MACHINE LEARNING APPLICATION FOR SALES FORECASTING IN E-COMMERCE

Thi Ha Do[1], Manh Quang Do[2,*]

DOI: http://doi.org/ 10.57001/huih5804.2025.351

ABSTRACT

In the digital era, data has become a key driver for enabling businesses to optimize operations and make informed decisions. The shift from traditional business models to e-commerce has resulted in the generation of massive data streams. When systematically collected and processed, such data allows enterprises to proactively develop more effective business strategies, reduce experimental costs, and enhance operational efficiency. In this study, we present a dataset on sales performance of food establishments in Ho Chi Minh City and apply three machine learning models - Linear Regression, XGBoost, and Random Forest - to forecast sales or product demand in the food and beverage sector on e-commerce platforms. The experimental results highlight the potential of modern machine learning approaches for demand forecasting, thereby supporting businesses in resource optimization and improving performance on digital platforms. The dataset and source code are publicly available and continuously updated at https://github.com/DoHa8705/sales_forecasting

Keywords: *Machine learning, e-commerce, sales forecasting.*

[1]School of Economics, Hanoi University of Industry, Vietnam

[2]School of Information and Communication Technology, Hanoi University of Industry, Vietnam

[*]Email: quang.do@fit-haui.edu.vn

Received: 07/7/2025

Revised: 15/9/2025

Accepted: 28/9/2025

1. INTRODUCTION

Recent studies have demonstrated the effectiveness of machine learning models such as Random Forest, Gradient Boosting (including XGBoost), and Neural Networks in forecasting product demand and sales on online platforms [1-3]. Specifically, XGBoost has been successfully applied in retail forecasting competitions such as Kaggle Rossmann Store Sales, achieving high accuracy and effective handling of missing data [4]. However, applying machine learning in e-commerce is still a developing topic, requiring empirical studies to verify its effectiveness in specific contexts - such as different product categories, business scales, or consumer behavior in each market. For this reason, this study focuses on applying popular machine learning models - Linear Regression, Random Forest, and XGBoost - to forecast product purchase counts in the food and beverage sector, which is highly volatile and heavily dependent on short-term consumption trends and market feedback. The research aims to contribute empirical evidence showing the potential application of machine learning in improving management efficiency and decision-making in e-commerce.

2. RELATED WORK

In recent years, sales forecasting in e-commerce has become a topic of considerable interest in the research community, particularly in the context of big data and the growing prevalence of machine learning. Studies have mainly focused on three directions:

1) Comparing modern machine learning models on e-commerce datasets.

2) Applying models in real-world business problems.

3) Optimizing models by integrating semantic data, events, or social media data.

Some studies focus on comparing the performance of modern machine learning models in retail forecasting. Bandara et al. provided a comprehensive review of LSTM models in time series forecasting, highlighting the advantage of deep networks in learning complex patterns in highly seasonal data [3]. More recently, Walter et al. conducted an empirical analysis comparing XGBoost, LightGBM, Temporal Fusion Transformer (TFT), and N-BEATS on e-commerce data, showing that tree-based models outperform others in both accuracy and training speed in noisy or incomplete datasets [5].

Another research direction focuses on optimizing forecasting models via hyperparameter search. Salim et al. found that Random Forest with hyperparameter tuning using Randomized Search outperformed Gradient Boosting, SVR, and XGBoost in retail sales forecasting, achieving R^2 up to 0.945 [6].

In Vietnam, several empirical studies have evaluated the effectiveness of machine learning models in specific contexts. Nguyen et al. compared SARIMAX, Prophet, and LSTM in e-commerce sales forecasting, showing SARIMAX had higher accuracy than LSTM in stable time series data [7]. Le and Pham applied LSTM to forecast FMCG product demand, showing model performance was heavily dependent on preprocessing and parameter tuning [8].

Recent studies have also focused on incorporating influencing factors such as customer sentiment or event information. Sutanto et al. analyzed customer feedback using the BERT model, integrating it into forecasting models to improve accuracy; this integration reduced RMSE by 12% to 23% [9]. Another new approach by Yildiz et al. embedded global event data (holidays, disasters, sports, etc.) into Transformer models, significantly improving performance during periods of major consumer behavior anomalies [10].

