

Optimization of Marketing Decisions Based on Machine Learning: Case for Telecommunications

Galyna Chornous and Yana Fareniuk

Taras Shevchenko National University of Kyiv, 90-A, Vasylykivska st., Kyiv, 03022, Ukraine

Abstract

Among the main marketing tasks are maintaining the clients and increasing their activity. Based on this, tasks are consumer segmentation and improving communication with them. Machine learning helps to analyze information about subscribers and their use of the company's services and find hidden insights to optimize marketing activities. The purposes of the research are to propose appropriate methods for solving the problem of clustering for customers segmentation and classification of clients who will respond positively to E-mail for the optimization of advertising mailings. The modeling was implemented based on data of the Ukrainian telecommunication company, and the article presents the results of constructing Self Organizing Map (SOM) with the g-means algorithm and k-means clustering to develop subscriber profiles by identifying their similar behavior in terms of frequency, duration of consumption, as well as expenses; determination of the most profitable customer segments. Such information will create a basis for the development of marketing activities aimed at certain groups of customers (personalized communications) and optimization of costs for targeted SMS/E-mail mailing. In order to minimize costs for clients who will not respond to advertising, such classification methods as JRip, DecisionTable, IBk, SMO, NaiveBayes, J48 (C4.5), RandomForest, Logistic regression and others were implemented for the Response to Mailing variable, considering the sampling imbalance, which was solved by an oversampling algorithm. A Cost-Sensitive Classifier has been demonstrated. The RandomForest, J48 and IBk models have the highest quality and are recommended for implementation in order to optimize advertising costs. Thus, based on the applied methods, the company can tailor the mailing to those customers who are more likely to respond. So, the research confirms the feasibility of using models in the clustering and classification of consumers to optimize marketing activities.

Keywords¹

Advertising, machine learning, consumer segmentation, optimizing mailings, E-mail, telecommunication.

1. Introduction

Marketing specialists around the world are trying to find the best solution for their marketing activities. Disruptive technologies such as data analytics and machine learning have changed the ways businesses operate. Of all the revolutionary technologies, artificial intelligence is the technological disruptor and has enormous potential for marketing transformation [1]. Data analytics provides more valuable insights to strengthen business success and make real-time business decisions by scrutinizing and deeply analyzing these data to choose a customized decision with a high level of sophistication [2]. Effective marketing can be built on the basis of high-quality and comprehensive information about the market, competition and consumers. Marketing research is necessary for an explanation of the behavior of the company's customers, determination of possible prospects development and increasing customer satisfaction, which will have a positive influence on the business results. Advanced analysis, mathematical tools and machine learning algorithms allow companies to build


Information Technology and Implementation (IT&I-2022), November 30 – December 02, 2022, Kyiv, Ukraine

EMAIL: Galyna.Chornous@knu.ua (G.Chornous); yfareniuk@gmail.com (Y. Fareniuk)

ORCID: 0000-0003-4889-1247 (G. Chornous); 0000-0001-6837-5042 (Y. Fareniuk)

© 2022 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

intelligent models and Decision Support Systems capable of learning from this data. Intelligent data analysis allows to solve such tasks as customer segmentation, management of customer outflow and determination of the best way to retain them; forecasting the response to the offer; effective attraction of new customers, etc. For businesses, the consumer is the subject of increased attention, because his behavior significantly affects the effectiveness of their marketing activities. Behavioral segmentation of consumers helps to increase sales, which affects consumer loyalty [3]. As a result, the marketing tasks of customer segmentation and the development of an effective customer relationship strategy are very important for all companies on the market. Segmentation of the client's base is a basis for forming effective communication with different target groups by the development of personal advertising propositions.

