İnsan ve Toplum Bilimleri Araştırmaları Dergisi ®
Jornal of the Human and Social Science Researches
[2147-1185]
11 th Years

[itobiad], 2022, 11 (3): 1493-1514

E-Ticaret Alanı İçin Sipariş İptallerini Tahmin Etme: Perakendecilik Deneyimine Dayalı Önerilen Bir Model

Predicting Order Cancellations for E-Commerce Domain: A Proposed Model Based on Retailing Experience

Video Link: https://youtu.be/scV8FA5_VwI

Kevser ŞAHİNBAŞ

Dr.Öğr.Üyesi, İstanbul Medipol Üniversitesi İşletme ve Yönetim Bilimleri Fakültesi, Yönetim Bilişim Sistemleri Bölümü

Asst. Prof, Istanbul Medipol University Business School, Department of Management Information System

ksahinbas@medipol.edu.tr

ORCID: 0000-0002-8076-3678

Makale Bilgisi / Article Information

Makale Türü / Article Type : Araştırma Makalesi / Research Article
Geliş Tarihi / Received : 07.06.2022
Kabul Tarihi / Accepted : 12.09.2022
Yayın Tarihi / Published : 19.09.2022
Yayın Sezonu : Temmuz-Ağustos-Eylül
Pub Date Season : July-August-September

Atıf/Cite as: Şahinbaş, K. (2022). Predicting Order Cancellations for E-Commerce Domain: A Proposed Model Based on Retailing Experience . İnsan ve Toplum Bilimleri Araştırmaları Dergisi , 11 (3) , 1493-1514 . doi: 10.15869/itobiad.1127578

İntihal-Plagiarism/Etik-Ethic: Bu makale, iTenticate yazılımınca taranmıştır. İntihal tespit edilmemiştir/This article has been scanned by iTenticate.

Etik Beyan/Ethical Statement: Bu çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur/It is declared that scientific and ethical principles have been followed while carrying out and writing this study and that all the sources used have been properly cited (Kevser ŞAHİNBAŞ)

Predicting Order Cancellations for E-Commerce Domain: A Proposed Model Based on Retailing Experience

Abstract

E-Commerce technologies enable contact between businesses and their suppliers for the aim of exchanging information such as purchase orders, invoices, and payments thank to the rapid development in information technologies. E-Commerce has become a particularly important concept and has revolutionized the retail space. Understanding customer behavior patterns is key to gaining competitive advantage and achieving business goals. Predicting the probability of order cancellations has become a very urgent need as it causes loss of revenue for the retailer. When dealing with day-to-day operations such as order processing, tracking and order cancellations, finding enough time to grow the business is difficult. Cancellations are an important aspect of retail industry revenue management. In fact, little is known about the factors that cause customers to cancel or how to avoid them. The aim of this study is to propose a model that predicts the tendency to cancel an order and the parameters that affect the cancellation of the order. This solution can identify key factors that cause orders to be canceled by analyzing historical transaction data. A custom modeling application has been created that helps automate the process of tracking order cancellations in real time and predict the probability of an order being cancelled. For this purpose, machine learning techniques (ML) such as Artificial Neural Network, Support Vector Machine, Linear and Logistic Regression, XGBoost, Random Forest are applied to provide a tool for predicting order cancellations. The Random Forest algorithm achieves the best performance with 86% accuracy and 88% F1-Score compared to the other algorithm. This work will help firms manage their inventories well and strengthen their actions regarding customer behavior.

Keywords: Classification in E-Commerce Cancellation, Marketing Strategies, Data Management, ANN, SVM, XGBoost, Logistic Regression, Parameter Tuning, Feature Importance

E-Ticaret Alanı İçin Sipariş İptallerini Tahmin Etme: Perakendecilik Deneyimine Dayalı Önerilen Bir Model

Öz

E-Ticaret teknolojileri, bilgi teknolojilerindeki hızlı gelişme sayesinde, işletmelerin satın alma siparişleri, faturalar, ödemeler gibi bilgi alışverişi amacıyla tedarikçileri ile iletişim kurmasını sağlamaktadır. E-Ticaret özellikle önemli bir kavram haline gelmiştir ve perakende alanında devrim yaratmıştır. Müşteri davranış kalıplarını anlamak, rekabet avantajı elde etmenin ve iş hedeflerine ulaşmanın anahtarıdır. Perakendeci için gelir kaybına neden olduğu için sipariş iptallerinin olasılığını tahmin etmek çok acil bir ihtiyaç haline gelmiştir. Sipariş işleme, takip ve sipariş iptalleri gibi günlük işlemlerle uğraşırken, işi büyütmek için yeterli zaman bulmak zordur. İptaller, perakende sektörü gelir yönetiminin önemli bir yönüdür. Aslında, müşterilerin iptal etmesine neden olan

faktörler veya bunlardan nasıl kaçınılacağı hakkında çok az şey bilinmektedir. Bu çalışmanın amacı, bir siparişi iptal etme eğilimini ve siparişin iptalini etkileyen parametreleri tahmin eden bir model önermektir. Bu çözüm, geçmiş işlem verilerini analiz ederek siparişlerin iptal edilmesine neden olan temel faktörleri belirleyebilir. Sipariş iptallerini gerçek zamanlı olarak izleme sürecini otomatikleştirmeye ve bir siparişin iptal edilme olasılığını tahmin etmeye yardımcı olan özel bir modelleme uygulaması oluşturulmuştur. Bu amaçla Yapay Sinir Ağı, Destek Vektör Makinesi, Doğrusal ve Lojistik Regresyon, XGBoost, Rastgele Orman gibi makine öğrenme teknikleri uygulanarak sipariş iptallerini tahmin etme aracı sağlanmıştır. Rastgele Orman algoritması diğer algoritmaya göre %86 doğruluk oranı ve %88 F1-Score ile en iyi performansı elde etmektedir. Bu çalışma, firmaların envanterlerini iyi yönetmelerine ve müşteri davranışlarıyla ilgili eylemlerini güçlendirmelerine yardımcı olacaktır.

Anahtar Kelimeler: E-Ticaret İptalinde Sınıflandırma, Pazarlama Stratejileri, Veri Yönetimi, YSA, SVM, XGBoost, Lojistik Regresyon, Parametre Ayarlaması, Nitelik Önemi

Introduction

The development of technology has provided innovation and convenience in many sectors. In this context, where the concept of speed and cost is emphasized, companies in all sectors have considered to keep up with technology. Otherwise, companies that do not keep up could be eliminated. With the spread of the Internet to the world, the idea of transferring the concept of commerce to the virtual environment has also emerged. In this way, the concept of e-commerce emerged, and this concept continues to increase its importance in coordination with technology (Gong, 2021, p. 694).

The concept of commerce has gained more importance with the development of technology today and has been called E-Commerce. E-Commerce is the transmission of goods and services by individuals and institutions in a network through certain parameters (Hamed & El-Deeb, 2020, pp. 242-244). The biggest reason it has gained much more importance today is the development of technology and the easier access to information. In addition, E-Commerce has become more preferred because customers can compare prices and access price performance products more easily.

Today E-commerce and retailing, which are the most common uses of Machine Learning (ML), are two of the most important industries. With easy access to information, understanding how to do extremely complex and difficult work faster on the Internet, and understanding the concept of cheap products and good service have now become one of the most important elements. As a result, e-commerce and retailing have become inextricably linked (Zhao, 2018, p. 1868).