Overall, these studies indicate that machine learning models, especially XGBoost and those integrating multiple data sources, hold great potential for sales forecasting in e-commerce. However, effectiveness depends heavily on data quality, industry structure, and specific market contexts.

3. DATA

In this study, data was collected from an e-commerce platform using the data collection method described in Quang et al. [11]. The dataset contains over 3,800 observations, each corresponding to a unique dish ordered by users through the platform. The collected samples ensure diversity in product categories, store sizes, and levels of customer interaction.

The dataset includes the following main groups of variables: Store characteristics (name, business type) and geographic location (district, city); Product identification information (product name, product category). Interaction and business performance indicators (number of customer reviews, average rating, number of likes, number of purchases - the dependent variable in the forecasting model); Price information.

Before building the models, the dataset was cleaned to remove duplicate observations and records with missing key information. String variables were standardized, while numeric variables were checked and handled for outliers. Table 1 and Table 2 present descriptive statistics of features collected for each observation. The results show that most explanatory variables have relatively high standard deviations compared to their means, reflecting large diversity in customer interaction behavior and transaction characteristics.

Table 1. Description of features in the dataset

Attribute	Description
Restaurant_id	Unique identifier of the restaurant
Restaurant_name	Name of the restaurant
Dish_name	Name of the dish
Dish_desc	Detailed description of the dish (e.g., ingredients, preparation)
Price	Price of the dish (VND)
Num_purchases	Number of purchases on the platform
Num_likes	Number of likes of the dish
Num_dislikes	Number of dislikes of the dish
Restaurant_district	District where the restaurant is located
Restaurant_city	City where the restaurant is located
Restaurant_address	Full address of the restaurant
Restaurant_rating	Average rating of the restaurant (0-5 scale)
Num_ratings	Number of ratings of the restaurant
Restaurant_type	Type of restaurant (e.g., eatery, restaurant, cafe)
City_id	City identifier code
Latitude	Latitude coordinate of the restaurant location
Longitude	Longitude coordinate of the restaurant location
Avg_price	Average price of dishes at the restaurant (VND)

Table 2. Statistics of features in the dataset

Feature	Min	Max	Mean	Std Dev
Price	1000	1950000	42423.73	40841.45
Num_purchases	0	71023	82.31	1355.87
Num_likes	0	42	0.61	2.11
Num_dislikes	0	4	0.06	0.30
Restaurant_rating	0	5	4.73	0.69

Num_ratings	0	2699	317.39	550.59
Latitude	10.75	10.79	10.77	0.01
Longitude	106.68	106.71	106.69	0.01
Avg_price	16000	70	40702.22	11830.62

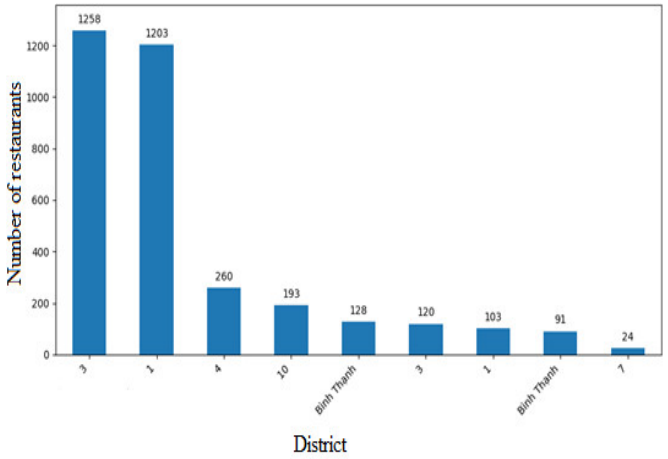

Figure 1. Distribution of the number of restaurants by district in Ho Chi Minh City

