Advances and Applications in Statistics
© 2022 Pushpa Publishing House, Prayagraj, India
http://www.pphmj.com
http://dx.doi.org/10.17654/0972361722029
Volume 75, 2022, Pages 91-109 **P-ISSN: 0972-3617**

DETERMINING THE BEST ESTIMATION MODEL WITH TREE-BASED MACHINE LEARNING METHODS: IMPLEMENTATION ON CUSTOMER SPENDINGS FOR E-COMMERCE WEBSITES

Mehmet Yalçin and Seda Bağdatli Kalkan

Department of Statistics
Istanbul Commerce University
Turkey

Abstract

Individuals who can easily access Internet have turned to online shopping instead of shopping in physical stores as a result of the development of technology. Individuals' tendency for online shopping has improved the e-commerce sector. Factors such as ability to realize global sales, reduction in physical store expenses, ability to realize 24/7 sales, online and low-cost stock tracking make e-commerce important. Since they use Internet channels in purchasing, the mobility of the customers is also monitored and the behavior of the customers is reflected as data. Thus, the number of studies that predicts the purchasing behavior of customers increases along with prediction models.

Received: January 31, 2022; Accepted: March 9, 2022
2020 Mathematics Subject Classification: 62P05.
Keywords and phrases: tree-based models, e-commerce, machine learning, ensemble learning.

How to cite this article: Mehmet Yalçin and Seda Bağdatli Kalkan, Determining the best estimation model with tree-based machine learning methods: implementation on customer spendings for e-commerce websites, Advances and Applications in Statistics 75 (2022), 91-109. http://dx.doi.org/10.17654/0972361722029

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Published Online: March 28, 2022

Within the scope of the study, it is aimed to determine the model that predicts how much time the customers entering the online purchasing platform will spend in shopping. Tree Based Methods, which are machine learning methods, were used for the established prediction models. In the study, Decision Trees, Random Forests, GBM, LightGBM, XGBoost and CatBoost models were used in addition to the explanation of e-commerce and machine learning methods. By comparing the success performances of the models, it was aimed to find the model that gives the best prediction results. As a result, the model with the best prediction performance was determined. In addition, it was determined which factors were more effective in customers' spendings with the best performing model.

1. Introduction

With the rapid growth of e-commerce platforms, there was a need for companies in the sector to better know and analyze their customers. For example, analyzes such as customer segmentation, customer-specific marketing ideas, correct positioning of the product, advertising strategies on the website, product-price impact analysis, product recommendation systems, association analyzes and product purchase estimates are among the frequently used methods to examine customer behavior in this field.

While the need of e-commerce companies to monitor their customers and take action accordingly is clear, it is not easy to put these strategies into operation. As a result of the rapid increase of data in e-commerce, it is becoming widespread to create strategies with machine learning methods used through software consisting of mathematics and statistics-based algorithms. Machine learning is a set of mathematical and statistical-based algorithms based on the idea that computerized systems learn similar to humans. With machine learning methods, it is possible to process and analyze data, create smart systems, examine customer behavior, follow trends and provide data-based predictions.

In this study, it was worked to create predictions on the spending behaviors of customers with machine learning methods. With tree-based models, which are among the machine learning methods, it is aimed to

determine the model that will find the spendings of the customers with the smallest rate of error. With this model to be determined and the inputs to be used, the spending amount of a new customer can be estimated. In addition, the variables that are effective in the spending amount were determined with the help of the model. As a result of this study, estimating spendings, which was not previously included in the literature, will help companies to take action by making predictions in the future. Thus, the company can offer special campaigns to customers, analyze customer behavior, perform segmentation using effective variables on spendings and data-driven decision-making stages. The data used in the study were obtained from the Google Merchandise Store e-commerce website, a global brand, via Google Analytics, which tracks and records every behavior of customers on the website. This website sells Google products, especially clothing and stationery.

