Predicting E-Commerce Sales with Three Machine Learning Models

Xinyan Li

International College Beijing, China Agricultural University, Beijing, 100083, China
2021314050113@cau.edu.cn

Abstract. Online shopping has gained popularity with the advent of e-commerce, owing to its convenience, wide range of choices, and reduced geographical limitations. At the same time, competition among e-commerce enterprises has become increasingly fierce, so enhancing the core competitiveness of e-commerce companies is now contingent upon accurately predicting future sales and devising rational sales strategies. Machine learning (ML) techniques play a pivotal role in this process, as they efficiently handle intricate data and reveal underlying patterns within sales figures, thereby enabling precise projections of upcoming trends. By harnessing the power of ML, e-commerce enterprises can gain a competitive edge and stay ahead of the curve in today's dynamic market. In this paper, sales data on the e-commerce platform of an online retail store registered in the United Kingdom are used to make e-commerce sales predictions employing three distinct ML models: Linear Regression (LR), Decision Tree (DT), and Random Forest (RF). Subsequently, the performance of these models is evaluated by calculating their Mean Absolute Error (MAE), Mean Square Error (MSE), and R-squared values. The selection of the optimal sales prediction model was based on the fitness of the prediction results obtained from each model. By comparing these three regression evaluation metrics, particularly R-squared, the model with the largest R-squared is selected as the one that predicts sales most accurately.

Keywords: Machine Learning Models, Sales Prediction, E-Commerce Sales.

1 Introduction

Along with the rise of the Internet, especially the popularity of mobile Internet based on cell phones and other terminal devices, more and more people are willing to spend their time on the Internet. As a result, e-commerce, which is based on the Internet, has begun to receive focused attention. In recent years, e-commerce has grown very rapidly, at an almost unbelievable rate, and the number of products and services offered to support sales has been enormous. According to an article published by Devi in 2021, the e-commerce industry grew by 36% year-on-year in the last quarter of 2020 [1]. Ecommerce facilitates consumers in engaging in shopping activities without being restricted by time or space, thereby enhancing their overall shopping experience and effectively reducing transaction costs. This advantageous platform offers a convenient and

B. Siuta-Tokarska et al. (eds.), *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, Advances in Economics, Business and Management Research 300,
https://doi.org/10.2991/978-94-6463-542-3_52

efficient way for consumers to purchase goods and services, making it a popular choice in today's digital era.

Compared with conventional trading methods, e-commerce makes the connection between consumers and businesses closer, thus allowing e-commerce companies to quickly learn a variety of market data, such as consumers' purchase lists, browsing records, and payment methods. The frequency of consumer shopping directly correlates with the volume of demand-related information available to e-commerce companies, leading to increasingly precise predictions of consumer purchasing demand. Since 2009, the business sector has increasingly recognized the significance of "big data" and "data mining" technologies, whose practical worth has been demonstrated through numerous examples. One notable instance is Amazon's e-commerce platform, which leverages users' shopping history to provide personalized recommendations [2, 3].

The prediction of goods sales has always been a pivotal concern in the retail sector. Nowadays, as information technology advances, ML is gradually gaining prominence, enabling it to easily access diverse factors that influence the sales of goods. Simultaneously, ML also faces multiple problems. The scale of data has expanded dramatically, but not all data are of high quality and value. Noisy data, missing values, and data imbalance negatively impact the training process and predictive performance of ML models. In addition, ML faces the challenge of model generalization ability. Many complex models perform well on the training set, but once applied to real-world scenarios, their prediction results are often greatly reduced, mainly due to model overfitting or underfitting.

These challenges are particularly prominent in the field of sales forecasting. Sales data are often affected by a variety of factors such as market fluctuations and changes in consumer behavior, and their complexity and uncertainty make the construction of ML models extremely difficult. Therefore, it is becoming increasingly essential to explore effective ML methods for sales forecasting. This helps to improve the efficiency of the store's stocking to minimize the loss of goods and inventory occupation, while effectively meeting market demand. It also can promote the development of ML technology to better adapt to complex and changing real-world scenarios.

