Utilizing Logistic Regression for Analyzing Customer Behavior in an E-Retail Company

HAKAN ALPARSLAN[1], SAFIYE TURGAY[1], RECEP YILMAZ[2]
[1]Department of Industrial Engineering
Faculty of Engineering
Sakarya University
54187, Esentepe Campus Serdivan-Sakarya
TURKEY
[2]Department of Business
Faculty of Business
Sakarya University
54187, Esentepe Campus Serdivan-Sakarya
TURKEY

Abstract: - The e-retail sector is growing day by day and the competitive environment is getting harder. Businesses have to compete with their competitors in order to survive. In parallel with the increasing internet penetration, the trade volume in E-Retail sites is also increasing therefore the data generated on these sites is enormous. Understanding these data with traditional analysis methods is difficult due to the size problem mentioned. Difficult to understand data causes loss of time, money and customers. In recent years, machine-learning algorithms have been frequently used to analyse these large-sized data and to use them in decision-making. This study aimed to perform predictive analysis for the product recommendation system established by using logistic regression, which is a supervised machine-learning algorithm. In addition, the binary classification algorithm preferred to predict whether customers make a purchase or not. As a result, the accuracy degree of the model was 79.73%. This study has the potential to affect the understanding of customers, ensuring customer satisfaction, increasing profit and market share, and contributes to a sustainable business purpose.

Key Words: Predictive Analysis, Logistic Regression, E-Retailing, Machine Learning, Binary Classification, Customer Behavior, Big Data

Received: April 11, 2023. Revised: February 14, 2024. Accepted: March 18, 2024. Published: May 14, 2024.

1 Introduction

Today, physical markets or market places digitized at a high speed. This digitalization is very important for businesses. In this case, the customers have been seen as a marketplace; generate huge amounts of data on the internet. These generated data can help businesses make decisions in their supply chain management processes.

E-Commerce includes transactions in many categories that take place on the internet. Retailing is one of these categories and is one of the leading areas of E-Commerce. Retailing defined as any activity that acts as a bridge between the producer/supplier and the end consumer and related to the presentation of goods and services to the end consumer. Electronic retailing emerged as a result of the transfer classical retailing activities to virtual and online platforms. With the growth in the e-retail sector and the increasing interest, the competitive environment is getting more and more challenging. Businesses aim to increase their sales by transforming their physical stores into digital/virtual stores.

In 2021, the ratio of e-commerce to general trade was 17.7%. Worldwide E-Commerce volume is around 4.9 Trillion USD in 2021. As a result, it is expected that this volume will reach 7.4 Trillion USD in 2025(in Fig.1).

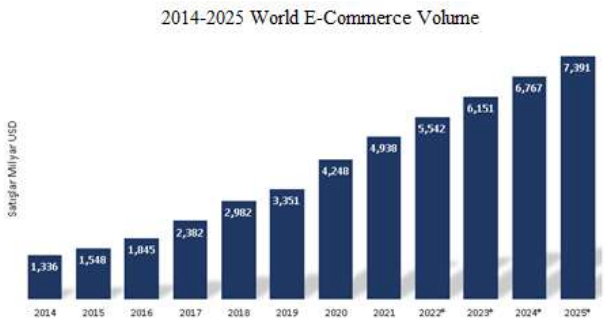

Figure 1. World E-Commerce Volume (Trillion USD) (https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/)

E-Retail sites aim to increase their sales in which the data created by the customers on the web pages is of great importance. Analysis of this data is useful for improving customer satisfaction and experience. Traditional analysis methods are insufficient due to the large size of the data generated on e-commerce sites. This situation causes customer dissatisfaction and loss of money and time for businesses. The large data size in machine learning increases the success of the trained model [1], [2], [3], [4], [5]. For this reason, machine-learning algorithms used in e-commerce web pages can be successful in understanding customer behaviors, making appropriate product recommendations or making sales forecasts.

This study aimed to measure the success of the machine-learning model based on customer behaviors by processing the raw data collected from a real E-Retail site with logistic regression. It will benefit the business in terms of the effectiveness of the established system and model by measuring the success of the established product recommendation system and machine-learning model.

The remaining of the paper organized as follows. The literature review presented in section 2. Section 3 presents the used techniques and. Section 4 applies and discusses the approach to logistic regression method to e-retail company. Finally, section 5 analysis the applied method. and finally last section summarizes some conclusions and discusses potential extensions of the research.

