ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

Predict a UK Customer's Likelihood of Making an Online-purchase Based on the Logistic Regression Model

Jiayi Gu

University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom

Abstract: *With the rapid development of e-commerce, the scale of e-commerce consumers and online transaction volume has increased rapidly. Before e-commerce, shopping was entirely based on in-person interactions. Conventional methods of purchasing and selling were followed due to the relatively low requirements of consumers. However, due to the rapid development of Internet technology, more and more consumers choose purchase online. This study proposes to review Logistic Regression model on customers' purchase likelihood analysis and prediction. Also, depending on the different advantages and disadvantages of Logistic Regression model, this study aims at building predict models which is suitable for the Wiggle Ltd. Finally, it will screen out the relationships between variables for recommendation and further evaluation.*

Keywords: *Logistic Regression, Predict Models, Purchase Likelihood, Customer Behaviours*

1. Introduction

The development of the internet has brought many benefits to our society. Firstly, the rapid development of the internet has significantly changed the way people obtain information. Secondly, the internet has revolutionized mobile communication technology. Thirdly, the emergence of the internet has greatly promoted the development of E-commerce. For example, the internet provides a sales platform for retailers. For retailers, on one hand, online sale effectively saves the cost. On the other hand, online sales expand the scope of sales for retailers.

From the perspective of consumption behaviour, the active degree of online shopping users continues to increase, and the proportion of online shopping consumption in daily expenditure significantly increases. Online shopping has become a consumption habit and a popular way of shopping.

With the popularisation of smart devices and the development of networks, a rich set of big data has been produced. McKinsey, a world-renowned consulting company, first proposed the concept of the era of big data. They believe that data will be used in every industry and business function area in the future, and will become a critical decision-making factor. With the beginning of the big data era, people will change their way of living, working and thinking gradually, as they begin to dig, analyse and use the massive data.

Therefore, finding out meaningful and valuable information from large-scale network user behaviour data has become the focus of Internet people and major Internet companies. Data mining and data analysis have also become hot research directions in recent years. Among the massive user behaviour information, the most commercially significant one is the analysis based on the shopping network user behaviour information, that is, the analysis based on the e-commerce website. According to the mining of the information of e-commerce websites customers, the information mined from the research can be applied to the marketing strategies of e-commerce, to provide a higher possibility for them to obtain profits.

2. Literature Review

2.1. The Value of Building a Predictive Model

Although e-commerce has made people's lives more convenient, the massive amounts of information also make it difficult for users to find the desired products because of the large variety of product

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

categories and various marketing methods. According to Savrul in e-commerce, people have to find the product from the huge online inventory represented by sellers, and thus it can be difficult for the customers to find the product they need. Websites incorporate products searching systems to save the time of customers and give them a better shopping experience. Product selection is a very important factor that affects the efficiency of e-commerce as supported by Sadagopan in 2008. E-commerce websites should be designed in such a way that users can easily find the product without wasting time on irrelevant or undesired products. According to Sivapalan, an improved customer journey on the website impacts the online e-commerce businesses positively.

Therefore, predicting customer preferences accurately and recommending suitable products for them through technology has become an important research topic in the current commercial data mining field, which is of great significance for improving the experience of online shopping and increasing e-commerce revenue.

2.2. The Data Set Used for Building a Predictive Model

Purchase forecast is of great significance to the improvement of economic efficiency of e-commerce websites and has been widely studied. As mentioned above, traditional recommendation technologies, such as collaborative filtering have limitations. Therefore, many researchers use machine learning and feature engineering methods to train customer preference purchase models based on big data of a large user sample. In e-commerce, the model has to deal with large data sets that need reliability and accuracy. Any errors in the model can affect the final decision and the overall functionality of the business is affected, as shown by Manning. Lehman stated that Logistic regression is an important method, which is used to predict customer behaviour by using customer logs.