Thanks to ML, retail companies have an opportunity to provide the models by processing the existing data and take more profitable processes with their goals. There are retail companies that do not correctly implement many success parameters such as knowing their consumers and giving the right product to the right consumer. Therefore, there is a lack of data analytics in most industries and most companies. A large amount of data is retained, but there are persistent gaps and deficiencies in the analysis of the data held (Ahmed, 2004, p. 455). The fact that the stored data is dysfunctional means a great waste

and cost for companies. For this reason, it is of great importance for companies in every sector that the data is processed, and the findings obtained from the processed data are modeled and delivered to the right consumer. Consequently, organization and planning are significant, and to have an accurate prediction model is crucial. This study is applied in order to fill the knowledge and method gap and is an idea for future studies.

Large consumer datasets can reveal hidden clues that can be used to strengthen customer relationships, optimize marketing campaigns, forecast sales, and boost competition, sales, and turnover. Businesses can deliver the correct targeted campaigns and locate the offer that has the most impact on the consumer with more precise data models (Yeung, 2014, pp. 943-944). When the strength of e-commerce is paired with retailing companies' high-volume sales, a massive data stack emerges. This demonstrates how critical and effective machine learning is in these industries. The main reasons behind this study and research problem are to process customer and order related data, reveal data analysis and propose a decision support system to managers to cope with the loss of revenue because of cancellation. Finding enough time for business expansion while dealing with day-to-day operations such as order processing, tracking and order cancellations is a big challenge for the business. Besides, the biggest financial loss for sellers in e-commerce is requesting the cancellation or return of the ordered product. While the revenues already earned are lost due to cancellations and returns, it may lead to large profit losses in the long run due to shipping costs and customer dissatisfaction.

The research subjects of this study are to analyze the cancellation / return behavior of customers, the factors that affect these behaviors and to discuss machine learning and its methods, e-commerce, e-retail, retail sector and machine learning studies in this sector. The topics covered were examined in detail, along with findings and examples, the behavior of customers in the retail industry to cancel/return their orders was analyzed using machine learning methods. In this study, a comparative study of ML algorithms is conducted to predict cancellation order of the company. The performance metrics are detailed, and the findings form experiments are indicated. For this study, supervised learning algorithms such as Support Vector Machine (SVM), XGBoost, Random Forest, Artificial Neural Networks (ANN) and Logistic Regression and tuned_Random Forest, tuned_logistic regression and tuned_XGBoost are applied. Among the algorithm RF algorithm achieved the best performance with 86% accuracy ratio and 88% F1-Score value.

The related studies about data analysis of ecommerce domain used low level parameters of dataset. Parameter tuning and feature importance have not been encountered in the current studies. In this study, the hyperparameters of each of the machine learning algorithms have been adjusted by parameter tuning to achieve better accuracy and measurements. The neural network is implemented using additional layers and some more training time. Comprehensive and comparative analysis has been made using five most used machine learning algorithms and three methods with parameter tuning. A large parameter scale containing 20 parameters is used. These constitute the originality of this study.

Since important features can be extracted with this study, companies in the field of e-commerce will have information about which variables have a significant effect on cancellation/return intention. Firms will have the ability to predict the probability of

cancellation, thus offering strategies to improve the ordering process and customer satisfaction and ultimately ensure customer loyalty.

The main contribution of this paper is to provide order cancellation model by different Machine Learning classification algorithms. Besides, feature importance by Logistic Regression and XGBoost are obtained. Consequently, firms are aware about the important factors that have influence on the cancel/return decision. This study could be used as a decision support system.

Retailing and E-Retail Sector

The concept of retailing has lost its old meaning, and now it has gained a new meaning and the concepts of e-retail have emerged. Technology has a significant impact on this issue because people now visit the online stores and decide accordingly, rather than shopping in stores (Erkent, 2006, p. 10).

There are many reasons for this situation. Consumers always prefer the cheapest and best quality before purchasing a product. In order to obtain this product, they go to many varied brands and stores and make comparisons. Going to a different physical store to make a comparison creates a disadvantage in terms of time, cost, and objectivity (Güllü & Tarhan, 2021, pp. 196-197). Technology has overcome all such disadvantages. Finding a product cheaper, better quality takes seconds now and at no extra cost. This situation pushes many retail businesses to understand the importance of the concept of e-retail. Most retail businesses keep up with technological advancement. Many businesses that do not comply are doomed to huge losses (Koçal, 2012, pp. 14-15). The e-retail sector, which offers the opportunity to get ideas about the product before purchasing by communicating with different consumers, continues to spread rapidly (Erkent, 2006, p.10). When consumers want to buy products, they prefer the physical store only to be quite close to the store or to be sure of the quality of the food products. Considering the latest innovations in this sector, the e-retail sector is now preferred even in food products.

Related Works

Özcan and Turna revealed a Decision Tree (DT) and ML study with the online shopping of consumers. The difference of this study is that the data obtained is through a survey study. While the number of data obtained by the authors in their survey was 250, the number of data used in this study exceeds 30 thousand. The amount of data is especially important when doing DM work. Therefore, the accuracy and Cohen's kappa rates obtained from survey studies are generally low. While the accuracy scores of their study are at most 68%, the highest accuracy score in this study is 82% with the DT method. Of course, other factors can also support this. In addition, a detailed DT study was conducted by using many different algorithms of the DT and the best algorithm was selected. However, in this study, ID3 was used only with gain ratio. This shows the weakness of this study, at least in terms of the DT. If better results are desired in the study, different algorithms should be used and the data should be manipulated less (Özcan & Turna, 2021, p. 94). Liu et al. (2020) propose a new analysis of online shopping behavior and forecasting model that overcomes the shortcomings of the old methods. In the study, decision tree based XGBoost model and linear model logistic regression are used. In their study, the model is optimized, and a single model is combined. The model fusion technique is applied. Finally, through two sets of contrast experiments, it has been

proven that the algorithm chosen in their study can filter out features effectively that simplify the complexity of the model to a certain extent and improve the accuracy rate of the model (Liu et al, 2020, pp. 1-5). Abhirami et al. focused on E-commerce fraud detection and compares various ML algorithms such as K-nearest neighbor, Decision Tree, LR and Random Forest. It has been observed that logistic regression gives better results than other algorithms researched on fraud detection. Kaggle dataset was used in this study. The dataset contains 284807 data, 492 of which are fraudulent. The Decision Tree Algorithm gave the minimum performance (Abhirami et al.,2021, pp. 827-828). Noor et al. proposed a model based on sentiment analysis. An aspect-level sensitivity analysis is performed on an e-commerce website. The dataset includes reviews of women on clothing sold on Amazon. Weka is used. SMO performed better than other algorithms in Sensitivity analysis. It was the best performing algorithm after SMO in NB (Noor et al.,2019, p.5). Mauritsius et al. present a study to decide whether some data mining approaches such as Random Forest Algorithm and J48 Algorithm can be applied to detect promotion abuse based on existing customer profiles. The data used in the study covers the transaction history recorded in 2018 and 2019. As a result of the study, it shows that using RFA, FDS functionality works with planning and is more successful than J48 Algorithm. RFA for fraud or non-fraud was chosen as the most accurate model for analyzing and predicting customers (Mauritsius et al.,2021, p. 6). Ballestar et al. develop a model that can predict the change of consumers' behavior over time. In the model, MLP and ANN are used. Their Feed Forward types have been used. Customer quality can be estimated using only data from references (Ballestar et al.,2018, p. 590). Rai et al. provide comparative analysis of the performance of Decision Tree ensembles, linear models and deep learning techniques in demand forecasting (SOTA) of e-commerce advertisements. A dataset including about 1.4 million C2C e-commerce advertisement training samples is processed. As a result, it is observed that deep learning is an effective method for demand forecasting studies (Rai et al., 2019, p. 5). Koehn et al. present a methodology compatible with the sequential structure of click data is proposed by using RNN instead of SML. Two models were evaluated to derive user behavior predictions from clickstream data (Koehn et al.,2020, p. 16). Pondel et al. present a multilayer perceptron (MLP) with one or two fully connected dense layers is applied. Besides, the repetitive layer is used as a first hidden layer (RNN) supported by an additional dense layer. The data were prepared using the Keras 2.3.1 library (Pondel et al.,2021, pp. 4-6). Singh et al. provide to predict sales for e-commerce platforms to find the best algorithm. Random Forest and Gradient boosting showed good results for the data. The most efficient method for the regression model was found to be Random forest. For the time series model, the most efficient model is selected as the SARIMA model (Singh et al.,2020, p. 6). Szabó et al. predict what users want to buy while using an e-commerce application. In the RMSE and MAE value results, the solution estimated the transformations with remarkable accuracy and improves traditional CF (Szabó et al.,2020, p. 6). Yin et al. develop a sales forecasting model for online products. In the research, the CNN model is compared with other models and as a result, it exhibits the most active performance. The unsupervised pre-training method can also be used for pre-training impact analysis with algorithms such as RBM and DAE (Yin et al., 2021, pp. 4-5).