Advertising is an important competitive tool. Media activity attracts consumers, increases brand awareness, creates loyalty to the company, distinguishes it from competitors or changes the taste of consumers. The company's share of the market drops in the case of the absence of advertising [4]. Media is a means of communication that generates various marketing results among consumers. With the occurrence of mobile communication and smartphones, consumer preferences can be pre-determined and therefore advertising can be delivered to consumers in a multimedia format at the right time and place with the right message. As marketing communications spread, the capability to target the right audience becomes increasingly important. Audience targeting practices in media tend to highlight the demographics, behavior and other consumer characteristics as a basis for selecting the right messages for each audience [5]. In the case of this new advertising opportunity, the development of personalized mobile advertising to meet consumer needs is becoming an important challenge [6].

The mobile advertising paradigm is shifting to personalized advertising services for each consumer in this era of data. In the telecommunications market, the growing demand for smart devices and the emergence of 4G mobile networks have increased the use of mobile services with increased competition among category players. Lately, as the mobile ecosystem has become more complex, marketing specialists are focusing on targeted marketing to clients to maximize ad impact [7] and increase revenue.

The purposes of the research are to propose appropriate machine learning methods for solving the problem of clustering for the segmentation of customers and classification of clients who will respond positively to E-mail for the optimization of advertising mailings. The results of this research can be used as a basis for customer relationship management.

2. Literature review

Well-done segmentation leads to a better understanding of the market and customer needs. The research [8] attempts to develop a new methodological approach combining Recency, Frequency and Monetary with the K-means clustering and provides a useful tool and valid methodology for marketing specialists and decision-makers to accurately identify the most profitable consumer segments.

Arunachalam and Kumar [9] evaluate the effectiveness of different clustering approaches for finding profitable consumer segments. The data are analyzed by hierarchical clustering, K-Medoids, fuzzy clustering and Self Organizing Maps (SOMs). The effectiveness of different clustering methods differs considerably in practice. The obtained results indicate that clustering based on Fuzzy and SOM are comparably more effective than traditional techniques to detect hidden structures in datasets. Segments derived from SOM have more potential to provide interesting and useful insights for data-based decision-making in business practice. Pukala R. [10] implements the artificial neural networks (ANNs) for quantifying the risks of an innovative company by the Kohonen network. Approaches to market segmentation and consumer diagnostics based on multivariate statistical analysis and ANN are considered in [11][11] (market segmentation using the SOM and further refinement of the results by the k-means algorithm, and consumer diagnosis by discriminant analysis and multilayer perceptron).

Zethmayr and Makhija [12] apply k-means clustering to customers according to demographics. Data analysis in the paper [13] was carried out by clustering according to Ward's methodological approach, which identified groups with different socio-demographic characteristics and hierarchical preferences. Ortiz et al. [14][14] explore the two-stage clustering for gaining a deeper understanding of consumers' consumption behavior by profiling and identification of consumer segments

considering their habits and lifestyles. The purpose of the paper [15] is the determination the motivational profile of consumers with a factor-cluster analysis using exploratory factor analysis with an unweighted least squares method. The discriminant analysis established factor importance and demographics.

One of the goals of training in predictive analytics is to create a model from a set of data. The goal of various strategies and experiments is to create a more accurate forecasting model. The aims of [16] are to take sequential steps to find an accurate model for data and save it for future implementation with Python. An integrated intelligent system for monitoring, modeling and managing the life cycle of the company's products has been developed in [17]. This system is presented in the form of a three equations structure, which functions in conditions of instability. The system is based on the concepts, principles, a set of nonlinear models and methods of decision-making and management.

One of the biggest challenges to creating more successful marketing strategies in the telecom market is understanding the diversity of consumer needs and the identification of consumer segments [18]. In consumer research, segmentation has been widely used to identify subsets of consumers based on their preferences. Since the last decade, a more comprehensive assessment of product performance has led to the consideration of a variety of information. Verain et al. [19][19] determine consumer segments and explore differences between them by consumers' perceptions. To determine consumer segments a three-way cluster analysis around the latent variables approach is examined in [20][20]. This method groups consumers into clusters and evaluate for each cluster an associated product latent variable, attribute weights and a set of consumer indicators that can help to determine product characteristics for the cluster. In studies where, in addition to preference indicators, external information about products as well as about consumers is available, the clustering by latent variables (CLV) methodology can be used for customer segmentation. A direct approach, L-CLV has demonstrated its competence to detect consumer segmentation related to a large number of sociological and behavioral parameters [21][21].