These logs consist of various factors which depend on the choices of the owners of the e-commerce business. Burns and Burns predicted that machine learning requires time and data to improve its decision-making ability. Therefore, the availability of sufficient data is necessary to produce results with a higher confidence level. The latest customer behaviour prediction models improve the machine learning by focusing on the logs of customers, which consist of every action taken by the customer during all the sessions as shown by Chang & Lin. However, Grégoire, predicted that conventional methodologies for prediction of customer behaviour are not capable of analysing these logs because of the complexities and amount of data. Therefore, it is preferable to use the latest customer behaviour prediction models to improve the accuracy and reputation of the e-commerce businesses in the market. Moshrefjavadi found that it makes online shopping easier for customers when businesses use the latest customer behaviour prediction models. One of the best examples of such an online store is Amazon.

After collating the above research, it can be seen that customer behaviour log data to predict purchases has a better effect.

2.3. Logistic Regression

Logistic Regression can be utilized for different arrangement issues, for example, spam location. It can be used to predict if a given client will buy a specific item or will they agitate another contender, regardless of whether the client will click on a given notice interface or not. Logistic Regression is one of the most basic and regularly utilized Machine Learning calculations for two-class characterization. It is not difficult to execute and can be utilized as the benchmark for any double order issue. This is an issue which arises due to multiple requests made by the customer. It produces unnecessary errors which affects the end results.

Following are the advantages and disadvantages of logistic regression:

It is one of the least difficult AI calculations and is not difficult to execute yet gives great preparation effectiveness. Additionally, because of these reasons, preparing a model with this calculation doesn't need high calculation power.

The anticipated boundaries (prepared loads) give induction about the significance of each component. The heading of affiliation, for example positive or negative, is likewise given. Therefore, we can utilize logistic regression relapse to discover the connection between the highlights. Refreshing of the model is necessary so that it represents the latest information. It is different from decision trees and SVM as refreshing is not permitted by calculations done within the model.

Logistic Regression yields all around aligned probabilities alongside characterization results. This

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

has a bit of leeway over models that only give the last arrangement as results. On the off chance that a preparation model has a 95% likelihood for a class, and another has a 55% likelihood for a similar class, we get a deduction about which preparing models are more exact for the detailed issue.

Logistic regression endeavours to anticipate results dependent on a lot of autonomous factors, however on the off chance that specialists incorporate inappropriate free factors, the model will have next to zero prescient worth. Logistic regression functions admirably for foreseeing all-out results like affirmation or dismissal at a specific school. It can likewise foresee multinomial results, similar to affirmation, dismissal, or hold up the list. Nonetheless, Logistic regression can't foresee persistent results. Logistic regression endeavours to foresee results dependent on a lot of free factors, however, logit models are helpless against pomposity. That is, the models can seem to have more prescient force than they really do because of inspecting inclination.

2.4. Evaluation Criteria for Predictive Models

The particular objectives of the rustic oral wellbeing system ought to be viewed as while deciding the measurements for program evaluation. Provincial oral wellbeing programs gather subjective and quantitative data on program cycles, yields, and results. A 10-overlap cross-approval test outfit is utilized to exhibit every measurement since this is the most probable situation where one will utilize diverse calculation assessment measurements. A confusion matrix is a strategy for summing up the exhibition related to a calculation of characterization. Gathering precision alone can be deceiving if the information has a conflicting number of observations in each class or in case the information has numerous classes in the dataset. There are a few important reasons that explain why one should use a confusion matrix for evaluation. It is this breakdown that beats the obstruction of using game plan precision alone.

The full text of the article must be typeset in single column.

3. Methodology

3.1. Build Logistic Regression

3.1.1. Linear Regression and Logistic Regression.

Logistic regression is an extension of the linear regression model for classification problems. It is widely used to model the probabilities for classification problems with two or more possible outcomes.

There are some problems when using linear regression model for classification tasks.

(1) A linear model does not output probabilities.

(2) The outputs of linear model do not always lie in the range (0, 1).

(3) Linear model does not extend to classification tasks with multiple classes.

