SETSCI Conference Proceedings
19, 17-21, 2024

International Conference on Advances in Electrical-Electronics Engineering and Computer Science
9-10 November 2024, Ankara, Türkiye

https://doi.org/10.36287/setsci.19.4.022

Development of Machine Learning-Based Sales Cancellation/Return Forecasting Models for the E-Commerce Industry

Zehra Sude Sarı[1], Batuhan Taşkapı[2] Gökay Dağdaş[3], Mehmet Fatih Akay[4]

[1]Dept. of Data Science, Inveon, İstanbul, Turkey
[2]Dept. of Customer Operations, Inveon, İstanbul, Turkey
[3]Dept. of Customer Operations, Inveon, İstanbul, Turkey
[4]Dept. of Computer Engineering, Çukurova University, Adana, Turkey
(sude.sari@inveon.com)

Abstract – E-commerce is evolving rapidly, creating a more competitive market environment. With this development, gaining a competitive advantage has become even more crucial. Companies implement various strategic moves to maintain their position in this competitive market. Strategies and forecasts have a high ranking among these strategies. Predicting sales cancellations and returns is crucial for companies to anticipate future challenges and stay ahead. The aim of this study is to develop sales cancellation/return prediction models for the e-commerce sector. To achieve this, sales cancellation/return prediction models have been generated using Multi-Layer Perceptron (MLP), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Logistic Regression (LR). A weekly dataset has been created using 4909 rows of sales cancellation/return data. The performance of the models has been evaluated using precision, recall, F1 score, and accuracy. Among all the methods, it has been observed that RF and XGBoost delivered the best performance.

Keywords – Machine Learning, Customer Behavior Analysis, Sales Cancellation/Return Prediction, E-commerce, E-commerce Analytics

I. Introduction

E-commerce sector has undergone a significant transformation with the widespread adoption of the Internet and the rapid development of digital technologies, resulting in substantial growth within the industry. Online shopping has become increasingly popular among consumers, replacing traditional retail shopping methods. In 2023, the volume of e-commerce in Turkey increased by 115.15% compared to the previous year, reaching 1.85 trillion Turkish Lira (approximately $77.89 billion). During the same period, the number of transactions increased by 22.25%, totaling 5.87 billion. The share of e-commerce within the total trade volume also rose significantly, from 10.1% in 2019 to 20.3% in 2023. Furthermore, the sector's growth potential remains strong, with the Ministry of Trade projecting that the e-commerce volume will reach 3.4 trillion Turkish Lira and the number of transactions will rise to 6.67 billion by 2024 [1].

In order to remain competitive in this dynamic sector, e-commerce companies need both short-term and long-term financial planning. Financial planning is a process that enables businesses to effectively manage their future financial resources based on an analysis of income and expenditure. However, sales cancellations and return processes pose significant challenges for companies. When sales are canceled or returned, financial plans based on expected revenues from sold products can result in budget deficits when sales are canceled or products are returned.

Sales cancellations and returns cause financial problems, increase costs, negatively impact customer satisfaction, and complicate inventory management. While sales cancellations generally do not incur shipping costs because the product has not yet been dispatched, returns incur shipping costs after the product has been delivered to the customer, which reduces profit margins. A study by Fast Company revealed that processing a return can cost as much as 59% of the original sales price [2]. Furthermore, tracking, storing, and restocking returned products necessitates additional resources and personnel. Ineffective management of cancellations and returns can lead to serious financial and operational consequences for businesses.

Particularly for products with expiration dates, such as perishable goods and seasonal fashion items, returns create significant challenges for reselling. These products require additional processes, making the return process even more complex. For instance, defective or damaged goods may need to be repaired or repackaged, adding labor and costs. In this context, anticipating cancellations and returns can provide significant advantages for companies. If a product has a high likelihood of being returned, the business can strategically delay its shipment or manage the return process more efficiently. These strategies help minimize unnecessary shipping costs and enable businesses to achieve financial savings.

This study is organized as follows: Section 2 covers relevant literature. The dataset, overview of methods and sales return/cancelation prediction models have been described in Section 3. Section 4 and 5 presents the results and discussion. Section 6 concludes the paper.