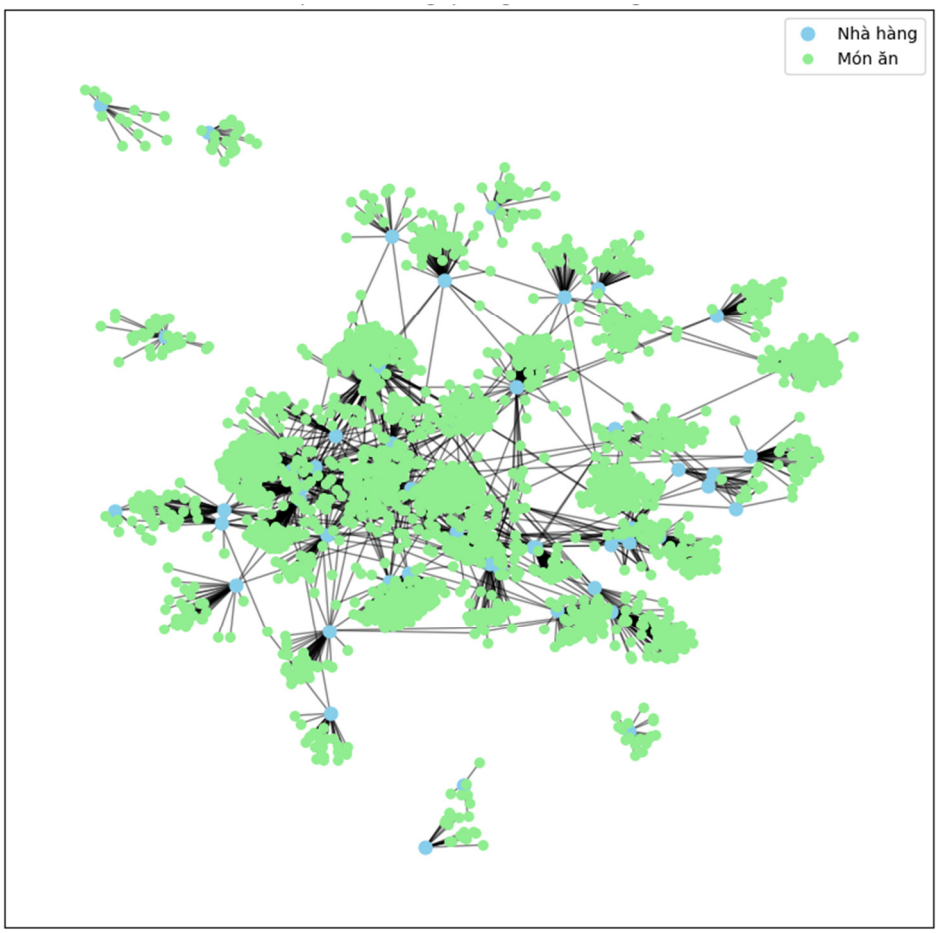

Figure 2. Graph of the relationship between restaurants (Nhà hàng) and dishes (Món ăn). Blue nodes represent restaurants, and green nodes represent dishes

The data reveals a significant disparity in the number of restaurants across districts, as illustrated in Figure 1. Specifically, District 3 and District 1 have the highest concentration of restaurants, with 1,258 and 1,203 establishments respectively, accounting for a large proportion of the city's total. Other districts, such as District 4 (260 restaurants) and District 10 (193 restaurants), have considerably fewer establishments. Some suburban districts or areas with a low concentration of food services have relatively few restaurants, such as District 7 with only 24. Notably, there is inconsistency in the naming of localities (e.g., "Quan 3" vs. "quan 3," "Quan Binh Thanh" vs "Binh Thanh"), which could lead to duplication and statistical errors if not standardized before analysis. This distribution reflects the concentration of food service businesses in central districts, where population density and commercial activity are high, and also suggests that suburban districts still have potential for expanding this type of business.

The correlation matrix among indicators, presented in Figure 3, reveals several notable relationships. The number of likes (num_likes) has a moderate positive correlation with the number of purchases (num_purchases, r = 0.48) and the number of dislikes (num_dislikes, r = 0.43), indicating that items receiving more positive interactions also tend to receive more negative interactions and higher purchase counts. The number of dislikes and the number of purchases have a weak positive correlation (r = 0.32).

The restaurant rating (restaurant_rating) shows almost no correlation with any other variable ($|r| < 0.05$), suggesting that this factor is relatively independent of both interaction and price indicators. The average price (avg_price) has a weak negative correlation with the number of likes (r = -0.09), number of dislikes (r = -0.07), and number of purchases

(r = -0.05), implying that higher prices tend to be associated with lower levels of interaction and purchases, although the effect is minimal.