In order to address these problems, the logistic regression scheme introduces a non-linear function (sigmoid function) instead of fitting a straight line in linear regression scheme. Sigmoid function is defined as:

$$\sigma(x) = 1/(1 + \exp(-x)) \quad (1)$$

It squeezes the outputs between zero and one and has a characteristic S-shaped curve like the following Figure 1

3.1.2. Theory of Logistic Regression.

The initial letters of words should be capitalized. Words like “is”, “or”, “then”, etc. should not be capitalized unless they are the first word of the subtitle. No formulas or special characters of any form or language are allowed in the subtitle.

(1) Logistic Distribution

According to Kisssell, the logistic distribution is a continuous distribution function, which is widely used in logistic regression, logit models and forward neural networks.

. The cumulative distribution function and density function of logistic distribution are shown as equations (2)-(3) respectively. Compared to normal distribution, logistic distribution has heavier tails and higher kurtosis.

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

$$F(x) = P(X \leq x) = \frac{1}{1 + exp\,-(x-\mu)\,/\gamma} \tag{2}$$

$$f(x) = F'(x) = \frac{\frac{exp\,-(x-\mu)}{\gamma}}{\gamma(1 + exp(-(x-\mu)\,/\gamma))^2} \tag{3}$$

Where $\mu,\ \gamma > 0$ denotes the location and scale parameters respectively.

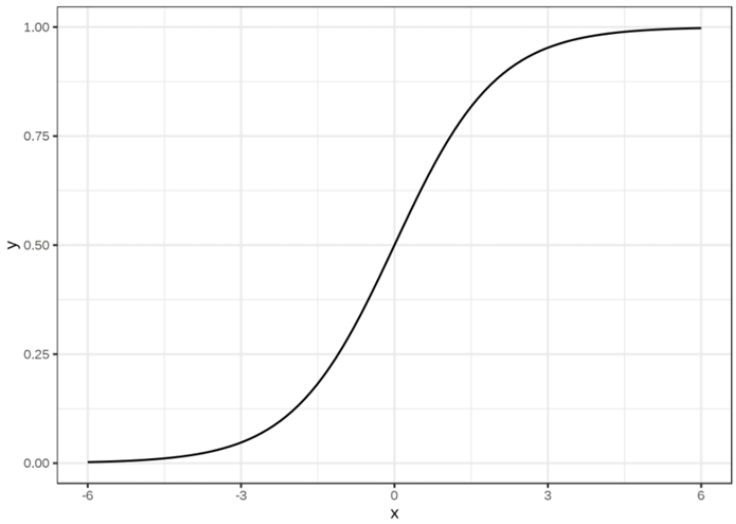

Figure 1: A characteristic S-shaped curve.

(2) Binomial Logistic Regression Model

A binomial logistic regression model is widely used in predicting the probability that an observation falls into one of two categories, based on one or more independent variables which can be either continuous or categorical.

The binomial logistic regression model can be described by the following equations (4)-(5):

$$P(Y = 0|x) = \frac{1}{1 + exp(w \cdot x + b)} \tag{4}$$

$$P(Y = 1|x) = \frac{exp(w \cdot x + b)}{1 + exp(w \cdot x + b)} \tag{5}$$

Where $x = (x^{(1)}, x^{(2)}, \ldots x^{(n)}) \in \boldsymbol{R}^n$ denotes the feature vector, and $Y \in \{\mathbf{0}, \mathbf{1}\}$ denotes two categories that outputs of the model will fall into. $w \in \boldsymbol{R}^n$ and $b \in \boldsymbol{R}^n$ are weights parameters which will be learnt in the training process.