2 Literature Survey

Machine learning uses large datasets with minimal human intervention and identify models that can support decision making. Machine learning develops computer algorithms that can imitate human intelligence, and it is useful for identifying and analyzing markets[6], [7], [8], [9]. Some of the reserachers preferred the logistic regression method for static representation learning to incremental learning of embedding dynamic networks[10], [11], [12], [13]. Some studies used the logistic regression method to obtain strong rules by addressing the multi-label classification problem [14], [15], [16]. Some researchers examined the logistic regression model for software error estimation and in distinguishing between faulty and error-free modules [17], [18], [19], [20]. Some of them examined the logistic regression model to develop the air quality index [21], [22]. Machine learning is divided into three subcategories depending on the learning paths [23,24].

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Logistic regression is a classification algorithm that falls under the category of supervised learning. In the learning process, predictions and decisions are made using the learned data. Linear Regression, Support Vector Machines, Neural Networks, Decision Trees, NaiveBayes and Nearest Neighbor algorithms are examples of supervised machine learning. By calculating the predictive value of the dependent variable with logistic regression binary classification, it classifies in the range of 1-Yes, 0- (in Fig.2)[6].

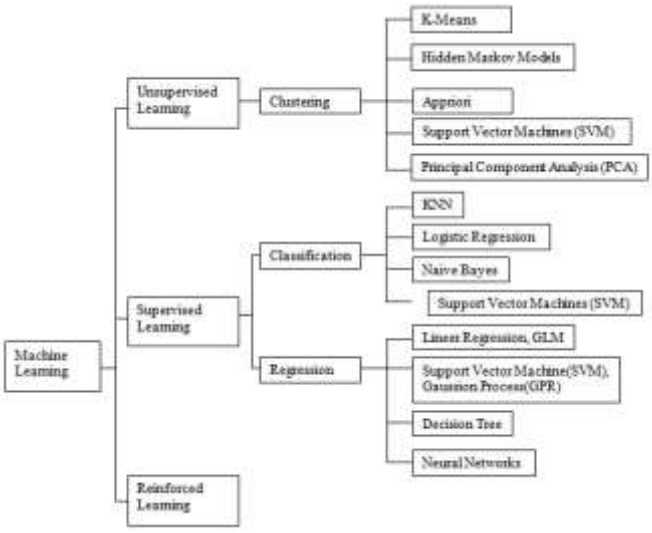

Figure 2. Machine Learning Categories by Learning Paths

The model for forecast probabilities is expressed as the natural logarithm of the odds ratio:

$$In\frac{P(Y)}{1-P(Y)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (1)$$

X_1, X_2, … X_k are the prediction variables, $\beta_0,\beta_1,\ldots,\beta_k$ are the regression (model) coefficients and $\beta 0$ is the intersection point. The odds ratio is defined as the ratio of the probability of occurrence to the probability of not happening. The purpose of

logistic regression is to estimate the βk+1 parameters.

The maximum likelihood requires finding the parameter set with the greatest probability of the observed data. The regression coefficients show the degree of relationship between each independent variable and the outcome. The purpose of Logistic Regression is to accurately predict the outcome category for individual cases using the best model [25], [26], [27], [28]. The goal of model includes all predictive variables which consists of the predicting the response variable. Logistic Regression calculates the probability of success based on the probability of failure.
As a result of the study, it was the job of mathematicians and consultants in the past, but with the effect of increasing technological developments in changing times, more senior managers and organizations are trying to use these techniques for long-term planning a five-day sales forecast was made.
It is important to analyze the target audience correctly and to give the customer what he wants with the support of analysis tools to compete in e-commerce is to understand the customer. At this point, forecasting or forecasting analysis is a savior for e-commerce sites. The process of predicting the visitor's next move in real time by understanding customer experiences and behaviors called predictive predictive analysis (in Fig.3).

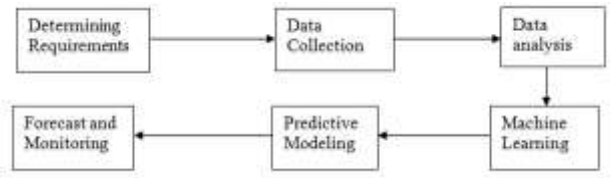

Figure 3. Predictive Analysis Process Steps

Today, due to the increasing customer interest in e-retail, the size of the competition among e-retail sites is growing. It is very difficult for companies to understand user behaviors and trends in an environment where competition is intense. The data generated on these sites is very large. Since traditional analysis methods are weak, different approaches needed. By using data science and machine learning, the business wants to understand the customers, to establish a product recommendation system and to predict real-time customer movements by analyzing the customer behavior data on the E-Retail site to increase sales by making personalized advertising campaigns. Finally, the success of the newly established system should be measure.