In this paper, different ML approaches are used for e-commerce sales forecasting based on sales data on the e-commerce platform. The study uses three ML models: LR, DT, and RF. The validity of the models is verified through MAE, MSE, and R-squared, the three statistical metrics used for regression evaluation. Through the obtained data, the fit of the three models is compared, and the size of the R-squared is analyzed to find out which is the best model to predict e-commerce sales.

2 Literature Review

Sales prediction necessitates the utilization of scientific forecasting methodologies and rational models. Scholars have diligently worked in this domain, conducting profound theoretical investigations and presenting diverse techniques for practical data analysis. For instance, in 2014, Wei et al. employed a Structural Time Series (STS) model to disentangle the trend and seasonal influences on sales volume [4]. They calculated the

residual series, constructed a search index, and ultimately developed a forecasting model grounded on search data and sales residuals. Additionally, Jiménez F et al. introduced a feature selection approach based on a multi-objective optimization algorithm in 2017, leveraging this framework to establish an online sales regression forecasting model [5]. Steinker et al. applied a Time Series (TS) method to forecast e-commerce retail sales based on European-related fashions, integrating weather into sales prediction [6]. It was found that weather, especially extreme weather, has a significant degree of influence on daily sales. Zhao et al. utilized a Convolutional Neural Network (CNN) model to automatically extract meaningful features from structured data [7]. These extracted features were subsequently employed for sales predictions, and the method's effectiveness was validated through testing on a practical dataset. Bandara et al. achieved competitive outcomes in 2019 by incorporating cross-series information into a globally trained Long Short-Term Memory Network (LSTM) model, thus creating a unified forecasting framework [8]. Furthermore, Ji et al. proposed a CA-XGBoost forecasting model, which took into account the sales characteristics of goods and the trends within the data series [9]. This model leveraged the ARIMA model for the linear component and the XGBoost model for the nonlinear component. More recently, in 2022, Li et al. introduced a sales prediction model for mobile e-commerce that integrated Bayesian optimization with TS segmentation to optimize the RF model [10]. In summary, e-commerce sales prediction involves many ML models, and the prediction methods are constantly improving by researchers.

3 Methodology

In this section, this article delves into the provenance of the dataset, examining its constituent parts and conducting a thorough analysis of the data. This paper outlines the fundamental principles guiding the implementation of three distinct ML models.

3.1 Data Collection

The dataset was obtained from Kaggle and contains all transactions that occurred on the platform of a registered non-physical online retail store based in the United Kingdom from December 1, 2010, to December 9, 2011, along with the specific details of the transactions.

This store sells gifts for various occasions. Many of its consumers are wholesalers. The entire dataset has 8 columns and 541,910 rows. This dataset comprises comprehensive information regarding transactions, encompassing details such as invoice numbers, stock codes, descriptions of the goods involved, quantities, dates, unit prices, customer identification numbers, and the respective countries.

3.2 Exploratory Data Analysis (EDA)

Every study of a dataset begins with EDA. The data types in the data framework were processed to prevent additional difficulties before making changes. Also, all date data

were transformed into “datetime” format. Then, the research dealt with missing values. In this dataset, the column called “CustomerID” has too many null values, so this column of data was removed to prevent it from affecting the modeling. In some rows, the numbers of price per unit are 0, so their items were categorized as “UNKNOWN ITEM” and reprocessed in subsequent analysis.

In the column “Description”, items’ names in all capitals are products sold successfully, while items’ names not in all capital letters are products canceled or returned. Based on the data in the “Description” column, the research created four bar charts. Fig. 1 lists the 15 most frequently purchased products. It can be seen that the best-selling product at online retail, which means the product that appears most frequently in orders, is a white hanging heart t-light holder. Fig. 2 enumerates the top 15 products that have experienced the highest frequency of canceled or returned orders. Using the same methodology, the research listed the 15 most frequently used stock codes and the 15 invoices with the most items. For the simplicity of the analysis, the research ignored negative numbers in the dataset. After cleaning the data and removing all suspicious records, the research newly added and calculated sales.