Delley and Brunner [22] investigate consumer segmentation using hierarchical cluster analysis. Consumer descriptions varied greatly between groups, indicating heterogeneity. These results indicate the need to study segmentation during data analysis. Techniques of hierarchical segmentation were used in [23] to determine different groups of consumers based on their values and lifestyles. The results contribute to the theoretical and practical aspects of customer segmentation. The recommendations and findings emphasize the importance of implementing different strategies for each segment. Cha and Park [24][24] show that the results of clustering can be used to find the appropriate strategy for each cluster.

The research by [25] presents an optimal predictive segmentation algorithm to identify subgroups that are homogeneous with regard to certain patterns in customer attributes and predictive to the desired result. The authors create an intuitive segmentation with high interpretability and an optimal targeting process for the company's clients. In this setting, the business develops a small number of messages that will be sent to appropriately selected customers who are most likely to respond to different message types. The proposed method uses consumption, demographics, and participation data to extract underlying predictive rules from the dataset using machine learning algorithms.

Marketers use marketing logic to target ads to specific consumer segments. However, there is not always a clear alignment between consumer segmentation and targeting, which can lead to a potential reduction in effectiveness [26]. Predicting the probability that a user will respond to a particular ad has been a common problem in advertising that has attracted much research attention. In recent years, a growing number of new learning models have emerged to improve ad CTR prediction [27].

Over the past decades, the rapid development in the area of information and communication technologies has led to the expansion of the Internet by broad segments of the population. Thanks to various technological advances, the Internet has made it possible for advertisers to reach their target audience [28][28]. Artificial intelligence technologies have numerous applications for online advertising, particularly to optimize the coverage of target audiences. Choi and Lim [29] investigate and categorize different techniques of machine learning used to improve targeted online advertising. A neural network classifier is proposed by Abrahams et al. [5] to assign ads to groups that represent different media channels. In its ability to classify unviewed ads, the model shows a higher performance of classification than the result generated by a random model by 100-300%. Authors suggest using ANN for automated media planning and advertising targeting. Mobile advertising has

evolved into a technology that allows an advertiser to effectively and efficiently promote products or services to target consumers.

Direct marketing is a crucial instrument for the company's promotion, among which direct mailing is quite important. One approach to improving direct mailing targeting is response modeling, which is predictive modeling that assigns the probability of future responses to customers based on their history with a company. Coussement et al. [30] present well-known statistical methods for data classification and analysis (logistic regression, linear and quadratic discriminant analysis, naive Bayes, neural networks, decision trees such as CHAID, CART and C4.5, and the kNN algorithm). The results show that data mining algorithms (CHAID, CART and neural networks) have well performance, followed by simplified statistical classifiers such as logistic regression and linear discriminant analysis. The research by Reynaldo [31] explores a social network user gender prediction model using AdaBoost, XGBoost, Support Vector Machine and Naive Bayes Classifier combined with grid search and K-Fold validation. Kaefer et al. [32] develop an alternative scoring approach for classifying new clients as "good" or "bad" prospects for direct marketing. The research proves that the approach of using only demographics to profile consumers can be enhanced by observing their purchases. The authors establish multinomial logit and neural network models, which can help to classify and target potential consumers.

Perception of SMS advertising has a significant direct or indirect impact on consumer purchase intention. However, there is a dearth of comprehensive research that suggests the predictors of SMS advertising perception and the process by which it impacts purchase intention. The research [33] focuses on developing a model based on the stimulus-organism-response framework with a two-stage hybrid model using PLS-SEM and ANN. Research benefits marketing specialists by facilitating better decision-making for developing effective advertising campaigns using mobile SMS advertising.