System Overview

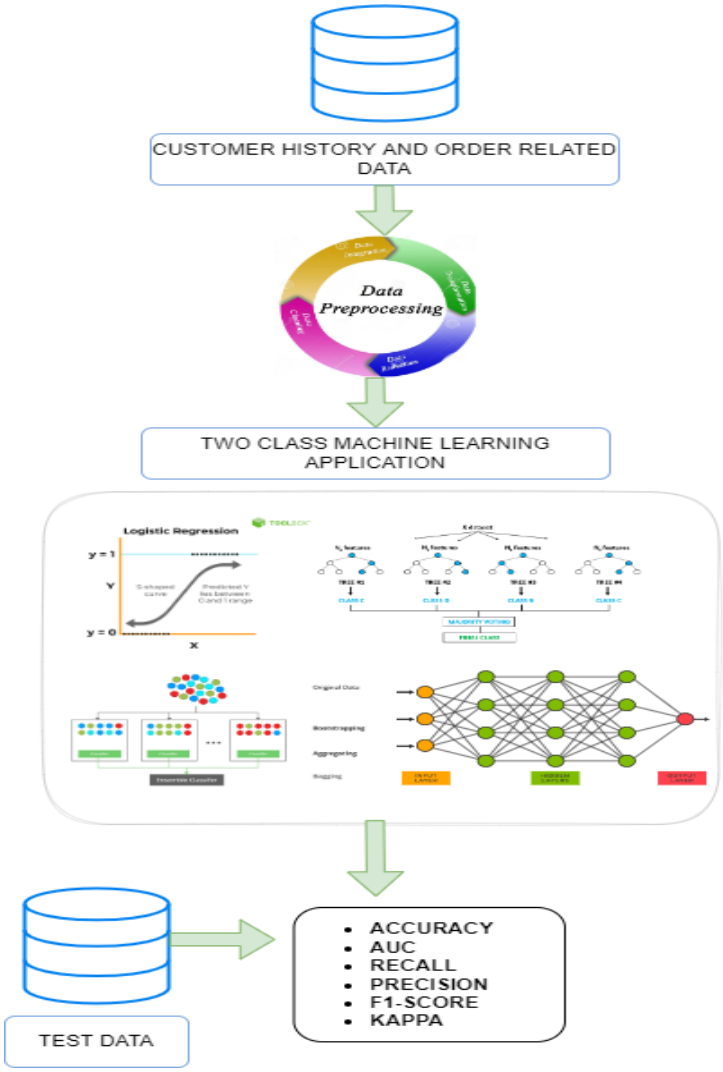


Figure 1. System Architecture

Data for order cancellation is gathered from a private ecommerce firm. The diagram of the system architecture is illustrated in Figure 1. Data is preprocessed firstly. Then XGBoost, ANN, SVM, LR, RF, tuned_RF, tuned_LR, tuned XGBoost are applied to preprocessed data. As a result, the model generates the classes that labeled ordered as 0 and cancelled/returned as 1 and accuracy and F1- Score values.

Materials and Methods

In this study, supervised learning algorithms such as Support Vector Machine (SVM) (Vapnic, 1995, pp. 123-125), XGBoost (Chen & He, 2017, p. 4), Random Forest (Breinman, 2001, pp. 10-12), Artificial Neural Networks (ANN) (Fritsch et al., 2019, p. 44) and Logistic Regression (Peduzzi, 1996, p. 1374) are applied by using Python.

Dataset

The data has been taken from the database of a private retail company in e-commerce sites. Data covers January, February, March, April and May of 2022. In the analysis, 30.231 includes customer and order history data are distributed according to the membership type, campaign type and gender of people living in different cities of Turkey. The data belonging to the enterprise are grouped mainly on textile products.

Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) is a computer system inspired by biological neural networks, built on interconnected artificial neurons to create artificial brains. It is designed to analyze information, like the way biological brains, especially humans, analyze it. It is one of the most used ML methods to deal with enormous amounts of data. It also has self-learning abilities. ANN structure consists of input layer, hidden layer, and output layer (Öztemel, 2012, pp. 60-64). Considering these layers, the program takes information as input. Then it is processed in the hidden layer, undergoes mathematical operations, and is optimized. Then the output is obtained, but the output may not be of the desired type. The output is returned to the hidden layer and subjected to some processing (Jiang et al., 2022, p. 10). Therefore, there is a continuous loop between the hidden and output layer.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm tries to fix a boundary surface between two classes based on the features of data (Romero Morales and Wang, 2010, p. 556) applying the structure risk minimization method. The support vector classifier method aims for the minimum generalization error among an infinite number of linear classifiers for two linearly separable classes by maximizing the margin defined as the smallest distance between the decision boundary and any classification (Bishop, 2006, p. 738). On the other hand, linearly separable classes are uncommon in nature, thus, to address this problem, the data space is transformed into a high-dimensional space where they can be segregated linearly (Amari and Wu, 1999, p. 785).

The main purpose of SVM is to separate the vectors belonging to different classes from each other and to obtain the optimal separation hyperplane (Cortes and Vapnik, 1995, pp. 280-282).

$$y = f(x) = \sum_{k=1}^{m} \bar{\alpha}_k \, .K(x, x_k) + b \qquad (1)$$

XGBoost

XGBoost (eXtreme Gradient Boosting) is a high-performance version of the Gradient Boosting algorithm optimized with various modifications. The most important features of the algorithm are that it can achieve high predictive power, prevent over-learning, manage empty data and do them quickly (Chen & He, 2017, p. 4). Among the machine learning applications that work smoothly that we use in our daily lives, there are applications such as spam blocker, ad advice, fraud detection (Chen & Guestrin, 2016, pp. 785-787). Among those who ensure the success of these applications are scalable learning structures that enable learning over large databases and models that find connections between data. Among the multiple algorithms for tree strengthening, XGBoost is the

most preferred algorithm by people for many classifications such as text classification, customer prediction, motion perception (Chen & Guestrin, 2016, p. 790). The purpose of the boosting algorithms is to find the hard learners from the data in the model and process them again and again to integrate them with the easier learners (Bentéjac, Csörgő, & Martínez-Muñoz, 2021, p. 1937). There are 3 different boosting techniques as XGBoost, gradient boost, regular boost and one of them can be used. It uses decision trees to use supervised learning as in the random forest algorithm, but the difference in the random forest algorithm is that it takes a different path during the training phase; XGBoost tries to obtain information about them by optimizing the decision trees. The best tree is selected until the best model is obtained and these trees are added until the goal is achieved (Dhaliwal, Nahid, & Abbas, 2018, p. 149).