II. LITERATURE REVIEW

Studies on predicting product returns and order cancellations in e-commerce are extensive, underscoring the growing need for accurate forecasting methods in this sector. Various methodologies have been applied to address these challenges, including probabilistic models, multimodal frameworks, machine learning techniques, and ensemble learning approaches. [3] analyzed sellers' optimal pricing and inventory policies in the context of order cancellations under cash-on-delivery (COD) systems. The study examined customer behaviors related to order cancellations when using the same payment method. Machine learning models, which offered new insights beyond earlier research that relied on periodic review systems and lacked experimental data analysis, were employed. The model's accuracy was assessed, with the Logistic Regression (LR) model having the highest prediction accuracy of 84% and a precision rate of 69%. In another study, [4] proposed an interpretable feature method to enrich the available Passenger Name Record (PNR) information. The prediction performance of two classes of models was empirically evaluated to determine whether they could cross-fertilize each other to improve cancellation prediction. The researchers combined Bayesian Networks (BN) and Lasso Regression (LR). They added to the body of research by suggesting an understandable feature interaction and a forecasting method that used both types of efficient models together. Further, [5] observed that the average return rate of fashion products purchased online ranged from 13% to 96%, based on a large dataset. Machine-based prediction was employed to automatically extract interpretable features from images, and the analysis demonstrated how to select and design fashion products that were less likely to be returned. Similarly, [6] aimed to develop personalized strategies to boost sales, employing current, frequency, and monetary (CFM) models based on digital transformation techniques to better understand and segment potential customers. The underlying reasons behind vendor behaviors were determined using K-means and hierarchical clustering. A real-world dataset was analyzed by [7] to predict accommodation order cancellations, employing XGBoost with optimized hyperparameters to achieve successful outcomes. Meanwhile, [8] proposed a multi-channel pricing and order optimization model with two return policies—full return and non-return. To overcome nonlinearity in the model, a linearization technique was used. The study used Support Vector Machine (SVM) to turn historical data into uncertainty clusters. This turned the model into an approximate mixed integer linear programming model that can be solved with commercial software. A comparison with the box uncertainty set indicated that the data-driven uncertainty set was less conservative and yielded higher profits for the retailer. In the context of online ride-hailing, [9] presented a deep learning model designed to predict the probability of order cancellations. The Didi Chengdu Express public dataset was used to test the Deep Residual Network-Based Ride-Hailing Order Cancellation Probability Prediction Model (DeepOCP). The results showed that the model was good at predicting how likely it was that a user would cancel an online ride-hailing order. A study by [10] suggested using a large dataset to show how important it is for probabilistic forecasters to be well-calibrated in order to make accurate predictions with high precision and good recall. The study recommended using calibrated models selectively, emphasizing the need to avoid predictions for some instances when confidence was low. The increasing trend of real-time order cancellations in live streaming e-commerce (LSE) was analyzed by [11], focusing on the factors that influenced these cancellations. Data were collected from 768 TikTok live streaming rooms in China, including 4,984 product promotion videos, 1.29 million comments, and information from 513,551 viewers. Using emotional contagion theory, the study examined the impact of emotional expressions by live-stream hosts on viewers and how these influenced real-time order cancellations. Finally, [12] proposed a multimodal analysis framework for predicting product returns in e-commerce environments. The model considered multimodal features and their correlations from different data sources, with experiments conducted using a dataset from Taobao live streaming. The findings highlighted the utility of multimodal signals in predicting product returns, allowing live streaming platforms and consumers to anticipate which products would likely have high return rates. Additionally, the study suggested that sellers and anchors could use these predictions to improve product descriptions and interaction strategies.

III. Materials and Method

A. DATASET

A dataset has been created with the sales cancellation/return data received from one of Inveon's customers. The dataset contains 4909 cancellation/return data across 10000 rows. Table 1 shows the attributes in the dataset with their descriptions.