SMS helps companies to make direct interactions with their target consumers at any time and location using their mobile phones. Using a modified technology acceptance model, the paper [34][34] explores the influencing factors of acceptance of SMS advertising by consumers. The usefulness is important in establishing favorable consumer attitudes toward SMS advertising. Authors show that consumers perceive SMS advertising differently. Email should be direct and personalized.

Measuring the effectiveness of email marketing is difficult. To maintain competitiveness, managers must maximize profits from mailings by deciding who should receive them. Before achieving the main purpose of converting sales, the intermediate goal of email campaigns is to capture interest and drive traffic to the website. The paper [35] examines the relevance of variables that impact recipient interest in promotional emails and provides companies with actionable and useful insights on how to plan and deploy email marketing strategies with higher efficiency. Paper [36] presents a two-step approach that allows companies to consider the dynamic consequences of mailing and make effective mailing decisions by maximizing the customer's long-term value. The authors suggest a heterogeneous hidden Markov model to capture the interactive dynamics between customers and mailings and use the resulting parameters to develop optimal mailing decisions using a Partial Observable Markov Decision Process. Both immediate and remote consequences of mailings are taken into account. Although email marketing is one of the most cost-effective tools, it remains problematic due to low email open rates and a high percentage of unsubscribed campaigns. The structure and content of the topic are investigated in [37] together with various machine learning techniques (Random Forest, Decision Trees, ANN, Naive Bayes, Support Vector Machines and Gradient Boosting). The results show that combining the data leads to more accurate classification.

Nowadays, mobile advertising focuses on powerful algorithms for personalized recommendations. Chen and Hsieh [6] propose the fuzzy Delphi method to determine the main personalized attributes in a personalized mobile advertising message for various products.

But many questions about the peculiarities of optimization marketing decisions in aspects of customer relationship management remain insufficiently studied, in particular, the customer base segmentation in the telecommunication market using clustering methods, as well as optimization of e-mail marketing via classification of clients who will positively or negatively respond to mailings. Solving such marketing tasks forms an information-analytical basis for marketing activity optimization and making effective decisions for further business development and marketing strategy.

3. Methodology and dataset description

In such a high-tech field as telecommunications, as well as in the field of marketing in general, machine learning methods and approaches have been widely used. Among the main problems that need to be solved are, first of all, those related to loyalty programs and maintenance of the existing client base, as well as the attraction of new consumers of services.

Internal systems of telecommunications companies accumulate large volumes of data every day. First of all, this is information about subscribers and statistics on their use of the company's services. The analysis of such information without using the capabilities of information technologies is ineffective, which creates significant opportunities for the use of approaches and methods of machine learning to optimize marketing activities and increase their effectiveness.

The research goals are to propose the relevant methods to solve the task of clustering (client segmentation) and classification for optimization of advertising mailing for the consumer base. The modeling was implemented based on a database of one of the Ukrainian telecommunications companies, which provides mobile communication and the Internet. The management decided on the need to segment the subscriber base considering the purpose of optimizing the marketing activities, in particular for building profiles of subscribers by identifying their similar behavior in terms of frequency, duration of service use, as well as the level of expenses; assessment and determination of the most profitable customer segments. This forms a scientific hypothesis about factors determining consumer segments. In the future, such information will create a basis for the development of marketing activities aimed at certain groups of customers (personalized promotional communications); development of new tariff plans; optimization of costs for addressed SMS/Viber/E-mail distribution in relation to new services and tariffs; predicting and avoiding the outflow of customers to other competing companies.

The data downloaded from the internal system, is a table with the following fields: age of the client, average monthly expenses (average amount of expenses per subscriber for mobile communication and mobile Internet), the average duration of calls (average number of minutes for outgoing calls by a subscriber per month), daytime/evening/night activity per month (number of activity (calls, messages, Internet connections) per month in the morning and daytime/evening/night time, respectively), activity with other cities/countries per month, the share of calls to landline phones (city numbers), the volume of Internet per month (number of Mb of Internet consumed). Only active subscribers of the company who regularly use mobile communication and/or mobile Internet services over the past few months were selected. The dataset for the experiment contains information about 4591 clients of this company and will be used to realize customer segmentation via Kohonen SOM with a g-means algorithm in Deductor Studio and k-mean clustering in Weka. Such methods were selected for solving the task of clustering, considering the specificity of the dataset, which contains client-related data about their consumption of the company's services, and the statement that it is useful techniques for the goal of segmentation, as we mentioned in the literature review. In addition, the task of optimizing mailings to customers with the goal of minimizing costs for those who do not respond to advertising activity often arises in marketing. Using the example of a database that contains information about customers and their activity, the application of classification methods for advertising response will be demonstrated to increase the effectiveness of advertising activity.