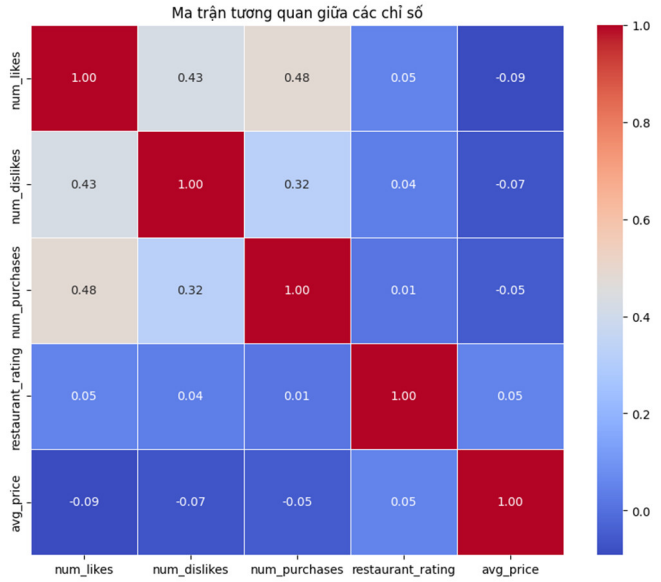

Figure 3. Correlation matrix among indicators such as number of likes, number of dislikes, number of purchases, average price, and restaurant rating

4. METHODOLOGY

The objective of this study is to build and compare the effectiveness of models for forecasting product revenue (number of purchases) on e-commerce platforms. The research implements and evaluates three popular machine learning algorithms: Linear Regression, Random Forest Regression (RFR), and Extreme Gradient Boosting (XGBoost). These models were chosen to leverage both the strengths of linear forecasting methods and modern nonlinear machine learning techniques [4, 12, 13].

A. Regression-Based Machine Learning Models

Linear Regression is one of the most basic and widely used forecasting methods, applied to estimate the linear relationship between a dependent variable (revenue) and independent variables (business features of the store) [12]. The model is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (1)$$

Where y is the target variable (number of purchases) forecasted on the e-commerce platform, X_i are input variables such as likes, dislikes, distance, etc., β_i re the estimated coefficients representing the impact of X_i on revenue, and, ϵ is the random error term.

Linear Regression assumes that the relationship between the dependent and independent variables is linear, meaning a change in an input variable will affect the output variable in a fixed proportion, with all other factors held constant. This method is simple, easy to implement, and easy to interpret, but is limited when the data contains nonlinear relationships or complex interactions between variables [14].

In the context of e-commerce data, which is often diverse, nonlinear, and influenced by many complex factors (e.g., seasonal consumption trends, promotional effects, rapidly changing user behavior), applying Linear Regression alone may lead to limited accuracy.

To overcome this, the study proposes the use of nonlinear machine learning models such as Random Forest Regressor and Gradient Boosting Regressor. These models can better capture nonlinear relationships, interactions between features, and handle noisy or missing data effectively.

Random Forest is an ensemble learning method in which multiple decision trees are constructed and combined to form a stronger forecasting model. The algorithm uses bootstrap sampling with replacement from the original dataset to create multiple training subsets. Each subset is used to train an independent decision tree, increasing diversity and reducing the risk of overfitting to specific samples.

By combining multiple independent decision trees built from bootstrap samples of the original dataset [13], this technique reduces variance, increases stability, and avoids overfitting. In the e-commerce sales forecasting task, each decision tree makes its own prediction based on the input variables, and the final result is the average prediction of all trees [15].

The advantages of Random Forest also lie in its ability to handle outliers and missing data [14], as well as its capacity to evaluate feature importance, which supports business strategy analysis and optimization [16].

Another method used in this study is Gradient Boosting Regressor. Unlike Random Forest - where decision trees are built independently and their results combined - XGBoost is an advanced algorithm based on Gradient Boosting, designed to improve efficiency and training speed on large datasets [4]. This algorithm builds a sequence of weak learners (shallow decision trees), each trained to correct the residual errors of the combined model from the previous step by optimizing a loss function [17].