Random Forest (RF)

Random Forest algorithm enables to create various models and classifications by training each decision tree on a different observation sample on more than one decision tree. Ease of use and flexibility; It has accelerated its adoption and widespread use as it addresses both classification and regression problems (Bonaccorso, 2017, pp. 167-168). Over-learning-over-fitting is one of the main issues with decision trees, which is one of the traditional methodologies. The random forest model generates multiple decision trees to solve this problem, thus avoiding this problem and at the same time aiming to increase the classification value during the classification process (Bonaccorso, 2017, p. 170). The random forest algorithm is the process of choosing the highest score among many independent decision trees. As the number of trees increases, our rate of obtaining accurate results also increases. The main difference between the decision tree algorithm and the random forest algorithm is that the process of finding the root node and splitting the nodes is random (Breiman, 2001, pp. 10-12).

Logistic Regression (LR)

Logistic Regression is a DM classification method used in classification and assignment processes (Peduzzi, 1996, p.1374). Unlike linear regression analysis, it is frequently used today because it also benefits from qualitative features. In this analysis method, it can be applied categorically on both ordered and unordered data. In the logistic regression model, all independent variables in the dataset must be included in the model. Likewise, all unsuitable independent variables should be excluded from the model. Normalization is important in this classification method as there should be no extreme values. Because excessive values will disrupt the structure of the model, healthy results will not occur. In the logistic regression model, some tests are needed to evaluate the compatibility of the model. These are Chi-Square and R^2 measurement methods.

Evaluation

Confusion Matrix

When analyzing results from classification algorithms, a confusion matrix, which is a contingency table that illustrates the difference between the actual class and the predicted one for the test set in a labeled table, is frequently used (Bradley, 1997, p 1145). The prominent parameters in the evaluation of classification models are accuracy, precision, recall and F1 score (Visa, Ramsay, Ralescu, & Knaap, 2011, pp. 120-127). The parameters

are calculated from the numbers produced by the confusion matrix called TP, TN, FP and FN. The equation and definition of the parameters are explained below.

The accuracy value, which is one of the performance measures of classification models, shows the ratio of correctly predicted data to all predicted values. It is obtained with the formula shown in Equation 2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is the ratio of positive samples classified correctly with this criterion to the total positive predicted samples are measured. It is given at Equation 3.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall is the ratio of correctly classified positive data to total positive data. From the data of the model created using this metric, the rate of finding positive class labels are determined. Formula is shown in 4.

$$Recall = \frac{TP}{DP+YN} \quad (4)$$

F value, which is a harmonic mean of sensitivity and precision performance measures, evaluates two different performance measures within itself thanks to this feature. This metric gives a single benchmark (Eq. 5).

$$F1\ Score = 2\frac{Precision*Recall}{Precision+Recall} \quad (5)$$

Experiments and Findings

The proposed model is applied in 7 steps that include Data Preparation, Data Exploration, Feature Engineering, Feature Reduction, Model, Model Interpreter, and Tune Model.

Data Preprocessing

Before performing ML, it is necessary to make sure that the dataset is ready. For example, missing information in the dataset, columns that are not wanted to be included in the analysis and extreme values can greatly affect the analysis results. It can even cause undesirable results. Therefore, data preprocessing is important for more accurate analysis.

Input Variables

In Table 1, the column names in the data set and whether these columns are numerical or categorical are specified. It is also shown in this table that the columns do not contain a missing value. Statistical values of the columns are also indicated on this table. In the following paragraph these categorical data are briefly explained. Platform includes which platforms the order was received from, such as iOS, Android and Website. City shows the city from which the order was received. Region indicates the region of the city where the order was taken. Day shows on which day the order was received. Campaign shows whether he has benefited from the campaign in that order. Member shows that the person who made the order is a registered or unregistered customer. Type shows that the products in that order are new season or regular products.

Table 1. Explanatory available variables of the dataset

Column	Type	Unique	Missing	Statistics						
platform	Numeric	3	0	min 1	max 3	median 2	mean 1,82	std 0,87	25% 1	75% 3
city	Numeric	81	0	min -1,57	max 2,24	median 0	mean -0,09	std 0,85	25% -0,67	75% 0,33
product_number	Numeric	16	0	min 1	max 20	median 1	mean 1,06	std 0,44	25% 1	75% 1
gender_category	Numeric	4	0	min 1	max 4	median 2	mean 1,68	std 0,75	25% 1	75% 2
product_category	Numeric	26	0	min 2	max 27	median 5	mean 6,26	std 3,1	25% 4	75% 8
price	Numeric	147	0	min -0,82	max 1,39	median 0	mean 0,11	std 0,62	25% -0,42	75% 0,58
day	Numeric	7	0	min 0	max 6	median 3	mean 2,8	std 2,03	25% 1	75% 5
region	Numeric	7	0	min -1	max 0,5	median 0	mean -0,01	std 0,54	25% -0,5	75% 0,5
time	Numeric	4	0	min 1	max 4	median 2	mean 1,98	std 0,96	25% 1	75% 3
member_type	Numeric	2	0	min 1	max 2	median 1	mean 1,29	std 0,45	25% 1	75% 2
type	Numeric	2	0	min 1	max 2	median 2	mean 1,52	std 0,5	25% 1	75% 2
campaign	Numeric	2	0	min 1	max 2	median 2	mean 1,6	std 0,49	25% 1	75% 2
order_women_total	Numeric	34	0	min 0	max 53	median 0	mean 1,02	std 3,49	25% 0	75% 1
order_man_total	Numeric	43	0	min 0	max 304	median 1	mean 5,28	std 24,45	25% 0	75% 3
order_kid_total	Numeric	37	0	min 0	max 173	median 0	mean 2,06	std 10,82	25% 0	75% 0
order_tototal	Numeric	60	0	min -0,5	max 76,25	median 0	mean 1,59	std 6,99	25% -0,25	75% 0,75
order_not_cancel_total	Numeric	283	0	min -0,5	max 152,5	median 0	mean 1,22	std 8,08	25% -0,5	75% 0,5
order_cancel_total	Numeric	271	0	min	max	median	mean	std	25%	75%

				-0,67	101	0	1,31	5,56	-0,33	0,67
cancel_percentage	Numeric	101	0	min -1,98	max 0,1	median 0	mean -0,36	std 0,56	25% -0,9	75% 0,1
cancel_return	Numeric	2	0	min 0	max 1	median 1	mean 0,64	std 0,48	25% 0	75% 1

Output Variable

The statistical values of the column used as the target are presented in Table 2.