The dataset, which was used for the experiment contains information on 13,500 customers, including known responses to the advertising E-mail and information such as gender, age, number of years the customer has been a company's client, the total value and total number of all purchases, the facts of service calls support, etc. In total, 9 independent and 1 dependent variables are available for analysis. The task is reduced to binary classification, where the variable "Response" was chosen as the class indicator and in the case of a positive answer, "1" means feedback, and "0" means no feedback.

The scientific hypothesis is the presented factors describe the probability of a positive or negative response to ad message and machine learning methods can effectively predict this response. The task is to classify consumers as clearly as possible according to the probability of responding to an advertising message. The results were analyzed in order to formulate recommendations for minimizing costs for new mailings to customers. In addition, the following information is also known for this company: costs for one mailing $CM = \text{UAH } 1$, costs for retaining 1 client $CR = \text{UAH } 9$, expected revenue from 1 client $R = \text{UAH } 20$, so it is assumed that the maximization of income from

communication with the consumer through the mailing. Consider the possible classification results (Table 1).

The total revenue will be $TP*(R-CM-CR)-FP*CM$. To evaluate the predictive power of the classification model, it is necessary to compare the expected revenue with that which can be obtained under the condition of mass mailing to all participants. In order to choose the best classifier, the ZeroR, PART, OneR, JRip, Decision Table, IBk, SMO, Naive Bayes, J48(C4.5), Random Forest, Logistic regression, and AdaBoostM1 methods will use. They were selected for solving the classification task, considering the scientific achievements of researchers, which prove the high performance of mentioned approaches, and the specific task of cost optimization with the relevant dataset. The Cost Sensitive option will also be implemented to the mentioned methods. All of them will be compared and the methods with the best performance will be selected for implementation according to the goal of minimizing costs for ineffective mailings and improving revenue.

Table 1

The possible results of consumer classification according to the variable “Response”

Forecast	Fact	Result	Economic essence	Income
No	No	TN	No Contact, No Cost	0
Yes	Yes	TP	Revenue excluding mailing costs	$(R-CM-CR)=10$
No	Yes	FN	Missing contact and expenses, but unearned revenues	0
Yes	No	FP	Costs for a mailing that does not bring results, but a person can react later as the advertising has a delayed effect	$CM*0.9=1*0.9=0.9$

The research was implemented through step-by-step analysis and modeling and the overall process of optimization of marketing decisions for telecommunications companies through machine learning technologies are look as shown in Figure 1.

