

Based on the results of the analysis, the XGBoost model was recommended as the best for forecasting retail resources.

Keywords: *machine learning models, retail, forecasting, retail resources, categorical data, model interpretability.*

Received: 20.09.2024
Received in revised form: 02.11.2024
Accepted: 15.11.2024
Published: 22.11.2024

How to cite
Teslenko, P., Barskyi, S. (2024). Analysis of machine learning models for forecasting retail resources. Technology Audit and Production Reserves, 6 (2 (80)), 11–15. https://doi.org/10.15587/2706-5448.2024.315495

1. Introduction

The retail system is considered as a set of processes and operations that ensure effective management of sales of goods or services to end consumers. It includes interaction with suppliers, inventory management, logistics and customer service [1].

The use of the term "retail" instead of "retail trade" is due to the expansion of the boundaries of this concept in modern business conditions. Retail covers not only classic retail trade, but also a wider range of processes and technologies that facilitate interaction between businesses and end consumers. Retail, as a traditional term, focuses on the physical sale of goods in stores or markets [2]. Instead, the term "retail" encompasses inventory management processes, marketing campaigns, logistics solutions and sales analytics.

Thus, the term "retail" reflects a complex approach to interaction with the external environment, which goes beyond traditional retail trade. Therefore, the use of the term "retail" is appropriate, as it more adequately describes modern business processes, which includes not only retail, but also other aspects of omnichannel strategies important for business management in the digital economy.

Conducting research with the aim of forecasting retail resources requires determining its components, their interaction and influence on each other. Such components are usually called retail resources or artifacts. They include:

– material resources (goods and inventory);
– human resources (personnel, in particular in the service sector);
– financial resources (capital for purchases and operations);
– information resources (data on demand, supply, storage and sales);
– technological resources (analysis, management and automation systems).

According to the authors, the terms "artifact" and "artifacts of retail" are more suitable for describing the subject

ISSN-L 2664-9969; E-ISSN 2706-5448

of research. Artifact comes from the Latin "arte factum", which means "artificially created object". An artifact is usually any object created or modified by a person for a specific purpose. The meaning of this term varies in different industries. In cultural studies, artifacts are material objects that reflect human culture and activity. In computer science, artifacts are objects created in the process of software development, such as: models, documents, prototypes.

In the context of retail, the term artifact can be used to denote any tangible or intangible objects created or used in retail processes that help realize business goals [3].

Artifacts of retail are tangible and intangible elements that ensure the functioning of retail trade, logistics and storage of goods stocks and shape the consumer experience. They are divided into several main groups: physical artifacts, information artifacts, and digital artifacts.

Physical artefacts include sales areas, exhibition stands, warehouses, means of transport and movement, inventories and shop interiors.

Informational artifacts such as price tags, information boards and promotional materials influence the decision-making process of shoppers by providing key information about products and promotions.

Digital artifacts include online stores, mobile applications, and e-commerce systems. The interaction between these artifacts ensures the integrity of the consumer experience: physical and informational artifacts are complemented by digital solutions, contributing to the improvement of customer service.

Thus, the use of the term "artifacts of retail" is correct.

The relevance of the research is substantiated by the business situation that has developed in a large corporation, which has a significant number of warehouses of various types and purposes, which are distributed territorially [4]. For its own production, project needs and for sale, the corporation needs a significant amount of commodity stocks, resource goods. There is a need to forecast several interdependent indicators:

– future business activity and need for commodity stocks or material resources;
– the number of necessary resources and the time of their use;
– quantity and quality of free warehouse spaces with reference to time;
– assessment of the expediency of early purchase of material resources and the cost of their preservation.

The aim of research is to conduct a comparative analysis of machine learning models that will ensure acceptable accuracy of the forecast of retail indicators. This will make it possible to forecast the amount of merchandise stocks for retail in conditions when data will come from different sources, will have different time intervals, categorical components and logistic reference.

2. Materials and Methods

Machine learning (ML) is a branch of artificial intelligence that deals with the development of algorithms that can automatically improve based on experience [5]. In [6] they emphasize that machine learning includes methods that allow computers to learn from data without explicit programming. The essence of ML is the ability of models and algorithms to learn from historical data and make predictions or classifications for new ones.

The paper considers three main types of machine learning*: supervised, unsupervised and reinforcement learning* [7]. Supervised learning requires the presence of labels or labeled data and aims to train models to predict output values based on input data. Unsupervised learning, in contrast, works with unlabeled data, focusing on discovering hidden structures or clustering. Reinforcement learning consists in optimizing actions through a system of rewards for correct actions or penalties for incorrect ones.

In the work, machine learning models are compared according to the following criteria: accuracy, ability to generalize, resistance to noise in data, learning speed and scalability [8]. Supervised learning usually provides high accuracy under well-labeled data, while unsupervised learning is suitable for discovering new patterns in large unlabeled datasets. Reinforcement learning is adaptive because learning occurs through interaction with the environment.