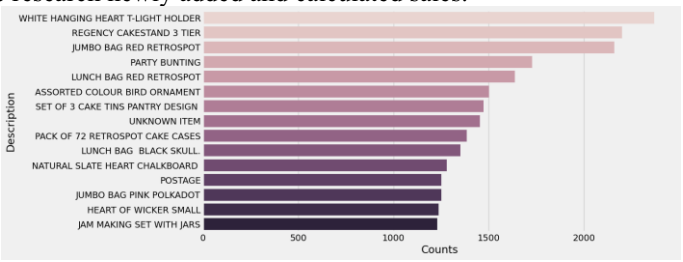


Fig. 1. The 15 most frequently purchased items (Picture credit: Original).

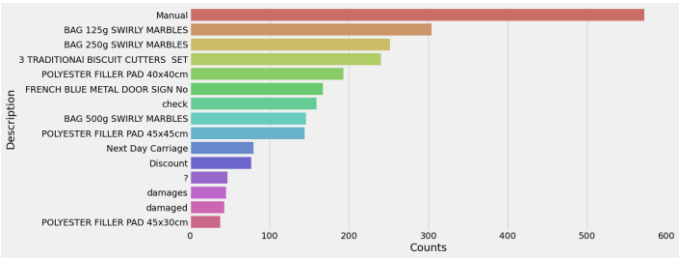
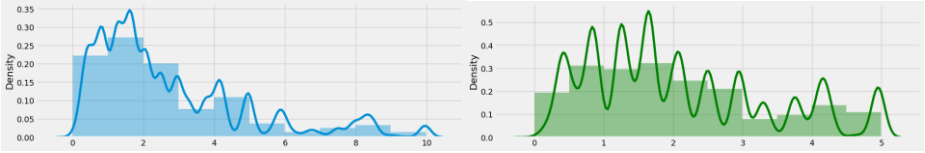


Fig. 2. The 15 most frequently canceled or returned items (Picture credit: Original).

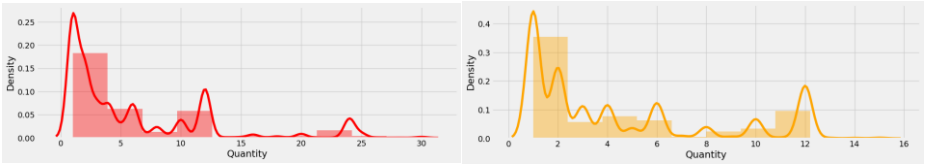
The research calculated that 91.51% of sales in this store were made in the UK and only 8.49% of sales went abroad. Additionally, the study used the scatter plot to plot the data and remove data points that were clearly out of range visually. From two histograms in Fig. 3, it can be concluded that the vast majority of items sold in this store are priced between £0 and £3. Also, from Fig. 4, it can be seen that people typically buy 1 to 5 items or 10 to 12 items at one time. From Fig. 5, it is clearly that most of the sales per order are between 1 to 15 pounds. Finally, the research resampled the temporal

data to analyze the change in sales over time in terms of weeks, using time on the horizontal and sales on the vertical (Fig. 6).



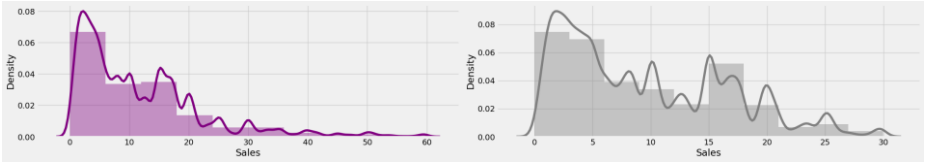
(a) (b)

Fig. 3. Quantities of goods at the same prices. (a) UnitPrice < 10; (b) UnitPrice < 5 (Picture credit: Original).