In the study, concepts of machine learning with e-commerce, tree-based models and application steps of models are explained. Tree-based machine learning models, which have recently been used frequently thanks to their speed and performance, have been used over e-commerce data as an application. Random Forests, which are among the ensemble learning models, GBM, LightGBM, Xgboost and Catboost models were applied in addition to Decision Trees. Models were trained on the learning set and compared over prediction scores using error metrics on the test data. The model that gives the most successful estimation results has given the variables that affect customers' spendings. The literature examination indicates that there is no study to determine a model that can predict customer spendings. However, many studies have been conducted in this field, especially in 2021. Behgounia and Zohuri [4] evaluated the effect of developments in machine learning in e-commerce in their study titled "Machine Learning Driven An e-Commerce". Within the scope of the study, the effects of machine learning techniques in following trends, determining promotional prices, reaching new customers, the appearance of websites and taking quick axioms in e-commerce were evaluated. Özdemir [21], in his/her thesis titled "Comparison of Machine Learning Classification Methods and

An Application on e-Commerce", made estimations for customers who made and did not make purchases with Logistic Regression, Naïve Bayes, Support Vector Machines and Random Forests classification methods using e-commerce website data. Karadeniz et al. [15], made classification estimation with ensemble learning models on heart disease risk in their study titled "Ensemble Methods for Heart Disease Prediction". Within the scope of the study, high success rates were achieved with the establishment of ensemble learning models using the Spectf data set. Çelik and Yelkikalan [9] discussed the improvement of warehouse processes, which is significant in e-commerce activities, in their study titled "Application of Machine Learning Methods in Warehouse Management Processes: Azure ML Studio Example". The company data kept in the warehouse processes were processed in Azure MLStudio environment, one of the machine learning platforms on cloud technologies. Prediction and demand forecasting applications were calculated with machine learning methods and artificial neural network models, and their model performances were compared.

2. Concept of e-Commerce

Traditional trade, which is defined as the purchase and sale of various goods and services, has created a new concept with the emergence of globalization and the Internet. This concept, which refers to the conduct of trade in electronic environment, has been described as 'e-commerce'.

The Organization for Economic Development and Cooperation (OECD) defines e-commerce as "An e-commerce transaction is the sale or purchase of goods or services made over Internet with specially designed systems for the purpose of receiving or placing orders. The goods or services are ordered according to this system, but the payment and consequent delivery of the goods and services does not need to be online." An e-commerce transaction can be between corporations, households, individuals, governments, and other public or private organizations (OECD [20]).

The advantages of e-commerce in terms of customers are the facility to use, wide selection range, diversity, saving up cost and time. In addition,

e-commerce facilitates the opportunity for businesses to fulfill the product and service demand of each customer one by one. Businesses have the opportunity to reach their customers, vendors, suppliers and employees all over the world thanks to e-commerce (Pitt et al. [22]).

The increasing volume of e-commerce has also increased its types, diversified and led to the formation of different new areas. Basically, e-commerce is handled under four different types in the literature. B2B, Business to Business e-commerce, B2C, Business to Consumer e-commerce, C2C, Consumer to Consumer e-commerce, B2G, Business to Government e-commerce.

3. Machine Learning

Machine learning is a scientific domain based on the idea that the computerized systems learn similar to humans in order to develop various algorithms, modeling, statistics and techniques (Alpaydın [1]). This machine learning system, consisting of software and statistical algorithms called *machines*, learns from data. Simply put, it is used both to create prediction models based on historical data and to detect patterns, breakdowns and differences within a data set. Data-based learning processes are carried out by using statistical-mathematical models and programming languages. Machine learning is classified in three basic types: Supervised Learning, Unsupervised Learning, and Reinforced Learning. In the study, tree-based machine learning models, which are one of the supervised learning models, are briefly explained.

4. Tree-based Machine Learning Models

Machine learning models generally consist of mathematics- and statistics-based structures. Therefore, it is necessary to provide some assumptions in the formation of some models. However, currently, there are no assumptions in tree-based algorithms, which are among the increasing use of supervised learning models, based on tree structures. It is not affected by outlier observations, missing values and multiple connection problems. Some

models have the possibility to set up a working model with incomplete data. There is no dispersion assumption for dependent and independent variables. They are also used when dependent and independent variables are both categorical and continuous. They are easy to understand and have swift structures in terms of performance.

4.1. Decision trees

Decision trees are a supervised learning model used to transform complex structures in the data set into simple decision structures. Decision trees create models that enable complex and unknown data to be overcome and easily interpreted. They can be used for both classification and regression problems (James et al. [14]).