Table 1. Attributes in the dataset

Attributes	Description
CustomerId	Unique Customer Identifier
OrderDiscount	Discount on the Order
OrderTotal	Total Order Value After Discount
CountryId	Customer's Country Identifier
ProductId	Unique Product Identifier
Price	Product Price Including Tax
Discount	Discount Amount Including Tax
AllowEmail	Permission for Email Communication
AllowSms	Permission for SMS Communication
OrderCount	Total Number of Orders by the Customer
ReturnCount	Number of Returns or Cancellations
Month	Month of the Order
Day	Day of the Order
Hour	Hour of the Order
Return	Status of the Order

B. OVERVIEW OF METHODS

1. Multi-Layer Perceptron

MLP function is a complex function that produces numerical outputs from numerical inputs. The input layer acquires raw data from the domain, the hidden layer extracts features, and the output layer generates a prediction. This represents the architecture of a MLP. A deep neural network comprises numerous hidden layers. Conversely, increasing the number of hidden layers may lead to vanishing gradient problems, requiring the implementation of specialized techniques. The hyperparameters of the MLP, encompassing the quantity of hidden layers and hidden neurons, must be chosen meticulously [13].

2. Random Forest

RF is a widely utilized ensemble learning method applicable to regression, classification, clustering, and interaction detection. In contrast to a solitary decision tree (DT), which is often unstable and biased, RF constructs a multitude of trees, mitigating these concerns through the utilization of multiple DTs. Each tree is generated from a bootstrap sample, and at each node, a random selection of variables is utilized. Out-of-Bag (OOB) error rates are employed to evaluate the precision of each tree. The ultimate classification is determined by a majority vote across all trees. Key error metrics comprise the mean reduction in the Gini coefficient and accuracy. To enhance the RF model, users must modify two critical parameters: the overall quantity of trees and the number of variables evaluated at each node [14].

3. Extreme Gradient Boosting

XGBoost is an efficient gradient tree boosting algorithm that builds decision trees sequentially. It is known for its speed and effectiveness, especially in handling structured datasets for label classification. XGBoost combines predictions from multiple decision trees using the "bagging" technique and enhances model performance by reducing sequential errors. Gradient descent further refines these errors. XGBoost also improves the traditional gradient boosting by eliminating missing data issues and using parallel processing to reduce overfitting [15].

4. Support Vector Machine

SVM is a technique for pattern recognition grounded in statistical learning theory. SVM were initially designed for classification; however, their primary applications now encompass regression and the classification of small, high-dimensional, non-linear datasets. SVM are founded on the VC-dimension of statistical learning theory and the principle of minimizing structural risk. Learning occurs without error identification utilizing a constrained sample size, and the model's precision is evaluated. The minimal deviation of the hyperplane from the sample points is employed to ascertain the optimal universal capability. SVM encompass both linear and non-linear regressions. The kernel function, which assesses the similarity between data points, and the cost loss function, or regularization parameter, are critical parameters. (i.e., among reflectance values.) [16].

5. Linear Regression

In technical and scientific applications, LR is one of the most prevalent models for determining the relationship between two variables. Two categories of statistical methods—Type I, known as Ordinary Least Squares (OLS), and Type II, referred to as Standard Major Axis (SMA)—have been developed for LR, contingent upon the characteristics of the data set. In optical oceanography, the rationale for selecting a specific method to calculate a linear regression fit is often overlooked and seldom substantiated by statistical evidence [17].

C. SALES CANCELLATION/RETURN PREDICTION MODELS

Results have been obtained with 10-fold cross-validation using MLP, RF, XGBoost, SVM and LR algorithms on the dataset. To maximize the performance of each algorithm, hyperparameter optimization has been performed with grid search. The hyperparameter ranges used for grid search are provided in Table 2.

Table 2. Hyperparameter Ranges Used for Grid Search

Method	**Hyperparameter Range**
MLP	Layer number: (1 - 3) Neuron count: (32 - 128) Epochs: (5 - 10)
RF	N Estimators: (50 - 200) Max Depth: (10) Min Samples Split: (2 - 10)
XGBoost	N Estimators: (50 - 100) Max Depth: (None - 10) Learning Rate: (0.1 – 0.2) Max Leaves: (25 - 100)
SVM	C: (1 - 1000) Degree: (4) Gamma: (0.1 - 0.5)
LR	C: (0.001 - 10) Max Iter: (1000) Penalty: (L1, L2)

IV. RESULTS

The confusion matrices of these models are shown in Figure 2 to Figure 6.