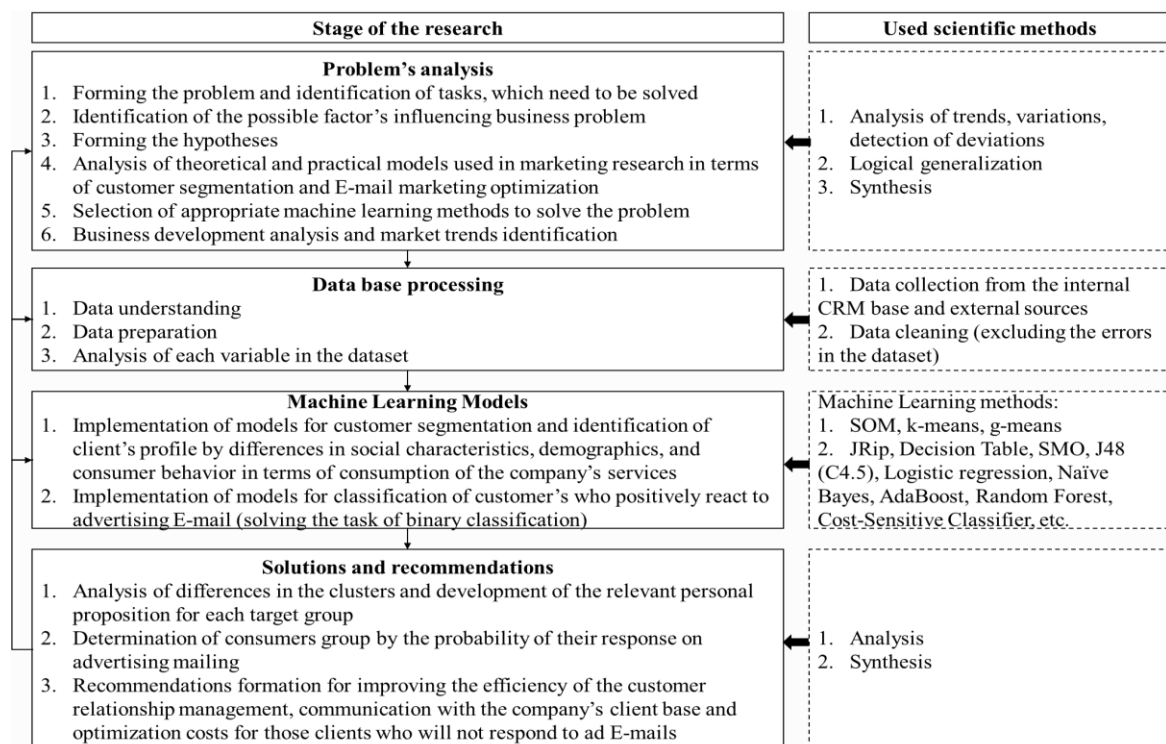


Figure 1: The proposed concept of the research

4. Results

At the initial stage of work with the clustering problem, the SOM algorithm was applied with automatic selection of the number of clusters. For implementation, the Deductor Studio software was used, as a result of which 9 clusters (0-8) with different profiles were formed.

On the basis of the obtained SOMs (Fig. 2), it is possible to analyze in detail the groups of consumers based on various characteristics and the formed customer clusters. Thus, analyzing the "Age" map, three age groups can be clearly distinguished: young people, middle-aged people, and people over 45 years old. Focusing on youth in more detail, we can understand that it is quite heterogeneous and several separate clusters can be distinguished among it. The first is located in the upper right corner and is characterized by those customers who actively use the company's services in the evening and at night, use Internet services to a large extent. As a result, they spend more on mobile communication and mobile Internet than other representatives of this age group. This segment includes most of those who prefer activity at night. It can be predicted that these are students and young people who often spend the evening outside the house or communicate with friends or watch video content. A small group of young people is concentrated below, which is not distinguished by the activity of using services neither during the day, nor in the evening, nor even at night, therefore, as a result, their monthly expenses for communication and the Internet in this cluster are small.