Among the important advantages of the decision trees model is that even when the data set is very complex, the decision trees model presents the variables affecting the dependent (target) variable and the importance of these variables in the model visually with a simple tree structure (Kayri et al. [16]). The disadvantage of decision trees can be shown as “overfitting”, which is defined as the test set performance being quite low while the performance of the learning set is high (Altaş and Gülpınar [2]). However, restrictions are imposed on the parameters of the model and methods such as pruning are used for solving this problem. In the decision trees method, afforestation, similar to the tree image to be a variable root in the data set, is started, and branching is provided with the points decided by the node. Each observation is classified with the help of nodes. The leaves provide the result and the processes continue until they reach the leaves (Bell [5]). The criteria determining the division, i.e., branching, differ according to the classification and regression trees. Pruning is performed in order to prevent the complexity of the constructed structure, classification error and memorization. The branches in the tree, which do not affect the result and do not contribute to the classification, are taken from the tree by pruning (Silahtaroğlu [24]). Decision trees use multiple algorithms to decide whether to divide a node into two or more sub-nodes. The selection of the algorithm is based on the type of the target variable. The most frequently used algorithms in decision

trees are Entropy, Gini, Classification Error for categorical variables and Least Squares for continuous variables.

According to the type of dependent (target) variable, AID, CHAID, CART, ID3, MARS, exhaustive CHAID, C4.5 - C5.0 Algorithm, SLIQ, SPRINT and QUEST algorithms are used according to the development of Decision Trees by years (Silahtaroğlu [24]).

4.1.1. CART algorithm

The Classification and Regression Trees (CART) method used within the scope of the study is a recursive method used for both regression and classification problems. The aim is to create homogeneous subgroups as wide as possible regarding the dependent (target) variable.

In cases where the dependent variable is categorical, it is defined as Classification Tree (CT); in cases where it is continuous, it is defined as Regression Tree (RT) (Chang and Wang [10]). CART algorithm was developed by Breiman et al. [7]. CART algorithm is an inverted tree view model that does not require the assumptions that parametric regression models have to provide and that predicts the class assignment of the categorical or continuous dependent variable (Temel [26]). The CART method divides each node into two small nodes starting from the root node containing the entire initial data set and creates binary trees. This method works on maximizing the homogeneity on the knot (Antipo and Pokryshevskaya [3]).

4.2. Ensemble learning methods

Ensemble learning methods are mainly based on the idea that the predictive power and performance of multiple models are better than the predictive performance of a model. Basically, the ensemble model combines the estimation results of many different models. Individual models are called “weak learners” because they have a single model. With the ensemble learning method, weak learners can come together to form a “strong learner” with high predictive performance (James et al. [14]). Numerous interesting ensemble learning methods that cluster and classify are used. In the literature,

there are three different ensemble learning algorithms developed on the basis of decision trees and providing stronger prediction results. These are Bagging, Random Forest and Boosting algorithms (James et al. [14]).

The learning set created to train the model in the data set with the bagging method is divided into smaller pieces by the bootstrap (randomly reselected sample) method. These different training sets create different models (trees) for training a single algorithm. The estimation results of the models are obtained by eliminating or averaging with voting. Accordingly, it is aimed that the model trained with a different training set will produce more successful estimation results than the model trained with a single training set (Breiman et al. [7] and James et al. [14]).

Although the Boosting method is very similar to the bagging method, it differs in terms of branching of trees. In the boosting method, trees grow in sequence using information obtained from previously growing trees. Thus, the tree tries not to repeat the guessing mistakes it made while growing up. Bootstrap sampling is not used, instead, each tree takes its roots on the subsequently modified data set in the boosting method (James et al. [14]). In each step in boosting method, the misclassified cases in the previous step are given an increase in weight and unsuccessful estimates are punished. The result estimates are based on the weighting of the estimation results produced iteratively. Although this method is created for classification, it can also be applied for the operation of weak learners (Sutton [25]).

4.2.1. Random forests

The basis of the Random Forests algorithm is based on the idea of creating a forest structure by combining multiple decision trees and eliminating or averaging the estimation results produced by the trees by voting (Breiman [6]).

In the Random Forests algorithm, as in the Bagging method, random sample selection method is used with observations bootstrap for trees. However, differently, a random subset is randomly selected from the variables by the Random Subspace method. This selected subset is used for branching of the respective node. This random selection reduces the

likelihood that the trees in the forest will resemble each other. In this respect, the random forest method is an improved version of the bagging method (Breiman [6]).