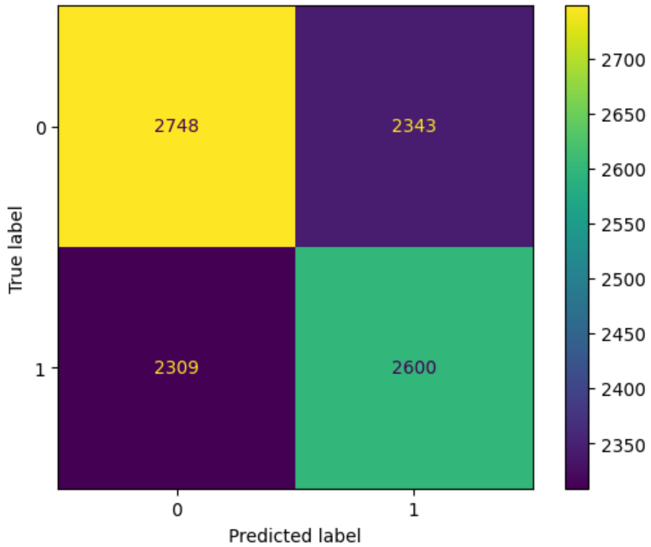

Figure 2. Confusion matrix obtained with MLP

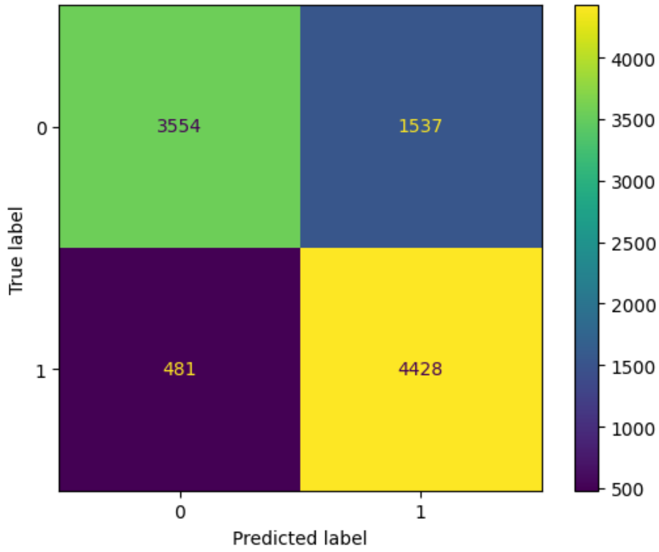

Figure 3. Confusion matrix obtained with RF

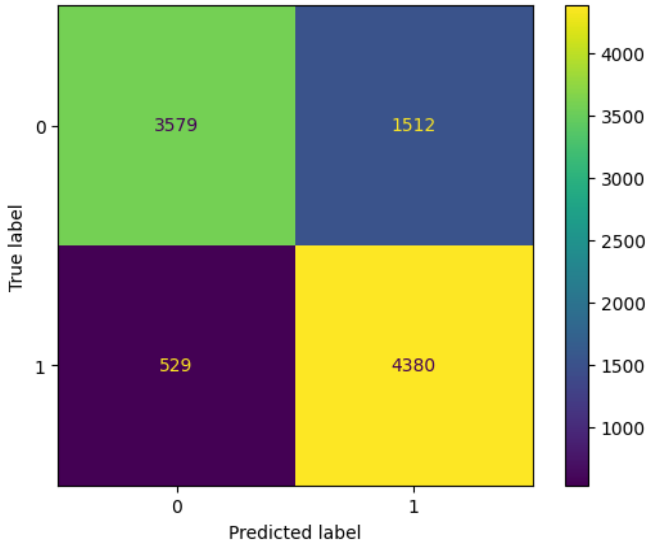

Figure 4. Confusion matrix obtained with XGBoost

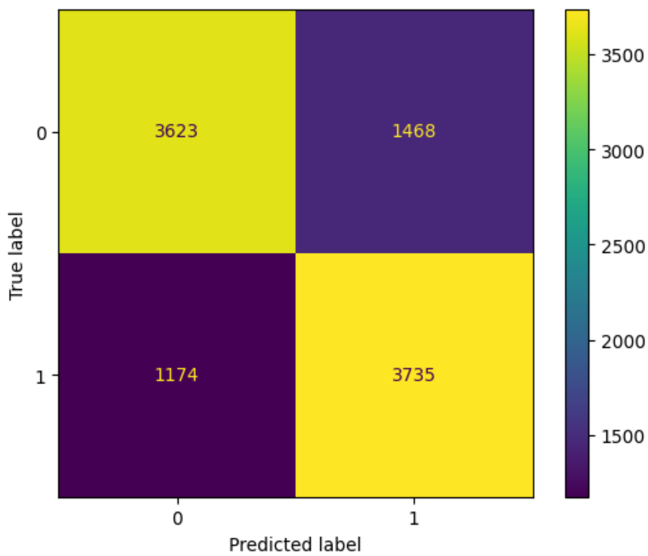

Figure 5. Confusion matrix obtained with SVM

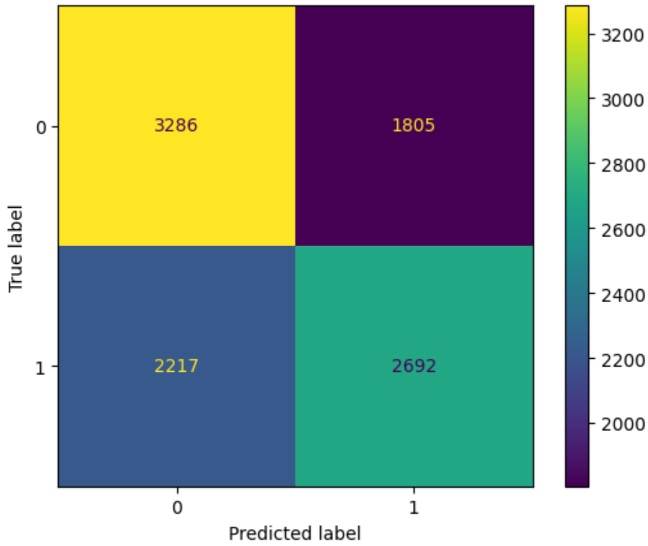

Figure 6. Confusion matrix obtained with LR

The comparison of precision, recall, F1-score, and accuracy values of the results obtained with the sales cancellation and return models developed using MLP, RF, XGBoost, SVM, and LR is given in Table 3.

Table 3. The Obtained Results

Algorithms	Precision	Recall	F1-Score	Accuracy
MLP	0.53	0.53	0.53	0.53
RF	0.74	0.90	0.81	0.80
XGBoost	0.74	0.89	0.81	0.80
SVM	0.72	0.76	0.74	0.74
LR	0.60	0.55	0.57	0.60

V. DISCUSSION

Based on the results,

- RF and XGBoost models demonstrated superior performance compared to other algorithms, achieving the highest F1-score and accuracy, along with high recall rates. These characteristics suggest that these models are the most suitable for predicting cancellations and returns in e-commerce.
- The MLP model produced the lowest results across all evaluation metrics, indicating that it struggles to capture the relevant features of the dataset effectively.
- The SVM model exhibited balanced performance in terms of F1-score and accuracy. However, it was not as effective as RF and XGBoost, indicating that while it is an acceptable option, it is not the optimal choice for this problem.
- The LR model achieved a reasonable level of accuracy and F1-score, but it underperformed in terms of recall compared to other models. This suggests that LR may be less suitable in scenarios where minimizing false negatives is critical.

VI. CONCLUSION

This study utilized several machine learning algorithms to develop forecasting models for predicting sales cancellations and returns. Among the models evaluated, XGBoost and RF demonstrated the best overall performance, particularly excelling in accuracy and recall. This suggests that these models are more robust in effectively identifying potential cancellations and returns. The SVM model also provided balanced results across most evaluation metrics, while the LR and MLP models exhibited comparatively lower effectiveness. Notably, the LR model had the lowest performance metrics, indicating a higher occurrence of false positives and false negatives.