Depending on the learning method and the type of problem they solve, the main machine learning models include linear regression, decision trees, support vectors (SVM), neural networks, ensemble methods, and clustering.

For forecasting in retail, when data comes from different sources, with different frequency and indexing intervals, models capable of processing heterogeneous, multidimensional and time-dependent data are desirable [9]. For this, recurrent neural networks (RNNs) and long short-term memories (LSTMs) are among the best options. They are well suited for time series analysis, allowing them to take into account historical data and trends from different sources of information.

RNNs and LSTMs are able to account for dependencies between data, even if these data are collected from different intervals. LSTM, in particular, works effectively with large volumes of data that have complex time structures, and is also able to store relevant information over long time intervals, which allows for more accurate forecasting of inventory needs.

Ensemble methods such as Random Forest or Gradient Boosting are also effective for such forecasting tasks. They combine the results of several simple models to improve overall accuracy. These models can handle data from different sources and formats, including categorical and quantitative variables, and deal well with noise in the data.

XGBoost is a variant of gradient boosting and is widely used in inventory forecasting due to its ability to work with large data sets, including heterogeneous and incomplete data. XGBoost uses a gradient descent method to minimize residual errors for each prior model. XGBoost has an optimized algorithm to speed up the learning process and increase accuracy.

The choice of model also depends on the available amount of data and their quality. When there is a large amount of data, neural networks can provide more accurate predictions, while ensemble methods may be better when working with smaller data sets or when computing resources are limited. All these models operate on the basis of available labeled data and try to predict the outcome using the available input values.

Therefore, *the object of research* is the processes of forecasting loosely structured data of retail artifacts by means of machine learning.

3. Results and Discussion

The description and characterization of the data, according to which it is necessary to forecast the amount of resources for the future period, includes several key

parameters. Data sources can be different: stores, warehouses and transport companies. Each source provides unique data, including sales, inventory, resource requirements for future projects, and shipment statuses. The time intervals are important because the data will come in at different frequencies – daily, weekly or monthly, which will make it difficult to reconcile and normalize them.

Quantitative indicators include data on sales, balances and planned projects and deliveries, which allows to assess the demand and need for inventory [10]. Resource categories help refine the analysis by considering different product types and their impact on inventory. Logistic parameters such as delivery times and shipment statuses allow to take into account potential delays and adjust the forecast quantity of goods. Such a comprehensive presentation of data is necessary for accurate forecasting of inventory needs for the future period.

The next step was to analyze the structure and properties of the data for the models and methods selected in the previous part of the article for forecasting retail resources. Each of the forecasting tools has different data requirements for the model to work in the best way [11].

To test the effectiveness of machine learning models, it is planned to use a real data set from a large retail network. It will include daily sales figures for the last two years, stock levels in warehouses and logistics information such as delivery times and delivery statuses.

The following signs were used to forecast the amount of required inventory in retail:

– X_1, X_2, ..., X_n – input characteristics (sales volumes, logistics data, stocks);

– Y – the target variable (the amount of necessary stocks for the next period).

Recurrent neural networks work well with data that comes from consistent sources, such as stores or warehouses, at regular time intervals. For effective work, it is important that these data are presented in time series. For RNN, it would be ideal if the data has a fixed time interval, daily or weekly sales. In a real-world situation, this can be problematic, as data from different sources may have different intervals, such as sales volume or product balances. RNNs work efficiently with quantitative data, while categorical data can be difficult to process and analyze. Such data must be converted into numerical formats. Logistics data can be processed if they are presented in time series. The RNN formula describes the relationship between the current state of the network and the previous state, taking into account the input data:

$$h_t = \sigma\left(W_h \cdot h_{t-1} + W_x \cdot X_t + b_h\right), \quad (1)$$

where h_t – current state; h_{t-1} – previous state; X_t – input data at time t; W_h and W_x – weight matrices; b_n – displacement vector; σ – non-linear activation function (for example, sigmoid).

LSTM models also, like RNNs, work well with data from sequential sources, but do better with long time series. These models are suitable for data coming from stores, warehouses and transport companies. This data can be at irregular time intervals, which is an important advantage, since data from different sources in reality may not coincide in frequency. LSTMs work well with quantitative data, while categorical ones, like RNNs, need to be converted to numeric formats. Logistic data can also be processed, but must be well structured to train the model. LSTM has a more complex structure that includes "memory cells". The basic formula for updating the state of a cell looks like this:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \quad (2)$$

where C_t – new state of the cell; f_t – vector of forgetting (forget gate); i_t – input vector (input gate); $\tilde{C}_t$ – new candidacy for the state of the cell.