In order to solve the above-mentioned problem, the solution reached by following the steps below. Firstly, the structure of the data obtained from the e-commerce site should be understand. Determining the data structure with the Pandas and Numpy libraries in Python and queries for the data structure is the initial step of the study.
So as to implement Machine Learning, the data must first be brought into an appropriate format with Pandas. In the second solution phase, it is aimed to conduct customer behavior research and establish a product recommendation system with the help of Pandas and Seaborn libraries, based on the data structures obtained during the determination of the structure of the dataset and the preparation of the data. In the third stage, it is aimed to measure the success of the model in predicting customer behaviors by using logistic regression.

3 Logistic Regression Model

Logistic regression is a multivariate analysis method that studies the relationship between two variables y and a series of influencing factors(Huang et al.) .

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \tag{2}$$

$$p = \frac{1}{1 - e^{-p}} \tag{3}$$

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

θ_0 is the constant term of the coefficient vector to be estimated, θ_i is the coefficient of the first i argument x_i in expression 1:

$$\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{5}$$

For $h_\theta(x)$, the value of a Function gives the following meaning: if represents the probability of taking 1 from the variability y. Thus, for the input argument vectors, the probability that the x results are Class 1 and Class o are:

$$h_\theta(x) \quad p\left(y=1 | x;\theta\right) = h_\theta(x) \tag{6}$$

$$p\left(y=0|x;\theta\right)=1-h_\theta\left(x\right) \tag{7}$$

The flowchart of the logistic regression model in this study is shown in Figure 2.

-Construct probability density likelihood functions of m samples:

$$L(\theta)=\prod_{i=1}^{m}p\left(y_i|x_i;\theta\right)=\prod_{i=1}^{m}\left(h_\theta\left(x_i\right)^{y_i}\left(h_\theta\left(x_i\right)^{1-y_i}\right)\right) \tag{8}$$

It's Log likelihood Function:

$$In(L(\theta))=\sum_{i=1}^{m}(y_i In\,(h_\theta x_i))+\left(1-y_i\right)In\,(1-h_\theta x_i) \tag{9}$$

$$h_\theta=\frac{1}{1-e^{-T_x}} \tag{10}$$

By estimating the y and x values of m test samples, the estimated values of the coefficient vectors that maximize the function are calculated θ. The probability of the sample y of 1 is estimated, in which x is the independent variable vector of the sample y.

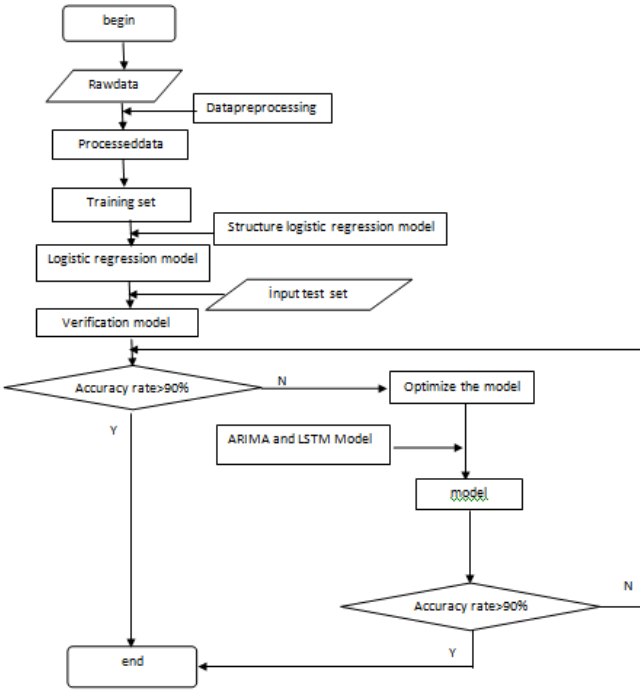

Figure 4. Flow chart of the logistic regression algorithm

4 Case Study

In this section, the logistic regression method applied to customer behavior data of e-retail company. Process steps, briefly, determining of the dataset, customer behavior research then measuring the prediction success of the model in 3 steps.