Let $W = \left(w^{(1)}, w^{(2)}, \ldots, w^{(n)}, b\right)^T$, $X = (x^{(1)}, x^{(2)}, \ldots x^{(n)}, 1)$, then the binomial logistic model can be rewritten as:

$$P(Y = 0|x) = \frac{1}{1 + exp(W \cdot X)} \tag{6}$$

$$P(Y = 1|x) = \frac{exp(W \cdot X)}{1 + exp(W \cdot X)} \tag{7}$$

In statistics, odds describe the relative probabilities, i.e., the ratio of the probability that the event will happen to the probability that the event will not. Thus, if one event will happen with a probability of p, then the odds of this event is calculated as $\frac{p}{1-p}$, and the log odds of the event is:

$$logit(p) = log\,\frac{p}{1-p} \tag{8}$$

The log odds of the binomial logistic regression model can be defined as:

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

$$log\ \frac{P(Y=1|x)}{1-P(Y=1|x)} = W \cdot X \tag{9}$$

Based on the equation (9), the log odds of binomial logistic regression model can be described as a linear function of feature vector X.

(3) Multi-nominal Logistic Regression Model

Compared to the binomial logistic regression model, the multi-nominal logistic regression model is an extension for Multi-class classification problem. Assuming that the target variable $Y = \{1,2,\dots,K\}$, where $K \geq 2$, then the multi-nominal logistic regression model can be described as:

$$P(Y=k|x) = \frac{\exp(W_k \cdot X)}{1+\sum_{k=1}^{K-1} exp(W_k \cdot X)}\ , \qquad k = 1,2,\dots,K-1 \tag{10}$$

$$P(Y=K|x) = \frac{1}{1+\sum_{k=1}^{K-1} exp(W_k \cdot X)} \tag{11}$$

Where $X \in \boldsymbol{R}^{n+1}$, and $W_k \in \boldsymbol{R}^{n+1}$.

3.1.3. Interpretation of Logistic Regression Model

The score function in logistic regression can be defined as:

$$\frac{1}{N}\sum_{i=1}^{N} -y_i * \ln\big(\sigma(\hat{w}x_i + \hat{c})\big) - (1-y_i) * \ln(1-\sigma(\hat{w}x_i + \hat{c})) \tag{12}$$

Where x_i denotes the feature vector, y_i denotes the labels for classification. $\hat{w}$ and $\hat{c}$ are the weight vector, which are learnt in logistic regression by minimizing the cost function (12).

In the case with two classes, y_i can either be 0 or 1. Considering these two cases separately, in the case that $y_i = 0$, the cost function becomes$\frac{1}{N}\sum_{i=1}^{N} -(1-y_i) * \ln(1-\sigma(\hat{w}x_i + \hat{c}))$, since the right term is zeroed out. It is clear that $0 \leq 1-\sigma(\hat{w}x_i + \hat{c}) \leq 1$, then $-\infty < \ln\big(1-\sigma(\hat{w}x_i + \hat{c})\big) \leq 0$ or equivalently $0 \leq -\ln\big(1-\sigma(\hat{w}x_i + \hat{c})\big) < \infty$. Thus, the optimal solution for minimizing the cost function is that $\sigma(\hat{w}x_i + \hat{c}) = 1$. In the logistic regression scheme, the $1-\sigma(\hat{w}x_i + \hat{c})$ can be treated as the probability that the $n-$ sample point will be not classified to belong to some class $z \in Z$, based on the feature vector x_i. If it really is the case this sample point does not belong to the class $z \in Z$, it will be expected that $\sigma(\hat{w}x_i + \hat{c})$ is close to 1.

In the case that $y_i = 1$, the cost function becomes$\frac{1}{N}\sum_{i=1}^{N} -y_i * \ln\big(\sigma(\hat{w}x_i + \hat{c})\big)$, since the right term is zeroed out. It is clear that $0 \leq \sigma(\hat{w}x_i + \hat{c}) \leq 1$, then $-\infty < \ln\big(\sigma(\hat{w}x_i + \hat{c})\big) \leq 0$ or equivalently $0 \leq -\ln\big(\sigma(\hat{w}x_i + \hat{c})\big) < \infty$. Thus, the optimal solution for minimizing the cost function is that $\sigma(\hat{w}x_i + \hat{c}) = 1$. In the logistic regression scheme, the $\sigma(\hat{w}x_i + \hat{c})$ is treated as the probability that the $n-$th sample point will be positively classified to belong to some class $z \in Z$, based on the feature vector x_i. If it really is the case this sample point belongs to the class $z \in Z$, it will be expected that $\sigma(\hat{w}x_i + \hat{c})$ is close to 1.