Table 2. Output Variable

Column	Type	Unique	Missing	Statistics							
				min	*max*	*median*	*mean*	*std*	*25%*	*75%*	*missing_ratio*
cancel_return	Numeric	2	0	0	1	1	0,64	0,48	0	1	0

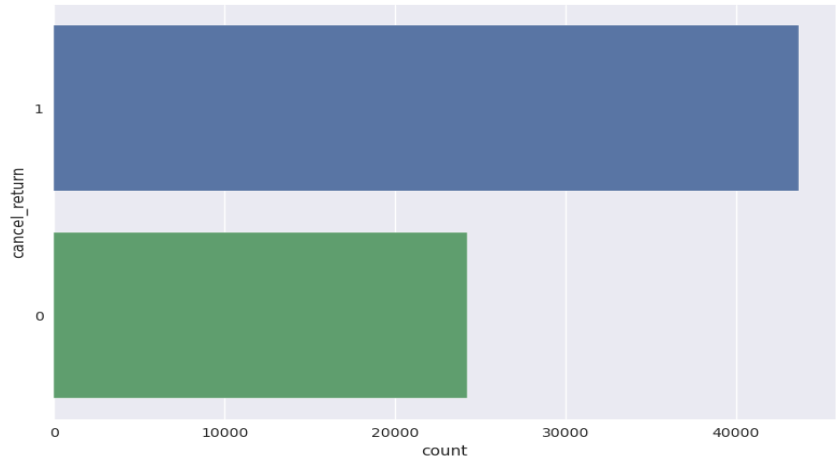


Figure 2. Number of Cancel/Return

Correlation

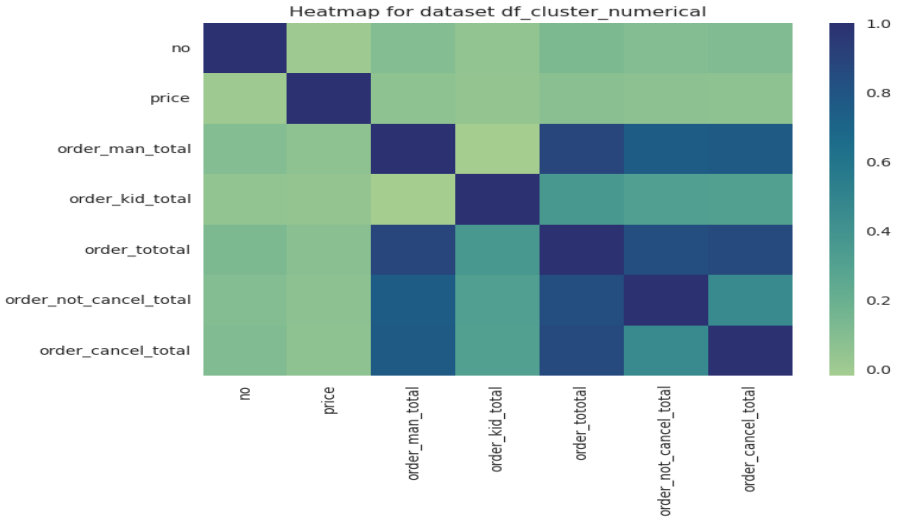


Figure 3. Variables Correlation

Values related to each other are specified in the correlation matrix. The correlation matrix is a table of correlation coefficients between multiple variables. In this table, the correlation between one variable and every other variable can be seen. By removing the variables that have very strong relationships with each other from the model, repetitive information entry can be avoided, and a simpler model can be obtained. Findings from Figure 3 show that the correlation coefficient between the "male order total" feature and the "total order" variable is 0.89, and that there is a strong and positive relationship between them. At the same time, the correlation coefficient between "man order total" and "order not cancellation total" and "order cancellation total" correlations are 0.81. There is a low and positive correlation between “price” and other variables, indicating that the price value is increasing, the other value is slightly increasing. This means that there is a positive and strong correlation between them. Accordingly, the "man order" value increased and the "order cancelled" value and the "order not cancelled" value increased as well. These relationships, which occurred as a result of the correlation analysis, should not be interpreted as a causal relationship.

Missing Value

Finding missing values is of immense importance for the ML process. Because the machine does not detect the missing, null information, even if it does, it may lead to wrong calculations while taking it into account. So, text, numerical etc. whatever the data is, missing values must be eliminated.

Encoding method

The Encoding method applied to the platform, city, product category, Day, cancel_return, campaign columns, which are categorical variable, are specified in the table.

Scaling

Table 3. Scaling of the columns

Column Name	Scaling Method
City	robust_scaler
Price	robust_scaler
Region	robust_scaler
order_women_total	robust_scaler
order_kid_total	robust_scaler
order_tototal	robust_scaler
order_not_cancel_total	robust_scaler
order_cancel_total	robust_scaler
cancel_percentage	robust_scaler

Column names of numerical variables and applied Scaling Methods are given in the Table 3.

Model and Parameter Tuning

The features of the models that will be applied to the dataset are indicated in the Table 4.

Table 4. Parameter Tuning for the Algorithms

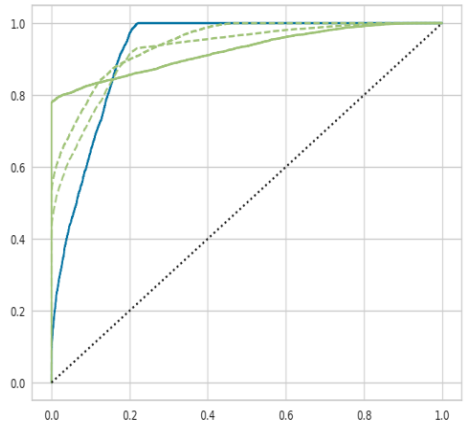
Algorithm	Params
logistic_regression	C: 1 tol : 0.0001 penalty : l1 l1_ratio : 0 max_iter : 1000 class_weight : balanced fit_intercept : True linesearch_max_iter : 50
Xgboost	eta : 0.3 alpha : 0 gamma : 0 lambda : 1 max_bin : 256 max_depth : 6 subsample : 1 grow_policy : depthwise max_delta_step : 0 sampling_method : uniform min_child_weight : 1 scale_pos_weight : 1 num_parallel_tree : 1
random_forest	n_bins : 128 bootstrap : True max_depth : 16 n_streams : 4 max_leaves : -1 max_samples : 1 max_features : auto n_estimators : 100 max_batch_size : 4096 split_criterion : 0 min_samples_leaf : 1 min_samples_split : 2 min_impurity_decrease : 0
ANN	hidden layer are between 50 and 100. Sigmoid for output layer and selu function for hidden layers are used as activation function. The epoch number is set at 5000.
SVM	The Radial basis kernel is utilized as the kernel function in the SVM model, and the gamma value () is set to 1 and C (cost parameter) is set to 1.

Findings from Algorithms

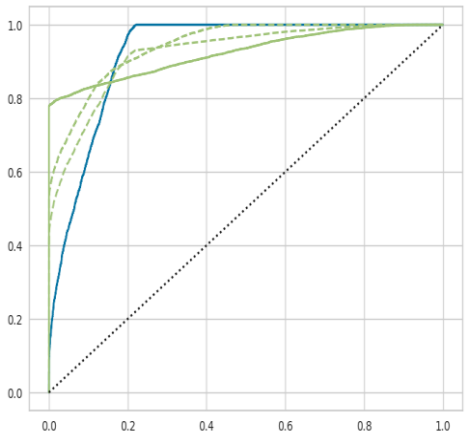
The classification accuracy, precision, recall, and F1 score are all important parameters to consider when evaluating classification models.

The ROC curve in the Figure 4, which is graphical assistance for evaluating the performance classifier and presenting the relationship between true positives and true negatives predicted by the model, is an easy approach to analyze anticipated outcomes (Bradley, 1997). For each approach, all the performance measures have been found and can be indicated in Table 5.