Specifically, the algorithm starts with a simple initial model (e.g., predicting all data with the mean value of the target variable). In each subsequent iteration, the residual

error between the actual and current predicted values is calculated, and a new shallow decision tree is trained to predict this residual. The output of the new tree is then added to the previous prediction with a scaling factor called the learning rate. This process repeats until the maximum number of iterations is reached or the model converges. XGBoost has been widely applied and shown outstanding performance in many forecasting and classification tasks [4]. In this study, Gradient Boosting Regressor uses likes and dislikes as input variables, which is particularly effective with data having nonlinear relationships and complex feature interactions.

B. Model evaluation metrics

In data science and machine learning research, choosing appropriate evaluation metrics is critical to ensuring the accuracy and objectivity of model comparisons. In quantitative forecasting tasks such as revenue estimation, evaluation metrics must reflect both the model's fit to training data and its predictive performance on unseen data.

Based on the nature of the problem and references from previous studies in time series forecasting and business data analysis, this research uses four widely recognized evaluation metrics: Coefficient of Determination (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics cover multiple aspects of model performance - from the ability to explain data variance to the average deviation, absolute deviation, and relative deviation expressed as a percentage.

The coefficient of determination R^2 reflects the proportion of variance in the target variable that is explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

The Mean Absolute Error (MAE) reflects the average absolute deviation between the forecasts and actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

The Mean Absolute Percentage Error (MAPE) expresses the average relative error as a percentage, allowing for comparison of error levels across datasets of different scales:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

Using all four metrics simultaneously provides a more comprehensive assessment of each model's accuracy and generalization capability, thereby supporting the selection of the optimal solution for the e-commerce sales forecasting problem.

5. RESULTS

In this study, the research team trained and evaluated three forecasting models: Linear Regression, Random Forest Regression, and Gradient Boosting Regression. The dataset used contains two input variables - number of likes (num_likes) and number of dislikes (num_dislikes) - and one output variable - number of purchases (num_purchases). To ensure objectivity, the data was split into two sets: 80% for training and 20% for testing.

For the Linear Regression model, the input data was standardized using the StandardScaler method before training, with weight estimation performed via the least squares method. The Random Forest Regression model was configured with 100 decision trees (n_estimators = 100) and a maximum depth of 10, balancing the ability to model complex relationships with the need to reduce overfitting. Meanwhile, the Gradient Boosting Regression model was set with 200 estimators (n_estimators = 200), a learning rate of 0.1, and a maximum depth of 5 for each base tree, enabling the model to iteratively learn from the residual errors of previous iterations..

Forecasting performance was evaluated using four metrics: coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The results are summarized in Table 3.

The findings show that Gradient Boosting Regression achieved the highest performance, with R^2 = 0.8704, MSE = 238180.0510, and MAE = 49.2491. Its MAPE was 313.49%, indicating that while the model explains most of the variance in the data (high R^2), the predicted values still have large deviations from actual values due to high relative errors in small-value samples. Random Forest Regression ranked second with R^2 = 0.7425, MSE = 473298.1544, MAE = 60.5663, and MAPE = 313.79%. Linear Regression performed worst,

with R^2 = 0.1133, MSE = 1629555.2794, MAE = 178.0954, and MAPE reaching 1728.02%, indicating that this model failed to capture nonlinear relationships and complex interactions between input variables and the target variable.

Feature importance analysis from the two ensemble models showed that the number of likes had a significantly stronger influence than the number of dislikes in predicting the number of purchases. This suggests that positive user feedback (likes) has a greater impact on purchasing behavior than negative feedback (dislikes)

Table 3. Model evaluation results

Model	R^2	MSE	MAE	MAPE
Linear Regression [18]	0.1133	1629555.2794	178.0954	1728.02%
Random Forest Regression [13]	0.7425	473298.1544	60.5663	313.79%
Gradient Boosting Regression [17]	0.8704	238180.0510	49.2491	313.49%

The above results indicate that nonlinear machine learning methods, particularly "boosting" algorithms, provide significantly higher accuracy and generalization capability compared to traditional linear methods. This aligns with the nature of e-commerce data, where the relationship between user interaction behavior (likes, dislikes) and purchasing decisions is often nonlinear and complex.

6. CONCLUSION AND DISCUSSION

Sales forecasting on e-commerce platforms is an important issue that supports sellers and managers in business strategy planning, resource optimization, and improving competitiveness. In the past, traditional statistical methods such as linear regression were widely applied. However, their limitation lies in modeling nonlinear relationships and handling complex data structures. The development of machine learning algorithms has opened up opportunities to improve forecasting accuracy and adaptability to various types of data.