Step 1 Determining and Editing the Structure of the Dataset

The dataset consists of four files: the event data, the product attributes divided into two files, and the category tree. The data was collected from a real e-commerce site. The data is raw, so no content conversion has been applied. The e-commerce web page in question has a site traffic of 1,407,580 unique visitors, with 2,756,101 events in total.

Figure 5. Number of Unique Visitors and Total Number of Visits

It contains the first 5 lines in the events.csv file that contains the event data in Figure 6. There are 5 variables: Timestamp, Visitor ID, Event, Item ID, and Transaction ID. The transaction ID variable takes a value if the transaction has been made, otherwise it does not take any value.

	timestamp	visitorid	event	itemid	transactionid
0	2015-06-02 05:02:12	257597	view	355908	NaN
1	2015-06-02 05:50:14	992329	view	248676	NaN
2	2015-06-02 05:13:19	111016	view	318965	NaN
3	2015-06-02 05:12:35	483717	view	253185	NaN
4	2015-06-02 05:02:17	951259	view	367447	NaN

Figure 6. Timestamps Converted to Familiar Date Format

Figure 7. Top 5 of User Events (Gestures)

The situation after the conversion of Unix/Epoch observation units taken by the timestamp variable in Figure 6 to the familiar date and time format

performed with the help of the codes below shown in Figure 7 again.

```
[37] times =[]
     for i in events['timestamp']:
         times.append(datetime.datetime.fromtimestamp(i//1000.0))

[38] events['timestamp'] = times
```

Figure 8. The Process of Converting Time stamps to Familiar Format

The event variable, i.e. events such as views, add-to-cart, and transactions, are interactions collected over a 4.5 month period. A visitor can perform three types of events: "view", "add to cart" or "action"(Fig.8 and Fig.9).

```
[51] print(events.shape)
     print(events['event'].unique())

     (2756101, 5)
     ['view' 'addtocart' 'transaction']
```

Figure 9. Demonstrating the Structure of the Events Dataset and the Actions Customers Can Take

There are events involving 2,664,312 views, 69,332 adding to carts and 22,457 transactions by 1,407,580 unique visitors(in Fig.10).

```
[57] print(events['event'].value_counts());

view           2664312
addtocart        69332
transaction      22457
Name: event, dtype: int64
```

Figure 10. Total Views, Add to Cart and Purchases.

When the events are analyzed over the pie chart, it is seen that 96.7% of the events are viewing, 2.5% are adding to cart and 0.8% are purchasing (Fig.11). In total, 11,719 visitors made purchases. The ratio of visitors purchasing products to the total number of unique visitors is 0.83%. Category IDs describe how different products relate to each other. For example, a product with CategoryID 1016 is a child of 213 ParentID. ItemIDs are a child of CategoryIDs. In Figure 12, the products in that category and their features are listed in the query made on the products with the CategoryID of 570.

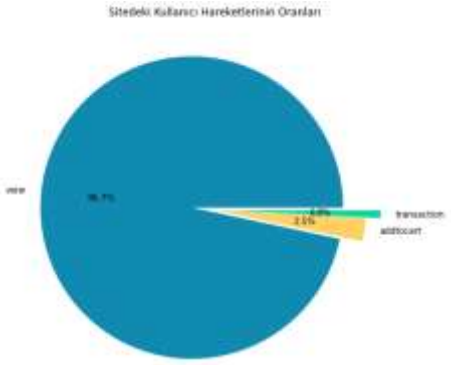

Figure 11. Pie Chart of Total Views, Add to Cart, and Purchases

Figure 12. Query for Number of Unique Items, Total Observation Units, and Number of Features

The data file containing the product properties contains 20275902 rows describing 417053 unique products. Because the characteristics of products can change over time (for example, their price changes over time), the data has the corresponding timestamp along with the behavior data. If the property of a product is constant during the observed period, there is only one snapshot in the file. "Property" contains the category ID and suitability of the product, some values have been hashed in the dataset for privacy reasons. "Value" represents the price. If its value is 0, it indicates that it is out of stock (in Fig.13).

Step 2: Customer Behavior Research and Suggestion System

It was determined that only 11719 out of 1407580 customers made a purchase. However, we do not have precise data on whether a customer makes more than one purchase. Therefore, we need to assume that 11,719 purchases are unique and at least once. Based on our assumption above, those who bought at least one product among 1,407,580 unique visitors removed with the following procedure. It known that the remaining 1,395,861 unique visitors made a definite viewing process.