Random Forest works well with many types of data, including mixed sources with irregular time intervals. It handles both quantitative and categorical data well, without the need for complex transformations, even with logistical parameters such as delivery times or shipment statuses. Random Forest is an ensemble model that consists of a set of decision trees. The prediction for random forests can be described as the average value of the predictions of individual trees:

$$\hat{Y} = \frac{1}{M}\sum_{m=1}^{M} T_m\left(X\right), \quad (3)$$

where $T_m(X)$ – forecast of the m-th tree for the input data X, and M – the number of trees in the model.

Gradient Boosting works with different data sources and with mixed data types. It is less sensitive to the regularity of time intervals than Random Forest. Categorical data can be processed using transformations. Logistics parameters can be integrated for processing. In gradient boosting, the forecast is added at each step as a correction to the previous forecast:

$$\hat{Y}_t = \hat{Y}_{t-1} + \eta \cdot g_t\left(X\right), \quad (4)$$

where η – learning rate; $g_t(X)$ – gradient of the loss function at step t.

XGBoost, as a variant of Gradient Boosting, handles data from different sources well, works with data that arrives at irregular time intervals, which is an advantage for real retail data. Works with quantitative metrics and easily integrates categorical data with minimal transformations. Logistic parameters can be integrated, making it versatile for complex data analysis. XGBoost is an optimized version of gradient boosting, and its prediction formula is as follows:

$$\hat{Y} = \sum_{m=1}^{M} \eta \cdot T_m\left(X\right), \quad (5)$$

where η – learning rate; $T_m(X)$ – forecast of the m-th tree; M – the number of trees.

Setting up machine learning models will involve several important steps common to all selected models. First of all, the data set must be divided into training and test samples: 80 % – for training the model, and 20 % – for evaluating its quality on unknown data.

The next step is to pre-process the data. Normalization of quantitative features is performed, as well as conversion of categorical features into numerical formats using the one-hot encoding method.

After that, hyperparameters of each model are selected. For this, cross-validation methods are used, which allow to evaluate the quality of the model on different fragments of the training sample. Hyperparameters such as tree depth,

ISSN-L 2664-9969; E-ISSN 2706-5448

learning rate, and number of neurons in layers will be selected to maximize accuracy and minimize model errors.

After training, the model is evaluated using metrics such as Mean Squared Error (MSE) or R^2 on the test sample. The final stage is the optimization of hyperparameters to improve performance and reduce training time using Grid Search and Random Search methods.

Based on the results of the analysis, Table 1 was created. Since all tools work effectively with quantitative data, this criterion for analysis from Table 1 is excluded to reduce its size. The following were selected as other criteria for comparative analysis:

– data sources;
– time interval;
– categorical data (CD);
– logistic parameters (LP);
– flexibility in working with heterogeneous data (Flexibility);
– requirements for computing resources (RCR);
– interpretability of the model (IM).

After analyzing all the parameters, XGBoost was chosen as the optimal model for forecasting the amount of inventory in retail. Because this model has maximum flexibility when dealing with heterogeneous data, supports irregular intervals, and requires moderate computing resources with high efficiency.

For the practical use of XGBoost to forecast the amount of inventory in retail, it is necessary to collect data from various sources: stores, warehouses, transport companies, combine them into a single dataset, aligning different time intervals. And also make sure that numerical and categorical features are presented correctly.

Divide the data into "training" and "test" samples, choose model parameters: number of trees, learning rate, depth of trees and choose optimal hyperparameters, use cross-validation. Next, train the model on the prepared data, using quantitative and logistic indicators as input variables, and sales and inventory balances to predict future needs.

Use the MSE, MAE, or R^2 metrics to test the accuracy of the model on a test sample and perform a forecast based on new data (for example, predict what inventory will be needed in the next period).

The limitation of the research can be considered the dependence of the forecasting model on the characteristics of the business taken as a basis as an example. For a different configuration of data sources, time intervals, and categorical indicators, the forecast accuracy results must be checked and the model adjusted anew.

The conditions of martial law in Ukraine, namely shelling and destruction of Ukrainian infrastructure by Russia, significantly limit both the retail business activity itself and the data that could be obtained for analysis and forecasting.

The perspective of further research is the formation of datasets for training the model and determining the accuracy of forecasting.

4. Conclusions

The paper analyzes data and models for forecasting retail resources. The analysis is carried out for a specific business situation and task, when a large corporation needs a more complete loading of its own warehouses with goods and resources that will be used in future periods for sales or in projects. The task is to reduce overall corporate costs by purchasing the necessary goods/resources in advance. The data required for forecasting, their sources and properties are defined. It is shown that the data will come from different sources, will have a different time interval, categorical component and logistic reference. RNN, LSTM, Random Forest, Gradient Boosting, XGBoost models and forecasting methods were chosen for such data. They were analyzed according to the criteria of data source, time interval, categorization of data, availability of a logistic component, flexibility of tools in working with heterogeneous data, requirements of tools for computing resources, interpretability of modeling results. Based on the results of the analysis, the XGBoost model was recommended as the best for forecasting retail resources.