3.2. Confusion Matrix

Confusion matrix is a very common and widely adopted evaluation criterion. The principle of the confusion matrix is to compare and calculate according to the corresponding positions and classifications of the positive and negative examples of the actual situation and the test situation. In this way, the confusion matrix can be used to better evaluate the data results.

In this study, each column of the confusion matrix represents the predicted category, and the total number of each column represents the number of data predicted to be that category. Each row represents the actual category of the data, and the total amount of data in each row represents the number of data actual to be that category.

The value in each column represents the number of real data predicted to be of that class. In the visualization results of the confusion matrix, the darker the color is, the better the prediction result of the model is. Therefore, the confusion matrix can clearly reflect the part where the true value and the predicted value are consistent with each other, and it can also reflect the part that does not coincide with the predicted value.

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

4. Results

After 10 variables were imported into the logistic regression model, the final results are shown in table 1-2 and figure 2-3.

Table 1: Accuracy of logistic regression model in TraintSet.

Accuracy:51.2%		
Accuracy and accuracy of each category		
Category	precision	recall
1	0.6463	0.8974
2	0.2921	0.2514
3	0.3617	0.2111
4	0.3024	0.1541

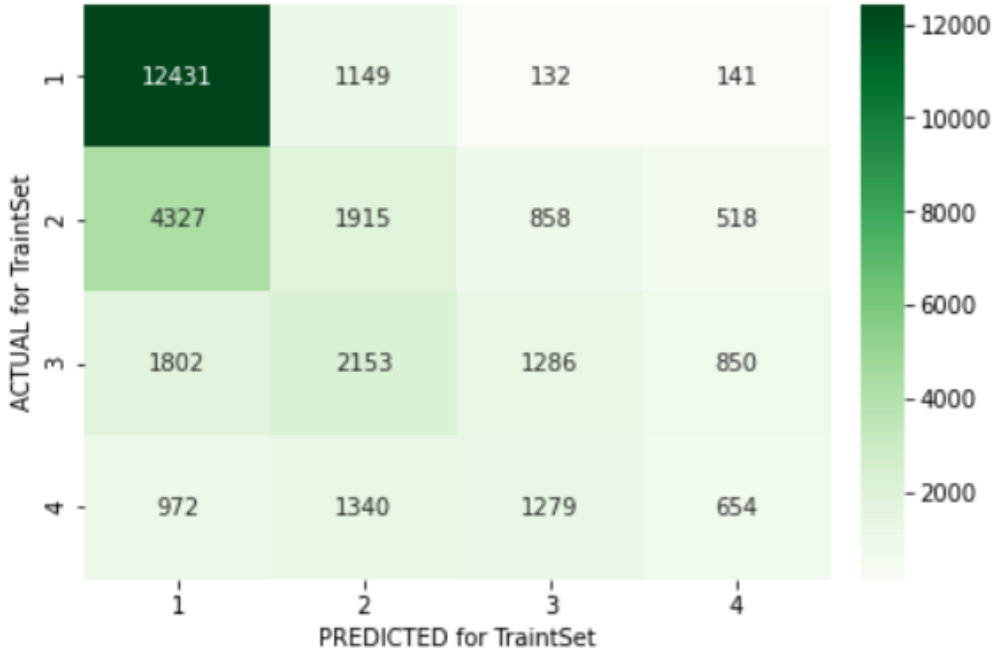

Figure 2: Confusion matrix of predicted for TraintSet in logistic regression model.

As can be viewed in Table 1, the accuracy of logistic regression TraintSet model is 51.2%. The prediction performance of the model is good, which means classification prediction with logistic regression is stable.