Figure 13. Number of Users Viewing But Not Making a Purchase

In Figure 14, 10 of the visitors who made a purchase observed. For example, "505565" ID numbers are known to make a purchase at least once. When the user's actions are queried with the visitor ID, it is seen that he displays the product with the ID number "243566", then adds it to the cart and buys it

Figure 14. The Movements of the First 10 Users Who Made a Purchase, Visitor ID number 505565

While users are viewing a product, it aimed to show a list of products purchased by previous visitors. For this, a list of the products purchased by these users created by making use of the list of users who made purchases, which we created in the previous stages. In short, a list of all purchased products has been created. The top five examples in the list of items sold are shown below (Fig. 15).

Figure 15. Creating the List of Products Sold and Top 5 Examples

In Figure 16, it aimed to create a new suggestion list in order to get rid of the repeated products and to remove the repeating products. Thus, a list of products recommended depending on the display of the product obtained. For example, customers who view the product with ID 105792 will be offered the products in the list specified as a result of the query.

Figure 16. Deletion of Duplicate Products, Creation of Suggestion List and Example of Product List to be Displayed to Users Viewing the Product with ID 105792

The data analysis made at the beginning of the application indicates that only 11719 out of 1407580 unique visitors purchased at least one product during the time the data was collected. A data frame was created in order to investigate the process of understanding user behaviors, which we obtained in our studies on the dataset during the application, according to the user ID, the number of products viewed, the total number of views, and whether they bought something. Users who made a purchase at least once subtracted from the number of unique users to obtain the number of users who made a viewing.

Figure 17. Creating the Master Dataset

Step 3: Measuring the Prediction Success of the Model

27821 random data were taken from the list of viewers. The reason for this is to catch the ratio of 70 to 30 in the training and test model. We aimed to create the 30% slice with 11719 purchase data and 70% with 27821 random viewing data. Then the first 27821 rows in the data set taken. 11719 rows of data that made the purchase and 27821 randomly obtained viewing data were combined and assigned to the list called main dataset. Finally, the main dataset list obtained shuffled again (Fig.17). In order to understand the relationship in the new dataset obtained, the following chart was created with product views, total views, total purchase data. As can be seen from the graph, the relationship between purchasing and viewing is linear. So the higher the number of views, the more likely the user is to buy. Due to the linearity of the relationship, a logistic regression model established in order to predict future user purchasing behavior. Firstly, the purchase, visitor ID and purchase totals removed from the features section and assigned to X. On the other hand, purchasing status assigned to Y as the target. The Logistic Regression estimation model used to predict the test features (in Fig.18).

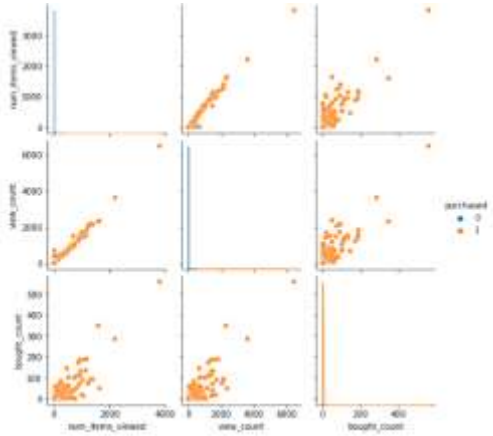

Figure 18. Linear Relationship Between Purchasing and Viewing

Figure 19. Creating the Model and Measuring the Accuracy Ratio

As a result, the rate of correctly estimating the users making purchases of the model accurated is 79.73% (in Fig. 19).

5 Analysis

In this section, the analysis of the model carried out by performing the Receiver Operating Characteristic (ROC) analysis. According to the model, we trained in the application part, the result shows that the accuracy rate of the model we built is 79.73%.

In the following process, with the help of predict_proba(), not a single prediction is created, but a prediction for each of the test observations. False Positive Ratio (FalsePositiveRatio, fpr) and True Positive Ratio (True PositiveRatio, tpr) are stored in vector form as they will be used in the graph. For the same purpose, Area Under the Curve (AUC) is stored in the variable named roc_auc(in Fig.20).

Figure 20. Construction of the ROC Curve.