XGBoost



Tuned XGBoost



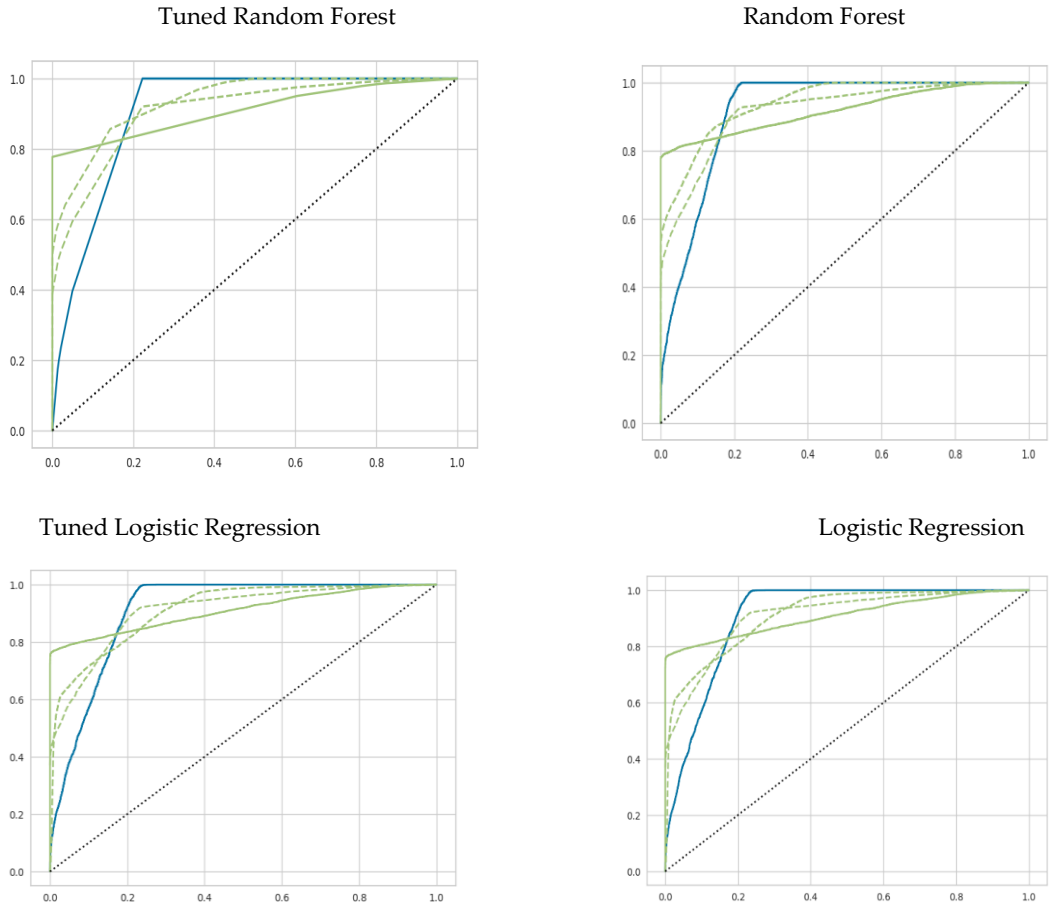
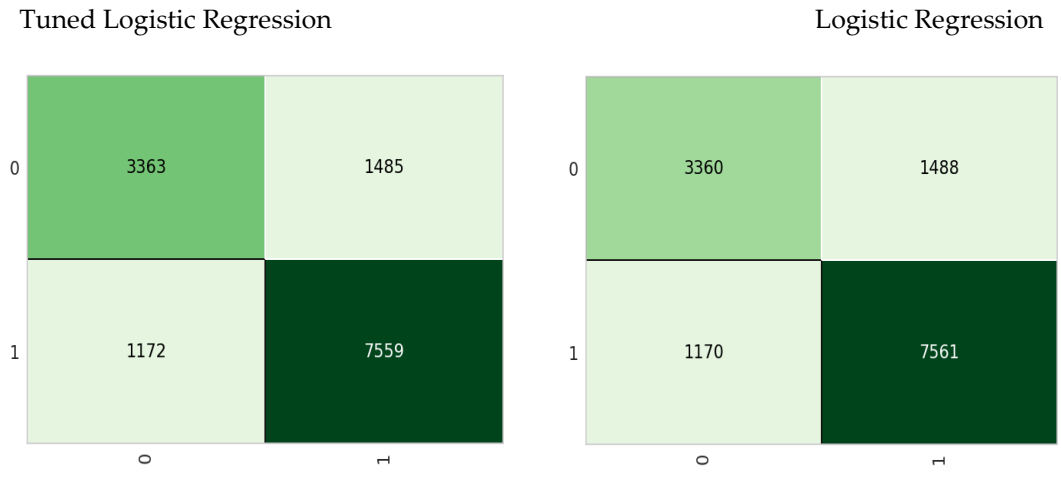


Figure 4. ROC Curves of Models



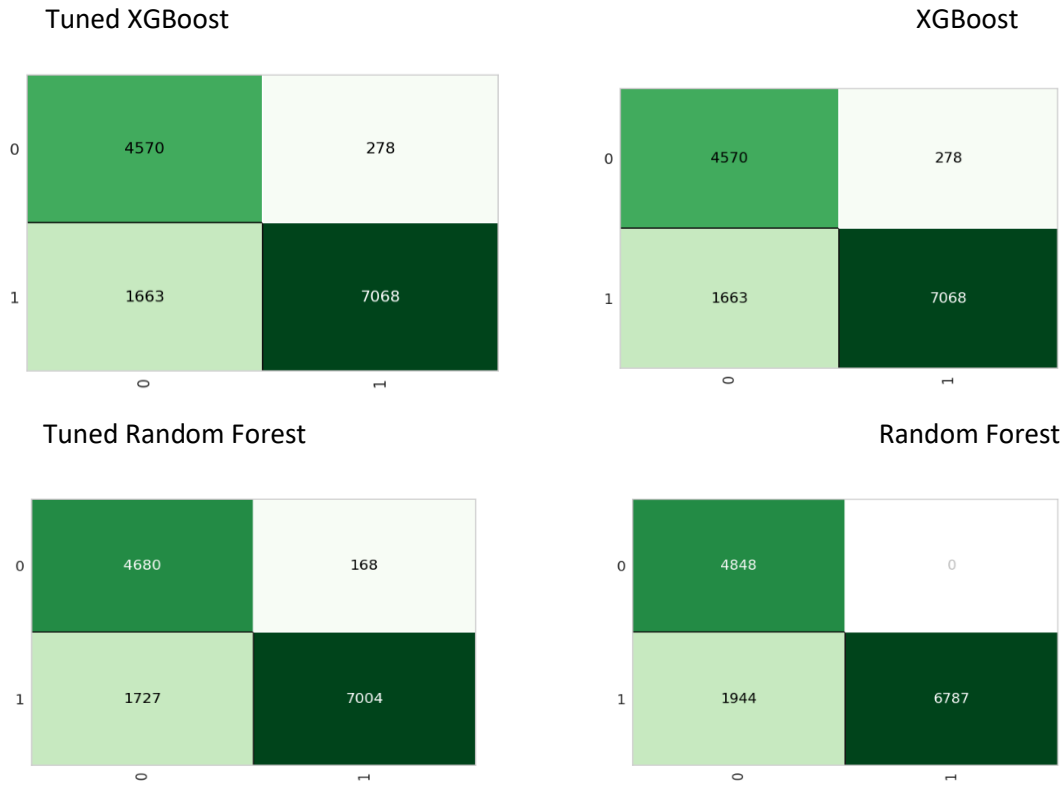


Figure 5. Confusion Matrix of Models

A table called a confusion matrix is used to show how well a classification model performs on test data for which the true values are known.