In the theoretical plane, the research result will be useful in the field of intelligent management models for forecasting loosely structured data. On a practical level, the research result can become an effective tool for data analysts and project managers when forecasting retail resources.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, including financial, personal, authorship, or any other, that could affect the study and its results presented in this article.

Financing

The study was conducted without financial support.

Table 1

Comparative analysis of tools for forecasting retail resources

Method/Model	Data sources	Time interval	CD	LP	Flexibility	RCR	IM
RNN	F	F	One-hot encoding	Need time series for logistic data	L	H	L
LSTM	F	Support for irregular intervals	One-hot encoding	Works with logistic data in sequences	Flexibility in working with different intervals	VH	L
Random Forest	Various sources	Does not depend on intervals	Handles categorical data easily	Easily integrates logistics parameters	Good flexibility with different sources	M	H
Gradient Boosting	Various sources	Does not depend on intervals	Handles categorical data easily	Easily integrates logistics parameters	High flexibility for complex data	M-H	M
XGBoost	Various sources	Supports irregular intervals	Integrates categories with minimal processing	Effectively processes logistics parameters	Max. flexibility for heterogeneous data	O	M

Notes: F – fixed; L – low; H – high; VH – very high; M – medium; M-H – medium-high; O – optimized

Data availability

Data will be provided upon reasonable request.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the presented work.

References

1. Vanessa Munoz Macas, C., Andres Espinoza Aguirre, J., Arcentales-Carrion, R., Pena, M. (2021). Inventory management for retail companies: A literature review and current trends. *2021 Second International Conference on Information Systems and Software Technologies (ICI2ST)*, 71–78. https://doi.org/10.1109/ici2st51859.2021.00018
2. Tereshchenko, S. I., Hrymailo, O. V. (2023). Innovatsiini formy orhanizatsii rozdribnoi torhivli. *MNPK Tsyfrova transformatsiia ta dydzhytal tekhnolohii dlia staloho rozvytku vsikh haluzei suchasnoi osvity, nauky i praktyky,* 281–285. Available at: https://repo.btu.kharkov.ua/bitstream/123456789/29872/1/Zbi%C3%B3r_prac_3_2023-281-285.pdf
3. Zimmermann, R., Mora, D., Cirqueira, D., Helfert, M., Bezbradica, M., Werth, D. et al. (2022). Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence. *Journal of Research in Interactive Marketing, 17 (2),* 273–298. https://doi.org/10.1108/jrim-09-2021-0237
4. Barskyi, S. Yu., Teslenko, P. O. (2024) Upravlinnia resursamy infrastrukturnykh proiektiv. *Informatsiini systemy v upravlinni proiektamy ta prohramamy.* KhNURE, 53–56.
5. Mitchell, T. M. (1997). Does machine learning really work? *AI magazine, 18 (3),* 11.
6. Bishop, C. M. (2016). *Pattern Recognition and Machine Learning*. New York: Springer, 778.
7. Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press. Available at: http://www.deeplearningbook.org
8. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Data Mining, Inference, and Prediction. https://doi.org/10.1007/978-0-387-84858-7
9. Chai, T.-Y., Haw, S. C., Jahangir, M., Hoe, K. B., Heng, L. E., Vellaisamy, M. (2023). Advancing Retail Operations: A Customizable IoT-Based Smart Inventory System. *International Journal of Membrane Science and Technology, 10 (3),* 1885–1897. https://doi.org/10.15379/ijmst.v10i3.1848
10. Barskyi, S., Kopaniev, A. (2023). Sutnist ta osoblyvosti upravlinnia infrastrukturnymy proiektamy. *Project, program, portfolio management. R3M-2023.* IShIR, 228–230.
11. Grewal, D., Benoit, S., Noble, S. M., Guha, A., Ahlbom, C.-P., Nordfält, J. (2023). Leveraging In-Store Technology and AI: Increasing Customer and Employee Efficiency and Enhancing their Experiences. *Journal of Retailing, 99 (4),* 487–504. https://doi.org/10.1016/j.jretai.2023.10.002

✉***Pavlo Teslenko,*** *PhD, Associate Professor, Department of Artificial Intelligence and Data Analysis, Odesa Polytechnic National University, Odesa, Ukraine, e-mail: teslenko@op.edu.ua, ORCID: https://orcid.org/0000-0001-6564-6185*

Serhii Barskyi, *PhD Student, Department of Artificial Intelligence and Data Analysis, Odesa Polytechnic National University, Odesa, Ukraine, ORCID: https://orcid.org/0009-0002-8012-3846*

✉*Corresponding author*