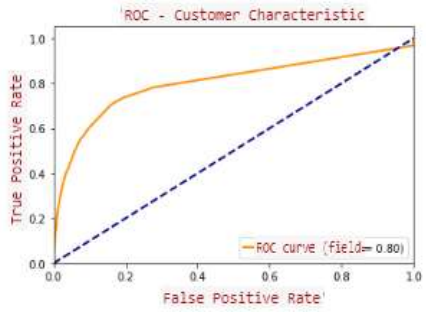

Figure 21. ROC(Receiver Transaction Characteristic) Line Chart

The chart above shows the accuracy of our binary classifier (Logistic Regression). The closer the orange curve is to the top left of the graph, the better the accuracy (in Fig.20 and Fig.21).

6 Conclusion

The e-retail industry is growing very rapidly. Parallel to this growth, the competition in the sector is increasing day by day. In the study, customers were classified, a product recommendation system was established and it was aimed to predict whether customers would make a purchase in real time. When we examine the largest companies in the sector in question, it shown that customer satisfaction is given the top priority. At the same time, customer satisfaction breeds loyalty. Again, when we look at the biggest companies in the sector, people want to feel valuable there rather than seeing an e-commerce site as just a place for shopping. For this reason, the importance of understanding your customer is quite large from a social point of view in Table 1.

The goal of companies to be sustainable is a difficult situation to achieve. Also, it is important for sustainability to give importance to customer satisfaction and understand the customer by keeping the customer in the foreground. Since the aim of the study is to understand the customer, it is thought that it contributes to the achievement of the sustainability goal of the companies.

Table 1. Current Situation, Work Done and New Situation Chart

Current Situation	Process	New Situation
The data generated on the e-Retail site has not been analyzed. As a result, it is possible to lose money and customers.	Customer segmentation was made by examining the Product, Customer and Customer Movements.	The firm can now understand its customers and interpret their behavior.
Lack of Product Recommendation System	A product recommendation system has been established by utilizing product data purchased by other customers.	With the established Product Suggestion System, product sales are expected to increase.
Customer movements cannot be predicted. According to the established system, it is required to predict whether the customer will make a purchase or not.	Predictive analysis was performed in the machine learning model with Logistic Regression, which is a binary classification algorithm.	Since the success of the model was found to be 79.73%, it is now possible to obtain information about whether 80 out of every 100 customers will make a purchase. It has been a useful study for advertisements and campaigns.

As a result of the study, the success of the machine learning model established in real time to predict whether the customer will make a purchase was 79.73%. The success rate of the model is quite high. The fact that the ROC Curve in the graph created as a result of the ROC Analysis is close to the y-axis shows that the model is a successful model. The established recommendation system and machine learning model are quite simple and applicable. Today, machine learning is actively used in the e-commerce sector. In fact, machine learning has served as an accelerator in the development of e-commerce for years. For this reason, the application of this study in the e-commerce sector contributes to the development of different approaches. Logistic Regression requires more data than other regression algorithms. Currently, the data produced on these sites reaches gigantic proportions. In other words, Foresight Analysis with Logistic Regression is a viable machine learning method for E-Commerce sites.

References:

[1] K. Barbé. Kurylyak, Y., Lamonaca, F., "Logistic ordinal regression for the calibration of oscillometric blood pressure monitors", Biomedical Signal Processing and Control 11, 89–96, 2014.

[2] E.Y. Boateng, D.A.Abaye, "A Review of the LogisticRegression Model with Emphasis on Medical Research", Journal of Data Analysis and Information Processing, 7, 190-207, 2019.

[3] P.Bielak, K.Tagowski, M.Falkiewicz, T.Kajdanowica, N.V. Chawla, "FILDNE: A Framework for Incremental Learning of Dynamic Networks Embeddings", Knowledge-Based Systems 236, 107453, 2022.

[4] D.Borges, M.C.V.Nascimento, "COVID-19 ICU demand forecasting: A two-stage Prophet-

LSTM Approach", Applied Soft Computing, 125,109181, 2022.

[5] B.Chen, X.Chen, B.Li, Z.He, H.Cao, G. Cai, "Reliability estimation for cutting tools based on logistic regression model using vibration signals", Mechanical Systems and Signal Processing, 25, 2526–2537, 2011.

[6] D.R.Cox, "The Regression Analysis of BinarySequences", Journal of the Royal Statistical Society, Series B (Methodological). 215-242, 1958.

[7] D. Das, P. Dutta, "Product return management through promotional offers: The role of consumers' loss aversion", Int. J. Production Economics 251, 108520, 2022.