In the Random Forests work steps, subspaces are primarily created with bootstrap sampling. However, not all observations are used. Approximately 63% of the observations are used as training data. Observations in the original dataset that were not selected are called *Out of Bag observations*. These Out of Bag observations are used to estimate the error rate. The error rate is often referred to as the Out of Bag error rate and is used to estimate the significance of the feature (variable). The random selection process of samples is repeated many times in each sub-sample, producing a single decision tree, and as a result, a forest of decision trees is formed. Trees grow without pruning and to the greatest extent possible. Individual trees make a classification or prediction decision. The last estimated class of an observation is made by majority vote for classification or by weighted averages for regression (Lewis [18] and Şahinarslan [23]).

One of the most important advantages of creating and running the Random Forests algorithm is that it is quite simple. Recent studies show that regression and classification errors in many of the data structures and problems can be reduced by using the Random Forests method. They are also faster to train because they have fewer parameters. They can work efficiently, requiring minimal data preparation and even in large databases. They are resistant to contradictory observations and over-learning that may be included in the data set (Lewis [18]).

4.2.2. Gradient boosting machines

The Gradient Increase model was developed before the Random Forests algorithm to improve the prediction of decision trees and is based on the boosting method. It can also be used in classification and regression problems (Freidman [12]).

A tree is created for each data set same as the decision trees. Before creating the next tree, the error rates of the trees created are taken into account. Model installation is completed by proceeding in interaction with

other trees. The purpose of the model is to continue to create trees by trying to minimize the error rate. A series of models in the form of a single predictive model is used on model residues (Freidman [12]).

The Gradient Increase Regression model consists of five stages in summary. First of all, a Regression Tree is created. In the second step, the error rate is calculated for the leaf of each tree. The calculated error rates are used as new observation data. The second tree installed aims to minimize the error rates in the second step. Progress is achieved by repeating the second and fourth steps (Şahinarslan [23]). GBM is a generalized and easily adaptable version of the AdaBoost (Freund and Schapire [13]) algorithm based on the idea that weak learners come together to form a strong learner.

4.2.3. Extreme gradient boost (XGBoost)

The Extreme Gradient Boost (XGBoost) algorithm is an improved version of the Gradient Boosting algorithm in terms of performance and operating speed of the results.

It is an algorithm package used to solve problems with a large data set in the field of machine learning. It was developed by Tianqi Chen in 2016 and added to the library called *DMLC* (*Distributed Machine Learning Community*) where open source machine learning algorithms are located (Chen and Guestrin [11]).

The XGBoost algorithm can analyze millions of datasets in a short time and with less processing power and temporary memory (ram) than other algorithms.

Therefore, the intensity of the recently increasing data and the orientation towards this model has increased to reduce the cost of processing power required to analyze the data. XGBoost divides the data into parts and works according to these parts instead of examining every value in the data. Therefore, it performs well according to the Gradient Increase algorithm in terms of speed. In addition, the XGBoost algorithm provides the opportunity to work with missing data.

4.2.4. Light gradient boosting (light gradient boosting, LightGBM)

The LightGBM algorithm is a different gradient boosting model that uses decision trees. It was developed by Microsoft in 2017. It is an algorithm developed to increase the training time performance of the XGBoost algorithm. While other decision trees branch vertically in the analysis process, LightGBM travels to the horizontal branching path over the leaves. Therefore, the training period is shorter in terms of performance (Ke et al. [17]).

Compared to other Gradient Increase models, LightGBM has advantages such as fast processing, processing larger data, having less computer memory (ram), and providing the opportunity to work with missing data. In studies conducted on different data sets, it was concluded that the process of learning the data of the LightGBM model was 20 times faster than other models (Ke et al. [17]).

4.2.5. Category boosting (CatBoost)

It is one of the recently preferred algorithms as a fast and successful Gradient Boosting algorithm developed by Yandex [27] to work with categorical variables. The code blocks that enable automatic modeling of categorical variables are embedded in the algorithm. As for Decision Trees models, digitization is not necessary to include categorical variables in the model. The fact that the data pre-preparation is short, can work with empty values, and enables fast and scalable GPU support brings the model to a good point in terms of performance. Catboost tries to achieve a high prediction rate without establishing deep trees by establishing symmetrical trees. Since it provides randomness with symmetrical trees, it also overcomes the problem of excessive learning.