(c) (d)

Fig. 4. Quantities of customer purchases at one time. (c) Quantity <= 30; (d) Quantity <= 15 (Picture credit: Original).



(e) (f)

Fig. 5. Sales per order. (e) Sales <60; (f) Sales <30 (Picture credit: Original).

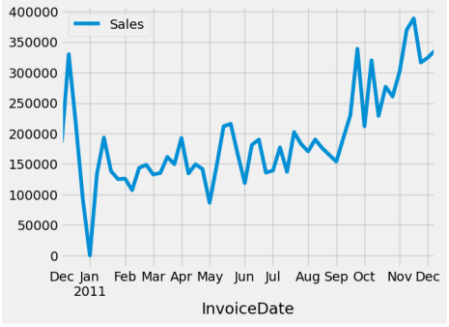


Fig. 6. Sales over time (Picture credit: Original).

3.3 Data Preparation for Modeling

The research found out the quantity per invoice. Based on the previously completed EDA, the cut method was used for binning, dividing these characteristics into 6 buckets

for “Quantity” and 5 buckets for “UnitePrice”. It can be concluded from above Figure 6 that sales vary from season to season, with a peak in the fourth quarter, then a sharp drop in the first quarter of the next year, and continuing to grow until a new peak again in the new fourth quarter. Therefore, the new feature, month, was created to improve the model.

3.4 Building Models

The UK and non-UK data were separated to provide more standardized data for the models. This is because some models may apply to other countries and not to the UK and vice versa. Then the features of the UK data were extracted and dummy variables were created. Since most of the features had values in the range 0 to 1, the “QuantityInv” feature was scaled. Finally, the dataset was partitioned into training and testing subsets to be able to train different models and validate their functionality.

Linear Regression (LR).

In the paper published by Schneider et al. in 2010, regression involves the identification of a line that best approximates the central tendency of all points depicted in the image [11]. When faced with a bunch of scattered points and no specific correlation can be seen, a straight line is a straightforward representation of the trend. Therefore, the principle of the LR model is to find a straight line that is as far as possible in the middle of all the points, representing the overall trend of a dataset, so that the overall relationship of the data can be more clearly visible, and it is convenient for the user to predict the future situation.

There are two types of LR: univariate LR and multiple LR. In this paper, the research uses multivariate LR. The equation for this type of LR is shown below:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n \tag{1}$$

Here, Y represents the dependent and target variable, while bi denotes a parameter of the model, which means it is the regression coefficient of the variable Xi, and a is the intercept. In the model, X1~Xi are actually different features on the sample data.

Decision Tree (DT).

According to Song et al. in 2015, a DT model is a predictive tool that establishes a mapping relationship between object attributes and their corresponding values [12]. Within this model, each root node, alternatively referred to as a decision node, signifies a specific classification. Each forked path, in turn, represents a division based on an attribute, segmenting the samples into subsets according to their distinct values for that particular attribute. Consequently, each leaf node, also known as an end node, corresponds to the object value determined by the path traversed from the root node to that specific leaf node. The crucial challenge in constructing a DT lies in selecting the appropriate attribute at each step to effectively divide the samples. The process of learning and building a DT model from training samples with known class labels follows a top-down, divide-and-conquer approach.

Random Forest (RF).

Belgiu et al. have shown in their research in 2016 that RF is an integrated reconstruction of multiple Categorical Regression Trees (CART) to produce a superior result. RF votes or takes the mean of all DT outputs, and sales forecasting, being a regression problem, can be accurately approached by calculating the mean value [13]. The RF model excels at capturing both linear and nonlinear relationships among variables. Its notable strengths include high prediction accuracy, swift computational speed, and a reduced likelihood of encountering the overfitting phenomenon.