[8] A.J. Dobson, An Introduction to Genaralized Linear Models, Chapman and Hall, 2001.

[9] X.Du, S.Chen, Z. Liu, J.Wang, "Multiple userids identification with deep learning", Expert Systems With Applications, 207, 117924, 2022.

[10] Z.Diang, H.Chen, L.Zhou, Z.Wang, "A forecasting system for deterministic and uncertain prediction of air pollution data", Expert Systems With Applications, 208, 118123, 2022.

[11] X.Hu, H. Luo, M.Guo, J.Wang, "Ecological technology evaluation model and its application based on Logistic Regression", Ecological Indicators, 136, 108641, 2022

[12] T. Huang, B.Li, D.Shen, J.Cao, B.Mao, "Analysis of the grain loss in harvest based on logistic regression", Procedia Computer Science, Vol.122, pp. 698-705, 2017.

[13] D.Jain, S.Makkar, L.Jindal, M.Gupta, "UncoveringEmployeeJobSatisfaction Using Machine Learning: A Case Study of Om Logistics Ltd. In: Gupta", D., Khanna, A., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. Advances in IntelligentSystemsand Computing, vol 1166. Springer, Singapore, 2020.

[14] J. Khaleeq, M. Amanullah , A.T. Abdulrahman, E.H. Hafez, M.M. Abd El-Raouf, "Influence diagnostics in Log-Logistic regression model with censored data", Alexandria Engineering Journal, 61, 2230–2241, 2022.

[15] J. Liu, S.Zhang, H.Fan, "A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network", Expert Systems With Applications 195, 116624, 2022.

[16] W. Liu, H. Fan, M. Xia, M. Xia," A focal-aware cost-sensitive boosted tree for imbalanced credit scoring", Expert Systems With Applications, 208, 118158, 2022.

[17] W. Liu, H.Fan, M.Xia, "Credit scoring based on tree-enhanced gradient boosting decision trees", Expert Systems With Applications, 189, 116034, 2022.

[18] X.Luo, X.Kong, T.Nie," Spline based survival model for credit risk modeling", European Journal of Operational Research 253, 869–879, 2016.

[19] K.Matuszelański, K.Kopczewska, "Customer Churn in Retail E-Commerce Business: Spatialand Machine Learning Approach", *Journal of TheoreticalandApplied Electronic Commerce Research*,17(1), 165-198, 2022.

[20] P. Manchala, M.Bisi, "Diversity based imbalance learning approach for software fault prediction using machine learning models", Applied Soft Computing 124, 109069, 2022.

[21] Y.Qiao, Q.Lan, Z.Zhou, C.Ma, "Privacy-preserving credit evaluation system based on blockchain", Expert Systems With Applications, 188, 115989, 2022.

[22] A.Robles-Velasco, P.Cortes, J.Munuzur, L. Onieva, "Prediction of pipe failures in water supply networks using logistic regression and support vector classification", Reliability Engineering and System Safety, 196, 106754, 2020.

[23] D.B.Sassi, A.Frini, M. Chaieb, W.B.A.Karaa, " A rough set-based Competitive Intelligence approach for anticipating competitor's action", Expert Systems With Applications 204, 117523, 2022.

[24] D. Simić, N. Ve Bačanin Džakula, "The Ethics of Machine Learning". InSinteza 2019-International Scientific Conference on Information Technologyand Data RelatedResearch,SingidunumUniversity, Serbia, 478-484, 2019.

[25] A. Strzelecka, A.Kurdys-Kujawska, D. Zawadzka, "Application of logistic regression models to assess household financial decisions regarding debt", 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science 00, 000–000, 2020.

[26] Y.Wu, Q.Zhang, Y.Hu, K.Sun-Woo, X.Zhang, H.Zhu, L. Jie, S.Y. Li," Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems", Future Generation Computer Systems, 129, 1–12, 2022.

[27] X.Yu, S.Guo, J.Guo, X.Huang, "An extended support vector machine forecasting framework for customer churn in e-commerce", Expert Systems with Applications, 38, 1425–1430, 2011.

[28] Y.Zou, C. Chun-An, "A combinatorial optimization approach for multi-label associative classification", Knowledge-Based Systems 240 , 108088, 2022.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

H.Alparslan, S.Turgay, R.Yılmaz – investigation,
H.Alparslan, R.Yılmaz - validation and
S.Turgay writing & editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.