The research results confirm the superiority of advanced machine learning algorithms over traditional methods in the context of e-commerce sales forecasting. This has practical implications for businesses and sellers, as choosing an appropriate model can enhance forecasting accuracy, thereby supporting decisions on inventory, product promotion, and pricing strategy.

In the future, research can be expanded in several directions: (1) Incorporating more features, including product price, delivery time, ratings, and review content to improve forecasting accuracy; (2) Applying automated hyperparameter optimization algorithms such as Bayesian Optimization or Genetic Algorithm to find the optimal model configuration; (3) Deploying trials on real platforms, such as enterprise product management systems or e-commerce platforms, to assess feasibility and effectiveness in operational environments. Additionally, combining powerful models like Gradient Boosting with deep learning approaches could be a potential direction in future research to further enhance forecasting performance and generalization ability.

ACKNOWLEDGMENTS

We would like to thank Samsung Innovation Campus and the School of Information and Communication Technology, Hanoi University of Industry, for organizing the Big Data course, which helped us complete this research. We also express our gratitude to the members of OptiVisionLab for their valuable feedback that contributed to improving this study.

REFERENCES

[1]. S. Makridakis, E. Spiliotis, V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PloS one*, 13, 3, e0194889, 2018.

[2]. G. Zhang, B. E. Patuwo, M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, 14, 1, 35-62, 2017.

[3]. K. Bandara, C. Bergmeir, H. Hewamalage, "LSTM networks for time series forecasting: A survey," *ACM Computing Surveys (CSUR)*, 53, 4, 1-35, 2020.

[4]. T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 785–794, 2016.

[5]. E. Walter, F. Janssen, H. Kuhlmann, "Comparative analysis of modern machine learning models for retail sales forecasting," *arXiv preprint arXiv:2506.05941*, 2025. Available: https://arxiv.org/abs/2506. 05941.

[6]. M. Salim, Z. Fatima, B. Ahmad, "Enhancing retail sales forecasting with optimized machine learning models," *arXiv preprint arXiv:2410.13773*, 2024. Available: https://arxiv.org/abs/2410.13773.

[7]. Pham Thuy, *A comparative analysis of demand forecasting models: A case study of a Vietnam e-commerce company*. Master's thesis, Aalto University, 2024. https://aaltodoc.aalto.fi/items/.

[8]. V. H. Le, T. N. Pham, "Ứng dụng mạng LSTM trong dự báo nhu cầu ngành hàng FMCG tại Việt Nam," *Tạp chí Khoa học và Công nghệ Hồng Bàng*, 12, 4, 2023.

[9]. A. Sutanto, E. Wulandari, F. Nasution, "Improving sales forecasting models by integrating customers' feedbacks: A case study of fashion products," in *Proceedings of ICECH 2023*, 2023.

[10]. S. Yildiz, C. Demir, S. Kara, "Forecasting retail demand using transformer models with external event embeddings," *arXiv preprint arXiv:2405.13995*, 2024. Available: https://arxiv.org/abs/2405.13995.

[11]. M.-Q. Do, T. L. Nguyen, D. D. Vu, et al., "End-to-end system for data crawling, monitoring, and analyzation of e-commerce websites," in *Advances in Information and Communication Technology*, P. T. Nghia, V. D. Thai, N. T. Thuy, V. N. Huynh, and N. Van Huan, Eds., Cham: Springer Nature Switzerland, 1037-1044, 2025. ISBN: 978-3-031-80943-9.

[12]. L. Breiman, "Random forests," *Machine learning*, 45, 1, 5-32, 2001.

[13]. A. Yousefi, et al., "A survey on machine learning techniques for e-commerce demand forecasting," *Journal of Big Data*, 7, 1, 1-23, 2020.

[14]. S. Yichuan, et al., "Improved random forest for classification," *Journal of Applied Mathematics*, 2014.

[15]. A. Maione, et al., "Feature importance measures in random forests," *International Journal of Data Mining*, 2016.

[16]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, 1189- 1232, 2001.

[17]. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009. DOI: 10.1007/978- 0-387-84858-7.