The prediction outcomes of the RF model are contingent upon both the choice of the training sample dataset and crucial parameter configurations within the model. Specifically, "n_estimators" pertains to the quantity of DT generated within the RF. As the value of this parameter increases, so does the number of DT within the model, leading to improved performance of the trained model. However, this augmentation also necessitates increased computational resources, potentially slowing down the processing speed of the code and heightening the risk of overfitting. Therefore, it is imperative to establish appropriate parameter values for the training of the RF prediction model.

4 Results

This article uses GridSearch and CrossValidation to test three types of models. The tuning parameters and best scores were calculated for all three models. Then the research also calculated the three metrics MAE, MSE, and R-squared for different models. The formulas for the three regression evaluation metrics are shown below [14]:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |(X_i - Y_i)| \tag{2}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m} (X_i - Y_i)^2 \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2} \tag{4}$$

In this context, m represents the total number of samples, Xi denotes the forecasted value, Yi signifies the actual value, and $\bar{Y}$ corresponds to the mean value of all actual values.

Table 1 summarizes the metrics for each of the three models. Since R-squared represents the fitting effect of the model, as the R-squared value increases, the model's fit improves accordingly. The research compared the R-squared values of three distinct models and graphically represented them in Fig. 7. The results indicate that the RF model exhibits the most optimal fitting performance, with an R-squared score approaching 0.6, which is considered satisfactory.

Table 1. Metrics for three models.

Metrics \ Models	LR	DT	RF
MAE	15.0914	6.7557	6.7265
MSE	3919.1064	2080.5751	1923.2869
R^2	0.1563	0.5521	0.5859

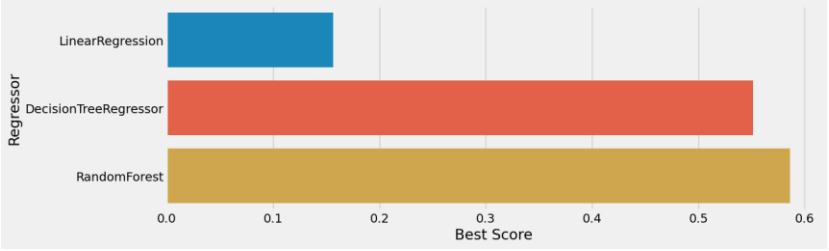


Fig. 7. R2 scores comparison among three models (Picture credit: Original).

5 Discussion

LR model, DT model, and RF model have significant differences in performance, which are mainly due to their different working principles and application scenarios. In the following, the advantages and disadvantages of these three models will be compared in detail, and their improvement directions and practical application values will also be discussed.

First of all, the LR model stands out for its simplicity and computational efficiency. It postulates a linear correlation between the independent and dependent variables, forecasting the target value through the fitting of a straight line. This aspect renders the LR model highly interpretable, straightforward to comprehend, and effortlessly applicable. Nevertheless, the limitations of the LR model are equally evident. It struggles to accommodate nonlinear data, thereby restricting its predictive capabilities in scenarios where the data exhibits nonlinear relationships.

DT model, on the other hand, serves as a classification and regression technique grounded in a tree-like structure. It effectively captures the relationship between inputs and outputs by partitioning the data into distinct regions, demonstrating a superior fitting capability for nonlinear challenges. However, the DT model is susceptible to overfitting, particularly when confronted with high-dimensional data, potentially compromising its performance. Besides, the DT model exhibits relatively less resilience to noise and outliers.

The RF model is a composite approach encompassing numerous DT. It arrives at the ultimate prediction outcomes through a voting mechanism, thus integrating the insights from various trees to yield a comprehensive prediction. The RF model has the advantages of being less prone to overfitting, robust to noise and outliers, and capable of dealing with high-dimensional datasets. The RF model is trained by randomly selecting features and samples, which effectively improves the model's generalization ability.

But the computational complexity of RF models is high and the training time may be long, especially when dealing with large-scale datasets.

To address the limitations of these models, some improvement directions can be proposed. For the LR model, its ability to fit nonlinear data can be enhanced by introducing nonlinear terms or using polynomial regression. To prevent overfitting in the DT model, the pruning technique can be employed. Moreover, incorporating integrated learning methods can enhance the model's stability. For the RF model, further improvements in performance and efficiency can be achieved by optimizing feature selection and refining sample sampling strategies.