5. Application

Within the scope of the study, Google Merchandise Store e-commerce site data, which is a global brand, was used. The data has been published on kaggle.com with permission to access by Google. The data were available

between September 2, 2016 and January 04, 2017, and included 903653 observations and 55 variables. Within the scope of this data, it is aimed to determine the model that shows the best predictive performance by using Tree-based machine learning methods. In the study, Python language and the libraries under it were used.

It is not possible to put all the variables in the data set into the model. For this reason, variable engineering was performed. Variables that could not be included in the model such as "Visitor ID, Customer ID, latitude, longitude", categorical variables with completely different characters and 36 variables with completely identical lines were removed from the data set. In addition, the variable containing date information was divided into variables as day and month and added to the data set. Transaction Revenue (Dependent Variable), Pageviews, Hits, Visit Number, Day, Month, New Visits, Channel Grouping, Bounces, Device Browser, Device Category, Operating System, City, Continent, Country, Subcontinent, TrafficSource.isTrueDirect, Campaign, Source, Medium, TrafficSource.adContent 1 dependent variable was included in 20 independent variable models.

Due to the machine learning process, the data set is divided into two as training and testing in order to prevent memorization before the model and to obtain better results in terms of performance. The data set for the study was separated by Holdout method as 70% training and 30% test data set. In the study, Decision Trees, Random Forests, GBM, LightGBM, XGboost and CatBoost models were trained on the training set with the help of Sklearn, one of the Python 3.7 language libraries, and their performance was achieved with model success metrics on the test data. In addition, the K-Fold Cross Validation method was used on the training set. K value was selected as 5 considering the size of the number of rows of the data set. With the K value, the training set was divided into 5 parts and the models were allowed to test themselves 5 times.

Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which are widely used as regression metrics

due to the continuity of the dependent variable, were selected as the error metrics of the created models. The error metrics of the created models on the test data are shown in Table 1 below.

Table 1. Error metrics

	Error Metrics		
Models	**MSE**	**RMSE**	**MAE**
Decision Trees	5,269	2,295	0,321
Random Forests	2,820	1,680	0,316
GBM	2,696	1,641	0,324
LightGBM	2,585	1,608	0,320
XGBoost	2,644	1,627	0,312
CatBoost	2,599	1,612	0,324

The decrease in the error rate due to regression error metrics increases the success of the model. The Average Absolute Error rates of the models seem to be close to each other. However, since the models are run according to the Least Squares, it would be more accurate to compare them through MSE and RMSE metrics. In cases where the average Square Error is close to each other or too large, it is preferable to apply to RMSE. Therefore, when compared according to RMSE, it is seen that the most successful model is LightGBM. Then there is CatBoost with a very small difference. Then comes XGBoost, GBM, Random Forests and Decision Trees, respectively. Tree-based Community Learning models all achieve more successful results than Decision Trees. If a prediction model is to be created with this data set, it is correct to use the LightGBM model with less error value. With the LightGBM model, estimation of the Expense Amount value can be carried out instantly and live using 20 independent variable inputs. Thus, it is possible to estimate customers' spendings by using Google Analytics data in the most realistic way.

Within the scope of the study, parameter optimization is made on this model since the LightGBM model is the best prediction algorithm. Parameter optimization is the process of making the parameters (number of trees, tree

depth, learning speed, etc.) carried by the models more suitable for the data set in accordance with the working principles. Models do not allow to optimize all the parameters they contain. Therefore, the important parameters discussed in general for the improvement of the LightGBM model are shown in Table 2.

Table 2. Parameters and their definitions (LightGBM [19])

Parameters	Definitions
Objective	Parameter for defining the type of problem. Classification according to the type of problem can be defined as regression or multiple regression.
Metric	It is the parameter with the evaluation criterion used to evaluate the success of the model. It is determined according to the Objective parameter. (MSE, RMSE, Accuracy, etc.)
Max_depth	It is used to determine the depth of the tree to be installed. If its value is increased too much, it causes overfitting, and if its value is determined as less than it should be, it causes underfitting.
Num_leaves	It is the parameter used to control the complexity of the tree. In order to avoid overfitting, the num_leaves value must be less than $2^{(max_depth)}$.
Min_data_in_sheet	It is the minimum number of observations on each leaf of the tree. It is one of the important parameters used to prevent excessive learning.
Num_titrations	It controls the number of boosting tours to be performed. Since LightGBM uses decision trees as learners, this parameter can also be considered as the “number of trees”.
Learning_rate	It is a value given between 0-1 to scale the established trees. As the learning rate value decreases, the predictive performance increases, but the duration of education and the likelihood of excessive learning also increase.
Colsample_by_tree	It is the ratio of randomly selected features (variables) used to train each tree.