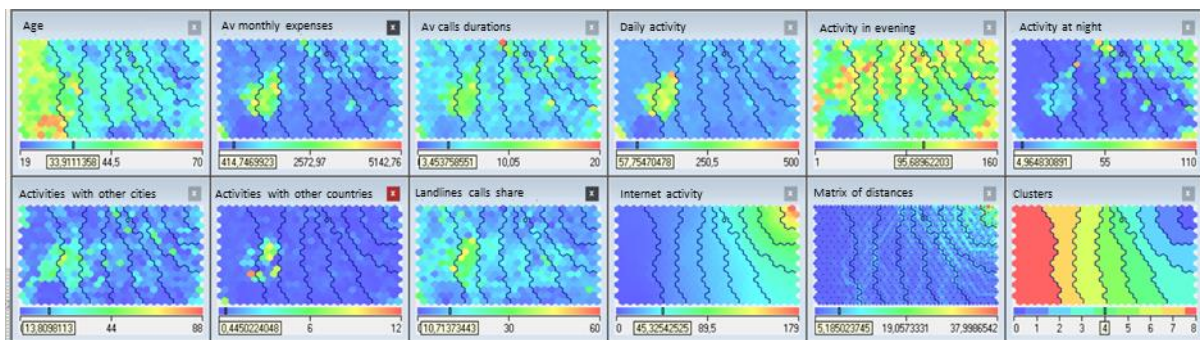


Figure 2: The SOM for client's segmentation constructed in Deductor Studio

The rest of the people of this age group are not distinguished by anything special: moderate expenses for communication and the Internet and to a greater extent activity in the evening. It can be predicted that most of the youth got here. Thus, we clearly identified three clusters in the youth age group. Continuing the interpretation of SOM, we will focus on people of mature and retirement age. Let's pay attention to the pronounced cluster in the lower left part, in which high values are observed for almost all indicators, except for the Internet, including activity with other cities and countries. These are so-called "VIP" clients: businessmen, executives, top managers. The vast majority of them are of mature age, they carry out a lot of activities during the day and in the evening (most likely, due to their work) and use the mobile Internet the least. Monthly expenses for communication and Internet in this category of subscribers are the highest among all the company's clients.

On the left, in one of the clusters, a completely opposite picture is observed: people practically do not use mobile communication and Internet services. Most likely, these are pensioners who need mobile communication and/or the Internet primarily to receive incoming calls, and their independent activity is minimal. Costs for this group of clients are the lowest, which may be due to the fact that the only source of income is a pension. The rest of the people in the mature and retirement age group are united by the fact that they are mainly active in the evening and do not use the Internet very actively. With a greater probability, it can be working pensioners, summer residents, and parents of adult children. The last cluster of middle-aged people includes working subscribers, but among them, there is a group of those who are not very active in the evening (perhaps these are employees with a non-standard work schedule – night/evening shifts, etc.).

However, the automatic determination of the number of clusters using the G-means algorithm produces 9 clusters in this case, which can create difficulties in practical application due to their large number. It is recommended to reduce the number of clusters to 6, and to apply the k-means algorithm. 6 clusters (0...5) were formed in the Weka software, each of which contains from 4% to 31% of customers, which indicates a sufficient number of cases for training and future application of the model. Each cluster is characterized by a unique centroid for each of the 10 indicators, which determines the differences between them and, as a result, determines the differences in the behavior of each group.

Based on the results, we will characterize the clusters shown in Figure 3. Cluster 5 includes mature and older people who have the lowest costs for communication and the Internet and minimal activity in using all services, that is, it can be assumed that this cluster is formed primarily by retirees and people with low incomes. Cluster 4 includes young people up to mature age, who have moderate expenses for Internet and communication, and their activity especially increases in the evening and at night, have the highest share of mobile Internet use, that is, it can be assumed that this group includes active young people who spend their free time outside the home or actively use the Internet for communication or entertainment. Cluster 3 includes young and mature people with moderate expenses and the highest levels of phone conversations, especially the activity increases in the evening. We assumed that the group includes people who actively use the company's services for interpersonal communication. Cluster 2 includes people under 35 years of age with moderate expenses and average indicators of activity. The group actively uses the Internet and different levels of activity in the evening. Cluster 1 includes mature and older people with moderate spending and average activity indicators, with minimal activity at night and on the Internet. Cluster 0 includes people of almost all age categories who have the highest costs for communication and the Internet, the average level of phone conversations and the highest activity during the day and evening, the highest activity with other cities and countries, that is, it can be assumed that this group is formed "VIP-clients", that is, business representatives who use communication and the Internet for business purposes.