Table 5. Performance measures for each method are presented

Algorithm		Metrics							
		Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	Gini
tuned_logistic_regression	Train	0.8019	0.9071	0.8651	0.8332	0.8489	0.5618	0.5627	0.8142
	Test	0.8043	0.9093	0.8658	0.8358	0.8505	0.5676	0.5683	0.8186
tuned_xgboost	Train	0.8535	0.9203	0.8097	0.9557	0.8767	0.6994	0.714	0.8406
	Test	0.8571	0.9241	0.8095	0.9622	0.8793	0.7073	0.7232	0.8482
tuned_random_forest	Train	0.8545	0.9041	0.7737	1.0	0.8724	0.7093	0.7413	0.8082
	Test	0.8568	0.9099	0.7773	1.0	0.8747	0.7137	0.7449	0.8198
random_forest	Train	0.8583	0.9151	0.8014	0.9736	0.8791	0.7118	0.7317	0.8302

	Test	0.8604	0.9183	0.8022	0.9766	0.8808	0.7163	0.7367	0.8366
xgboost	Train	0.8535	0.9203	0.8097	0.9557	0.8767	0.6994	0.714	0.8406
	Test	0.8571	0.9241	0.8095	0.9622	0.8793	0.7073	0.7232	0.8482
logistic_regression	Train	0.8019	0.9071	0.865	0.8332	0.8488	0.5617	0.5625	0.8142
	Test	0.8043	0.9093	0.866	0.8356	0.8505	0.5673	0.5681	0.8186
ANN	Train	0.8560	0.8790	099	0.71	0.83	0.71	0.6625	0.8042
	Test	0.8561	0.8785	0.78	0.99	0.87	0.70	0.6681	0.8186
SVM	Train	0.851	0.8763	0.97	0.71	0.82	0.72	0.5525	0.8042
	Test	0.85	0.8785	0.79	0.98	0.87	0.70	0.5681	0.8186

The metrics of the improved models for each algorithm can also accessed by the Table 5. Accuracy, Auc, Recall, Precision, F1, Kappa, MCC and Gini were determined as performance metrics. The metrics selected for each algorithm are calculated separately on the test and train score and are indicated in the Table 5.

Findings from Table 5 demonstrate that the SVM algorithm achieved good results, while RF-based technique improves the accuracy even more. Random Forest has superior accuracy than tuned XGBoost in terms of the area under the ROC curve (AUC). Despite this, the tuned XGBoost approach outperforms the Random Forest method according to AUC, although it has worse accuracy. For all ML algorithms, it can be stated that they achieved good findings. The RF approach produces good values for all measures and has the best overall performance with an accuracy of 86%, as seen in the ROC curve. For tuned RF, the precision value is 1.0, indicating that it operates in accordance with the other parameters, indicating a high rate of true positive items expected by the models over the total positives predicted. Lastly, the f-score, which is a harmonic measure of precision and recall, indicates how well these variables are balanced. Logistic regression has the lowest f-score value, while RF has the greatest value as 88%, based on the same values as the other performance metrics.

Feature Importance

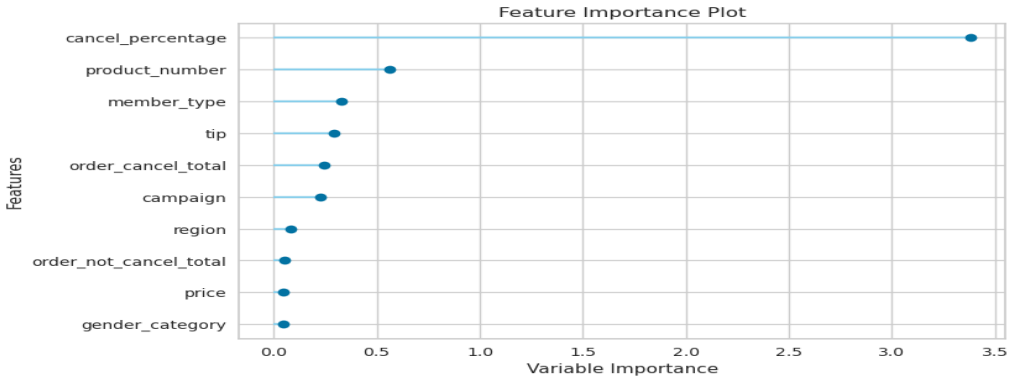


Figure 6. Feature Importance Plot for Logistic Regression

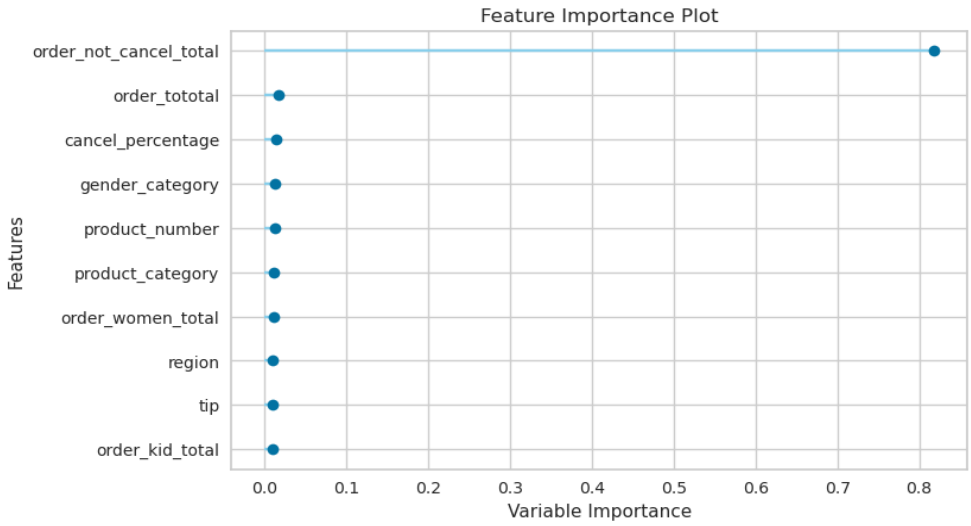


Figure 7. Feature Importance Plot for XGBoost

An attribute with a value between 0 and 100 for each feature, representing how useful the features are when the model is trying to predict the target.

It gives the opportunity to analyze which feature contributes to the accuracy of the model and which features are just noise. If we feel that the features do not add any value, we can discard the features and use the model to hypothesize about new features that we can develop. The attribute importance figure shows that the total number of orders that are not canceled has the most impact on the order or cancellation prediction. Ordered kid total parameter has the least effect on classification.

Discussion and Conclusion

The retail sector is in direct communication with the consumer, it is very important indicator for the entire economy. For this reason, the retail sector has an important place not only in its own value chain, but also in the entire economy. We can observe that it is critical for the general course of the economy to analyze the direction of the developments in the sector and to take measures accordingly. As one of the sectors with the highest number of customers, there is countless data for retail, and it is very important to evaluate this data correctly at the right points. Especially the location of the stores, why some of them are successful or unsuccessful, and a discipline-oriented approach make data analytics even more important in this period. By using data analytics in the retail sector, stores that are at risk of potential closure of companies can be revealed and a special strategy can be applied for these stores. This approach to data analytics helps retailers see the success and failure of their particular stores.