It is not possible to estimate each of the parameters. The interaction and multiplication of the parameters to be used within themselves also affect the model. Therefore, the Grid Search Algorithm is used. With this algorithm, separate models are established on the training data with all combinations for

the parameters and values that are defined as dictionaries and desired to be tested in the model. The parameter values that are successful according to the determined metric are found with the help of the algorithm (Burkov [8]). Within the scope of the study, the parameters tested and found on the Grid Search algorithm and LightGBM model are shown in Table 3.

Table 3. Tested and found values of parameters

Parameters	Tested Values	Found Values
Colsample_bytree	[0.4, 0.5, 1]	[1]
Learning_rate	[0.01, 0.1, 0, 15, 0.5]	[0.15]
Num_leaves	[5, 10, 50, 100]	[64]
Max_depth	[5, 10, 50]	[10]
Num_titrations	[40, 100, 200, 1000]	[1000]

As a result of parameter optimization, when the model is established again, the RMSE value of the LightGBM model is 1.608, while the RMSE value is 1.562 as a result of the optimization of the parameter values. Thus, the model was made more suitable for the data set by lowering the error value. This improvement in error estimation value also increased the application performance of the model.

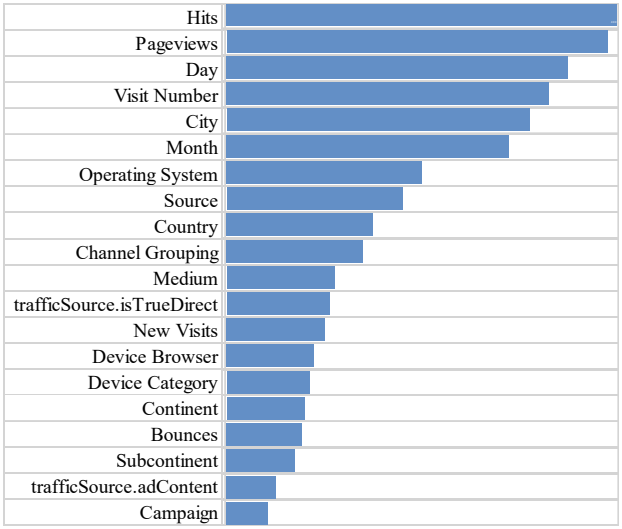


Figure 1. Ranking of variables affecting expenditure.

Within the scope of the study, the factors that affect the dependent variable the most were also reached as a result of the models established among the randomly selected variables through the LightGBM model, which showed the best predictive performance. Figure 1 shows the order of the variables that affect expenditure. The variable that was most effective on spending was the Number of Clicks variable. In other words, the fact that the customer spends a lot of time on the website and clicks on the page has a significant effect on his/her spending.

6. Result

With the development of e-commerce, companies have provided many opportunities to obtain information about their customers. Companies developed different models using these data and tried to increase both their profitability and customer profitability by applying these models. It is very important to use machine learning methods that have been increasingly trendy recently on the basis of processing data and drawing meaningful conclusions. The fact that companies use methods such as machine learning will provide companies with very vital advantages such as time, automation and speed.

Within the scope of the study, the model that provides the estimation of customers' spendings with the least error value by using tree-based machine learning methods by using Google Merchandise Store e-commerce website data was determined. In addition, with the help of the established model, the variables that affect the spendings were determined. Considering all model results, it was determined that ensemble learning models yielded more successful results than Decision Trees.

It was determined that the model that showed the best predictive performance was the LightGBM model. With this model, the most important variables affecting expenditure were determined. Accordingly, the most effective variable on expenditure was the number of Clicks variable. The Number of Views variable, which is associated with the Number of Clicks

variable, was determined as the second in importance. Then, the Day variable was important in 3rd order. Visiting Frequency, City and Month variables also ranked high in importance level.

As a result, within the framework of these important variables, actions can be taken that will have an impact on the behavior of customers. Companies will be able to develop actions that will affect their customers' purchases such as special campaigns, city-specific opportunities and discounted days campaigns and increase their spending rates even more. Thus, the company will be able to increase its market share in the e-commerce field by continuing data-based decision-making processes. They will have the opportunity to take a quick position within their own processes by making a quick decision with the machine learning models created.