In practice applications, these models exhibit a diverse range of utilities and are widely employed in various scenarios. LR model is suitable for problems with linear relationships between features, such as house price prediction and stock price analysis. DT model is used in both classification and regression problems, especially in scenarios that require visual presentation and easy understanding. RF model, on the other hand, has advantages in handling complex datasets, enhancing the stability and generalization capabilities of the model renders it apt for diverse classification and regression tasks.

6 Conclusion

In summary, the RF model demonstrates superior performance in predicting e-commerce sales compared to the other two models, achieving an R-squared score of approximately 0.6, the highest among the three. Nevertheless, due to constraints in professional knowledge and various other factors, the research still exhibits several shortcomings. Firstly, exploring additional methods for feature construction and selection would be beneficial, enabling to determine the optimal feature set through comparative analysis. Secondly, the current study only incorporates three ML models for prediction, which is insufficient for comprehensive big data analysis. Given that the R-squared scores of all three models remain well below 1, further training of additional ML models is necessary to enhance prediction accuracy. And exploring the integration of deep learning models with multiple ML models could potentially further improve prediction precision. What's more, the hyperparameters of the model can be adjusted to give it more training time to improve the model fitting effect. All in all, LR, DT, and RF models have their own characteristics, which need to be selected according to specific problems and dataset characteristics in practical applications. By continuously optimizing and improving these models, the challenges facing ML can be better addressed and more accurate and efficient solutions can be provided for practical applications.

References

1. Devi, S.: Growth and emerging trends in e-commerce: an overview. Peer Reviewed and Refereed Journal, 10(11(2)), (2021).
2. Liu, X. F., Zhan, Z. H., Deng, J. D., Li, Y., Gu, T., Zhang, J.: An energy efficient ant colony system for virtual machine placement in cloud computing. IEEE Transactions on Evolutionary Computation, 22(1), 113–128 (2018).

3. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (2003).
4. Wei, D., Geng, P., Ying, L., Shuaipeng, L.: A prediction study on e-commerce sales based on structure time series model and web search data. In: The 26th Chinese Control and Decision Conference (2014 CCDC), pp. 5346–5351. IEEE, Changsha (2014).
5. Jiménez, F., Sánchez, G., García, J. M., Sciavicco, G., Miralles, L.: Multi-objective evolutionary feature selection for online sales forecasting. Neurocomputing 234, 75–92 (2017).
6. Steinker, S., Hoberg, K., Thonemann, U. W.: The value of weather information for e-commerce operations. Production and Operations Management 26(10), 1854–1874 (2017).
7. Zhao, K., Wang, C.: Sales forecast in e-commerce using convolutional neural network. arXiv: 1708.07946 (2017).
8. Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., Seaman, B: Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In: Gedeon, T., Wong, K., Lee, M. (eds) Neural Information Processing, ICONIP 2019, vol. 11955, pp. 462–474. Springer, Cham (2019).
9. Ji, S., Wang, X., Zhao, W., Guo, D.: An application of a three-stage XGBoost-based model to sales forecasting of a cross-border e-commerce enterprise. Mathematical Problems in Engineering 2019, (2019).
10. Li, S.: Sales forecasting model of e-commerce activities based on improved random forest algorithm. In: 2022 2nd International Conference on Computer Graphics, Image and Virtualization (ICCGIV), pp. 195–198. IEEE, Chongqing (2022).
11. Schneider, A., Hommel, G., Blettner, M.: Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International 107(44), 776–782 (2010).
12. Song, Y. Y., Ying, L. U.: Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry 27(2), 130–135 (2015).
13. Belgiu, M., Drăguţ, L.: Random forest in remote sensing: a review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114, 24–31 (2016).
14. Chicco, D., Warrens, M. J., Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. PeerJ Computer Science 7, e623 (2021).