To further enhance the accuracy of these forecasting models, future work will focus on leveraging the strengths of various feature selection algorithms to better assess the impact of features on model performance. Instead of changing the feature set, this approach aims to provide a more accurate evaluation of the contribution of existing features, ultimately helping e-commerce companies make more precise and proactive decisions, thereby reducing operational inefficiencies and increasing overall profitability.

REFERENCES

[1] Ministry of Trade of the Republic of Turkey. (May 2024). E-commerce outlook report in Turkey.
[2] Chaturvedi, H. (2022, May 30). Product returns are wasteful for companies and the planet. Here's how to change that. Fast Company
[3] Adebanjo, S., & Banchani, E. (2023). DEVELOPMENT OF MACHINE LEARNING TECHNIQUES FOR DETERMINING OPTIMAL PRICING AND INVENTORY POLICY WITH ORDER CANCELLATION UNDER THE CASH-ON-DELIVERY PAYMENT SCHEME. Advance Journal of Current Research, 8(12), 18-41.
[4] Chen, S., Ngai, E. W., Ku, Y., Xu, Z., Gou, X., & Zhang, C. (2023). Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. *Decision Support Systems*, 170, 113959.
[5] Dzyabura, D., El Kihal, S., Hauser, J. R., & Ibragimov, M. (2023). Leveraging the power of images in managing product return rates. *Marketing Science*, 42(6), 1125-1142
[6] Jauhar, S. K., Chakma, B. R., Kamble, S. S., & Belhadi, A. (2024). Digital transformation technologies to analyze product returns in the *e-commerce* industry. *Journal of Enterprise Information Management*, 37(2), 456-487.
[7] Liu, Z., Jiang, P., Wang, J., Du, Z., Niu, X., & Zhang, L. (2023). Hospitality order cancellation prediction from a profit-driven perspective. *International Journal of Contemporary Hospitality Management*, 35(6), 2084-2112.
[8] Qiu, R., Ma, L., & Sun, M. (2023). A robust omnichannel pricing and ordering optimization approach with return policies based on data-driven support vector clustering. *European Journal of Operational Research*, 305(3), 1337-1354.
[9] Sun, H., Lv, Z., Li, J., Xu, Z., & Sheng, Z. (2023). Will the Order Be Canceled? Order Cancellation Probability Prediction Based on Deep Residual Model. *Transportation Research Record*, 2677(6), 142-160.
[10] Sweidan, D., Johansson, U., Alenljung, B., & Gidenstam, A. (2023, July). Improved Decision Support for Product Returns using Probabilistic Prediction. *In 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)* (pp. 1567-1573). IEEE.
[11] Wang, Z., Luo, C., Luo, X. R., & Xu, X. (2024). Understanding the effect of group emotions on consumer instant order cancellation behavior in livestreaming E-commerce: Empirical evidence from TikTok. *Decision Support Systems*, 179, 114147.
[12] Xu, W., Zhang, X., Chen, R., & Yang, Z. (2023). How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce. Decision Support Systems, 172, 113984.
[13] Al Bataineh, A., & Manacek, S. (2022). MLP-PSO hybrid algorithm for heart disease prediction. *Journal of Personalized Medicine*, 12(8), 1208.
[14] Park, S., & Kim, J., Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. *Applied Sciences*, 9(5), 942, 2019.
[15] Park, S., & Kim, J., Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. *Applied Sciences*, 9(5), 942, 2019.
[16] Yuan, H., Yang, G., Li, C., Wang, Y., Liu, J., Yu, H., ... & Yang, X. (2017). Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: Analysis of RF, ANN, and SVM regression models. *Remote Sensing*, 9(4), 309.
[17] Bellacicco, M., Vellucci, V., Scardi, M., Barbieux, M., Marullo, S., & D'Ortenzio, F. (2019). Quantifying the impact of linear regression model in deriving bio-optical relationships: the implications on ocean carbon estimations. *Sensors*, 19(13), 3032.