This study is derived from the first author's master's thesis.

References

[1] E. Alpaydın, Introduction to Machine Learning, USA: Massachusetts Institute of Technology Press, 2010.

[2] D. Altaş and V. Gülpınar, Karar Ağaçları ve Yapay Sinir Ağlarının Sınıflandırma Performanslarının Karşılaştırılması: Avrupa Birliği Örneği, Trakya Üniversitesi Sosyal Bilimler Dergisi 14(1) (2012), 1-22.

[3] E. Antipov and E. Pokryshevskaya, Applying CHAID for logistic regression diagnostics and classification accuracy improvement, Journal of Targeting, Measurement and Analysis for Marketing 18 (2010), 109-117.

[4] F. Behgounia and B. Zohuri, Machine learning driven an e-commerce, International Journal of Computer Science and Information Security (2020), 61-70.

[5] J. Bell, Machine Learning: Hands-on for Developers and Technical Professional, Somerset, John Wiley and Sons, Incorporated, USA, 2014.

[6] L. Breiman, Random forests, Machine Learning 45 (2001), 5-32.

[7] L. Breiman, J. Freidman, R. Olshen and C. Stone, Classification and Regression Trees, Chapman and Hall, New York, 1993.

[8] A. Burkov, The Hundred-page Machine Learning Book, Canada, 2019.

[9] C. Çelik and N. Yelkikalan, Makine Öğrenme Yöntemlerinin Depo Yönetim Süreçlerinde Uygulanması: Azure ML Studio Örneği, Journal of Yasar University, 16(62) (2021), 659-682.

[10] L. Chang and H. Wang, Analysis of Traffic Injury: An Application of Non-parametric Classification Tree Techniques, Accident Analysis and Prevention 38(5) (2006), 1019-1027.

[11] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, The 22nd ACM SIGKDD International Conference, 2016.

[12] J. Freidman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29(5) (2001), 1189-1232.

[13] Y. Freund and R. Schapire, Experiments with a new boosting algorithm, In Machine Learning, Proceedings of the Thirteenth International Conference, Bari, İtaly, 1996, pp. 148-156.

[14] G. James, D. Witten, T. Hastie and R. Tibshirani, Introduction to Statistical Learning, Springer, 2013.

[15] T. Karadeniz, H. Maraş and G. Tokdemir, Ensemble methods for heart disease prediction, New Generation Computing (2021), 569-581.

[16] M. Kayri, F. Elkonca, H. Şevgin and G. Ceyhan, Ortaokul Öğrencilerinin Fen ve Teknoloji Dersine Yönelik Tutumlarının CHAID Analizi ile İncelenmesi, Eğitim Bilimleri Araştırması Dergisi 4(1) (2014), 301-316.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma and T. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, NIPS, 2017.

[18] D. Lewis, Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners, CreateSpace Independent Publishing Platform, 2017.

[19] LightGBM, LightGBM Library, data: 01.05.2021, 2021. lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html

[20] OECD, OECD Guide to Measuring the Information Society, OECD Publishing, 2011. https://doi.org/10.1787/9789264113541-en

[21] R. Özdemir, Makine öğrenmesindeki sınıflandırma yöntemlerinin karşılaştırılması ve e-ticaret üzerinde bir uygulama, İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul, 2021.

[22] F. Pitt, P. Berthon, R. Watson and G. Zinkhan, The Internet and the birth of real consumer power, Business Horizons 45(4) (2002), 7-14.

[23] F. Şahinarslan, Makine Öğrenmesi Algoritmaları ile Nüfus Tahmini: Türkiye Örneği, İstanbul Teknik Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, İstanbul, 2019.

[24] G. Silahtaroğlu, Veri Madenciliği (Kavram ve Algoritmaları, Papatya Yayıncılık Eğitim, İstanbul, 2013.

[25] D. Sutton, Classification and Regression Trees, Bagging and Boosting, Handbook of Statistics 24 (2005), 303-329.

[26] G. Temel, Sınıflama ve Regresyon Ağaçları, Mersin: Mersin Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2004.

[27] Yandex, CatBoost, 2017. http://catboost.ai/en.docs/concepts/educational-materials-papers