In this study, a model is proposed to predict whether products purchased will be canceled or returned using e-commerce site data. The contribution of this work is to predict order cancellation with different Machine Learning classification algorithms. Accuracy, Auc, Recall, Precision, F1, Kappa, MCC and Gini metrics were used as performance metrics in this modeling. Also, feature importance was obtained with

Logistic Regression and XGBoost. Among the ANN, LR, RF and SVM algorithms used in this study, the RF algorithm achieved the best performance with an accuracy rate of 86%. The RF model with an F1 score of 0.8808 shows that the model prediction gives good results. The total number of non-cancelled orders feature has the most impact on model prediction.

From a management perspective, the findings suggest that a customer's track record is critical to retail businesses and should be viewed as an asset. In this context, the processing of these data using RF algorithms adds significant value to the organization due to the difficulties caused by cancellations and loss of income. On the one hand, accurate cancellation prediction enables businesses to make sound managerial decisions and benefits the industry. These methods allow management to obtain advance information, allowing them to develop acceptable cancellation policies and leverage appropriate pricing tactics, among others.

The previous studies suggested more parameters and dataset and parameter tuning to achieve more accurate model. This study confirms the previous research by adding more attributes, data and parameter tuning.

For the future work, this work is likely to be developed in different areas. Because nowadays it is of great importance whether a customer can cancel or return an order. Predicting the user's behavior and acting accordingly is an action most companies want. In other words, predicting when a customer might cancel or return an order means providing that customer with a better sales experience and ensuring that they don't cancel or return the order. For most companies, this means big profits and a loyal company that understands its customers. In terms of the scope of the study, more data should be used and applied in different sectors.

It is quite difficult to obtain real company data and make it ready for analysis. In this study, it takes a lot of time to integrate the data obtained from different database files and complete the missing values. By adding cargo data and more customer data, the study can be applied to other sectors as well.

Kaynakça / References

Abhirami, K., Pani, A. K., Manohar, M., & Kumar, P. (2021). An Approach for Detecting Frauds in E-Commerce Transactions using Machine Learning Techniques. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 826-831). IEEE.

Ahmed, S. R. (2004). Applications of data mining in retail business. *International Conference on Information Technology: Coding and Computing* (pp. 455-459). Las Vegas: IEEE.

Amari, S. I., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6), 783-789.

Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2019). Predicting customer quality in e-commerce social networks: a machine learning approach. *Review of Managerial Science*, 13(3), 589-603.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54(3)*, 1937-1967.

Breiman, L. (2001). Random forests, *Machine Learning* , 45, 5–32.

Bonaccorso, G. (2017). Machine Learning Algorithms, pp.167-170.

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 30 (7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142.2.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, (pp. 785-794). San Francisco.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Cortes, C. and Vapnik, V. (1995), Support-vector networks, Machine Learning, 20, 273-97.

Dhaliwal, S. S., Nahid, A. A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. *Information, 9(7)*, 149.

Erkent, E. E. (2006). Elektronik Perakendecilik ve Elektronik Alışveriş. *Ege Akademik Bakış*, 10-16.

Fritsch, S., Guenther, F., Wright, M.N., Suling, M., Mueller, S.M. (2019). neuralnet:Training of Neural Networks (Version 1.44.2).

Gong, J. (2021). In-depth Data Mining Method of Network Shared Resources Based on K-means Clustering. *13th International Conference on Measuring Technology and Mechatronics Automation* (pp. 694-698). Beihai: IEEE.

Güllü, K., & Tarhan, M. (2021). Satış sonrası hizmetler ve tüketicilerin yeniden satın alma niyetleri arasındaki ilişkiye yönelik e-perakende sektöründe bir uygulama. *Turkish Journal of Marketing*, 192-205.

Hamed, S., & El-Deeb, S. (2020). Cash on Delivery as a Determinant of E-Commerce Growth in Emerging Markets. *Journal of Global Marketing*, 242-265.

Jiang, P., Zhu, K., Shang, S., Jin, W., Yu, W., Li, S., et al. (2022). Application of Artificial Neural Network in the Baking Process of Salmon. *Journal of Food Quality*, 1-12.

KAYAKUŞ, M., & ÇEVİK, K. K. (2020). Estimation the Number of Visitor of E-Commerce Website by Artificial Neural Networks During Covid19 in Turkey. *Electronic Turkish Studies*, 615-631.

KOÇAL, C. (2012). Uluslararası perakendecilikte rekabet stratejileri ve e-ticaretin önemi. Izmir, Turkey: DEÜ Sosyal Bilimleri Enstitüsü.

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, 150, 113342, pp. 1-16.

Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. R News 2 (3), 18–22. [R News]. Retrieved from https://CRAN.R-project.org/doc/Rnews/.

Liu, C. J., Huang, T. S., Ho, P. T., Huang, J. C., & Hsieh, C. T. (2020). Machine learning-based e-commerce platform repurchase customer prediction model. *Plos one*, 15(12), e0243105, pp. 1-15.

Mauritsius, T., Alatas, S., Binsar, F., Jayadi, R., & Legowo, N. (2020, December). Promo abuse modeling in e-commerce using machine learning approach. *In 2020 8th International Conference on Orange Technology (ICOT)* (pp. 1-6). IEEE.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chih-Chung, C., Chih-Chen, L. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (Version 1.7-2). Retrieved from https://CRAN.R-project.org/package=e1071.

Noor, A., & Islam, M. (2019). Sentiment Analysis for Women's E-commerce Reviews using Machine Learning Algorithms. In 2019 10th International Conference on Computing, *Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Özcan, b., & Turna, c. (2021). karar ağaçlari ile internet alişverişlerinde tüketiciyi etkileyen faktörlerin analizi. *journal of business in the digital age*, 94-105.

Öztemel, E. (2012). Yapay Sinir Ağları (Vol. 3). *İstanbul: Papatya Yayıncılık Eğitim*.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12), 1373-1379.

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep learning for customer churn prediction in e-commerce decision support. *In Business Information Systems* (pp. 3-12).

Rai, S., Gupta, A., Anand, A., Trivedi, A., & Bhadauria, S. (2019). Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods. In 2019 10th International Conference on Computing, *Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Romero Morales, D., Wang, J. (2010). Forecasting cancellation rates for services bookingrevenue management using data mining. *Eur. J. Oper. Res*. 202 (2), 554–562. https://doi.org/10.1016/j.ejor.2009.06.006.

Singh, K., Booma, P. M., & Eaganathan, U. (2020). E-Commerce System for Sale Prediction Using Machine Learning Technique. *In Journal of Physics: Conference Series* (Vol. 1712, No. 1, p. 012042). IOP Publishing.

Szabó, P., & Genge, B. (2020). Efficient conversion prediction in E-Commerce applications with unsupervised learning. In 2020 International Conference on Software, *Telecommunications and Computer Networks (SoftCOM)* (pp. 1-6). IEEE.

Vanneschi, L., Horn, D. M., Castelli, M., & Popovič, A. (2018). An artificial intelligence system for predicting customer default in e-commerce. *Expert Systems with Applications*, 104, 1-21.

Vapnik, V. N. (1995). The nature of statistical learning theory, 2nd ed., *Springer-Verlag New York*, USA, pp. 1-279.

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, 710(1), 120-127.

Yeung, W. L. (2014). Applications of data mining in online retailing: A case for mining prefix-ordered web site navigation paths. *2nd International Conference on Systems and Informatics (ICSAI 2014) Systems and Informatics (ICSAI)* (pp. 943-947). Shanghai: IEEE.

Yin, X., & Tao, X. (2021). Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. *Scientific Programming*, 2021, pp. 1-9.

Zhao, X. (2018). A Study on the Application of Big Data Mining in e-Commerce. *IEEE 4th International Conference on Computer and Communications* (pp. 1867-1871). Chengdu: IEEE.