It also can be seen in Figure 2, which is the confusion matrix of predicted for TraintSet in logistic regression model. The confusion matrix can clearly reflect the part where the true value and the predicted value are consistent with each other, and it can also reflect the part that does not coincide with the predicted value. The darker the colour, the better the prediction performance. This logistic regression model trains 555,678 values of session quality dim. The first category performs best, which is 12,431, but the other three categories perform unsatisfactorily.

Table 2: Accuracy of logistic regression model in TestSet.

Accuracy:51.339999999999996%		
Accuracy and accuracy of each category		
Category	precision	recall
1	0.6434	0.9052
2	0.2907	0.2524
3	0.3533	0.2067
4	0.2905	0.1469

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

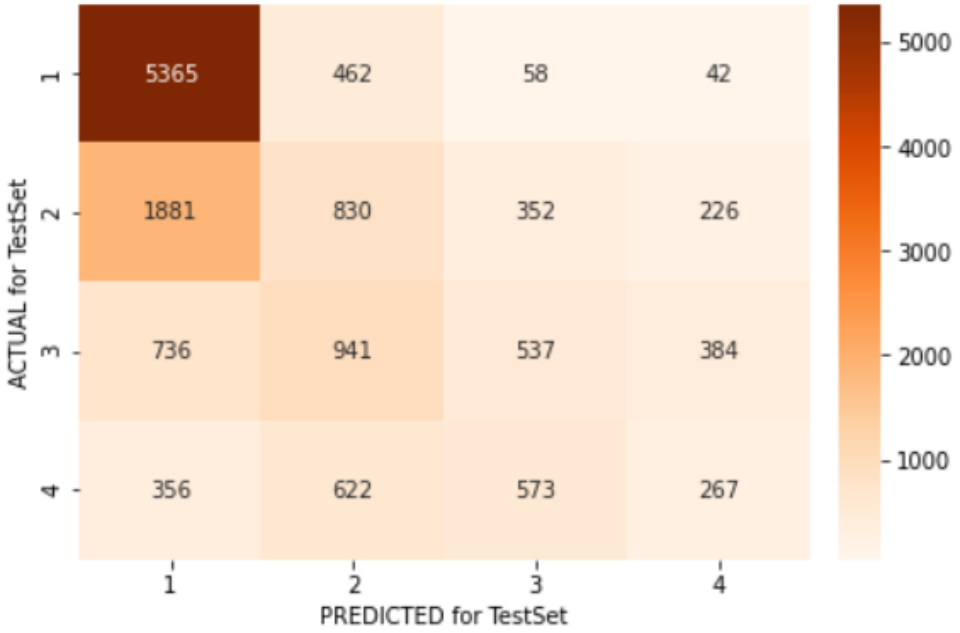

Figure 3: Confusion matrix of predicted for TestSet in logistic regression model.

It is clear in Table 2 and Figure 3, the accuracy of logistic regression model in TestSet is 51.339999999999996%, similar to the result of prediction set.

5. Conclusion

5.1. Research Significance

With the rapid development of information technology, we have entered the era of big data on the Internet, and online shopping for young people has become a fashion. E-commerce platforms gather a large number of consumer purchase data. Many e-commerce platforms use big data such as Hadoop and Spark and cloud computing technologies to extract useful information from high-dimensional massive data, analyze and model online user consumption behavior, and predict consumer demand. Its main purpose lies in three aspects. First of all, it is used for product display, personalized recommendation and accurate advertising. Secondly, it is used to influence users' purchasing decisions through commodity prices, sales volume, and commodity evaluations. Finally, it is used to support the decision-making of countries, regions and enterprises. They can understand consumers' consumption behaviors based on the analysis results of purchase data, and adjust the industrial structure in time, so that the economy can develop in a coordinated, stable and sustainable manner and benefit the society.

5.2. Difficulties and Limitations in Analysis

Despite of theoretical basis, a quantity of difficulties and limitations would happen during the study.

(1) Missing values and data outliers: Although data preprocessing overcomes the issues of incomplete data, excessive missing values or outliers still lead to deviations between predicted outcomes and real ones. This study does not conduct data outliers track because customer data are collected from google analytics and data size is not large. The author cannot check which values are abnormal in customers' behavior field.

(2) Data inadequacy: Because the data set is collected from the Google analytics, it may not be adequate and comprehensive. For example, by correlation analysis, it is clear that Customers who don't use mobile phones to shop are more likely to buy goods than customers who use mobile phones to shop. However, do not provide further data on customers who do not use mobile phones that what they use for shopping. If provide the data sufficiently detailed, the outcome could be different.

ISSN 2522-3488 Vol. 6, Issue 1: 20-27, DOI: 10.25236/IJNDES.2022.060104

5.3. Further Work

This research has built a model and conducted test process. Although the model on the training set is similar to the model on the Test set, it requires further work to do:

(1) It needs larger customer to be scored. For a useful and long-term model, before being into operation, scoring is a necessity to ensure that the model performs good.

(2) Listed variables may not cover all aspects of customers. To predict the likelihood of customers, the conditions of the customers themselves are also important, for example, the economic level of the customers' needs to be considered.

References

[1] Shen, B. and Chan, H.L., 2017. Forecast information sharing for managing supply chains in the big data era: Recent development and future research. Asia-Pacific Journal of Operational Research, 34(01), p.1740001.

[2] Savrul, M., Incekara, A., & Sener, S. (2014). The Potential of E-commerce for SMEs in a Globalizing Business Environment. Procedia - Social and Behavioral Sciences. https://doi.org/10.1016/j.sbspro.2014.09.005

[3] Shen, Y., Jiang, Y., Liu, W., & Liu, Y. (2015). Multi-class AdaBoost ELM. https://doi.org/10.1007/978-3-319-14066-7_18

[4] Sadagopan, S. (2008). E-commerce. In Operations Research Applications. https://doi.org/10.4018/ijwp.2014070104

[5] Sivapalan, S., Sadeghian, A., Rahnama, H., & Madni, A. M. (2014). Recommender systems in e-commerce. World Automation Congress Proceedings.

[6] Manning, C. (2007). Logistic regression (with R). Changes. https://doi.org/10.1017/CBO9781107415324.004

[7] Lehman, M. ., Ramil, J. F., Wernick, P. D., Perry, D. E., & Turski, W. M. (1997). Metrics and laws of software evolution. Software Metrics Symposium. https://doi.org/10.1109/METRIC.1997.637156

[8] Burns, R. P., & Burns, R. (2008). Chapter 24 Logistic Regression. Business Research Methods and Statistics Using SPSS.

[9] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for support vector machines. ACM Transactions on Intelligent Systems and Technology. https://doi.org/10.1145/1961189.1961199

[10] Grégoire,G.(2015).Logistic regression. EAS Publications Series. https://doi.org/10.1051/eas/1466008

[11] Moshrefjavadi, M. H., Rezaie Dolatabadi, H., Nourbakhsh, M., Poursaeedi, A., & Asadollahi, A. (2012). An Analysis of Factors Affecting on Online Shopping Behavior of Consumers. International Journal of Marketing Studies. https://doi.org/10.5539/ijms.v4n5p81

[12] Davis, L. J., & Offord, K. P. (2013). Logistic regression. In Emerging Issues and Methods in Personality Assessment. https://doi.org/10.4324/9780203774618-23

[13] Pampel, F. (2011). Probit Analysis. In Logistic Regression. https://doi.org/10.4135/9781412984805.n4

[14] Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia Medica. https://doi.org/10.11613/BM.2014.003

[15] LaValley,M.P.(2008).Logistic regression.In Circulation. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

[16] Sarkis, J. (2001). Benchmarking for agility. Benchmarking: An International Journal. https://doi.org/10.1108/14635770110389816

[17] Brownlee, J., 2016. Supervised and unsupervised machine learning algorithms. Machine Learning Mastery, 16(03).

[18] Kissell, R. L., & Poserina, J. (2017). Optimal Sports Math, Statistics, and Fantasy. Academic Press.