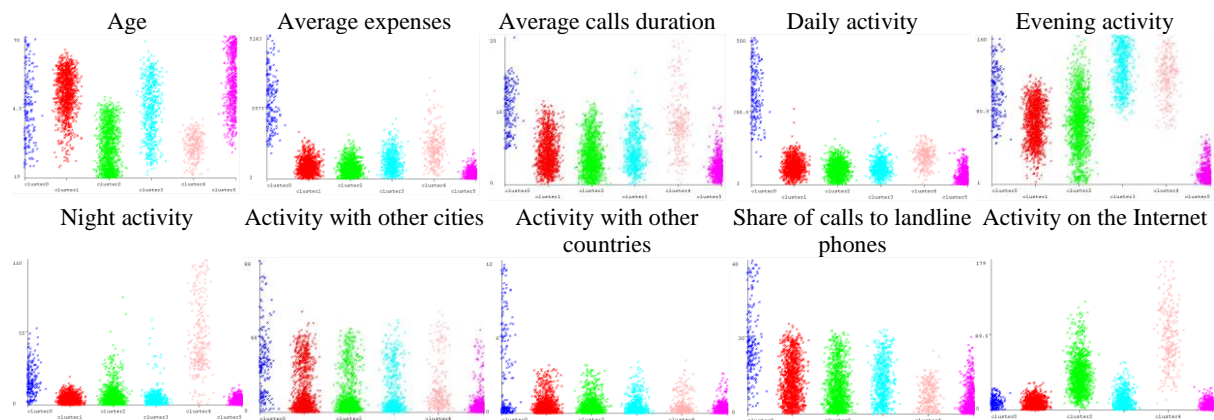


Figure 3: Characteristics of customer clusters of a telecommunications company

Thus, the results of this clustering create a basis for improving the product offer and the company's transition to personalized communication with their subscribers to increase revenue per 1 user, as well as ensure loyalty to the company. These problems will be solved later.

It is important to investigate the question of whether the division of clusters is preserved in different periods of time and the time horizon at which it is appropriate to apply these results. The analysis proves that in the short term (1-3 years) the behavior of the customers is relatively stable because they consume mobile communication and mobile Internet services in almost unchanged volumes, have a relatively unchanged standard of living and income, etc. However, if we talk about the long-term perspective (5-10 years and more), the results of clustering may have significant deviations from reality, as it is affected by a number of factors and general market trends. In this regard, the results of clustering must be updated during the period of annual strategic and tactical planning, so that the results correspond to the actual behavior of customers, and marketing activities are effectively adapted to modern conditions. However, despite the fact that the results require regular verification, the potential and effectiveness of the application of machine learning and Data Science methods for business are constantly growing. The conclusions formed are part of complex research for the formation of an effective marketing strategy and, as a result, the management of marketing activities in general. Segmentation of customers based on machine learning increases the quality of advertising planning as it makes it possible to launch personalized communication with digital placement tools, as well as to ensure quality management of customer loyalty due to product improvement and its relevant offer to interested segments.

The next marketing task, that needs to be solved is the optimization of e-mail mailings to customers in order to maximize response and minimize costs for those who do not respond to advertising activity. The application of different classification methods will help to increase the efficiency of advertising activity via the development of effective recommendations for future advertising mailing. The results of various classification approaches were analyzed to assess the quality of the classification.

According to the classification results, it can be seen that OneR, JRip and Decision Table show the best accuracy (compared to the baseline Zero R classifier, the accuracy increased from 85.5% to 92.0-92.2%). Among them, the Decision Table approach showed the highest results, but if we take into account the business goal of minimizing costs, then the error brings greater losses when there is no feedback, and the model predicted that there should have been feedback. This error is the smallest in the OneR method (none of the clients were incorrectly clustered for this problem). In this regard, it is recommended to use a combination of Decision Table and OneR.

Among the methods based on decision trees, the J48(C4.5) algorithm shows slightly worse results (accuracy - 92.0%, 129 observations are classified as positive, but there was no feedback, which implies additional costs for the company). The decision tree with standard settings is the best option for the algorithm (highest accuracy without overtraining). The decision tree is quite extensive, but it provides a clear understanding of the factors that influence whether a customer will respond to an advertising message. The tree turns into the so-called "golden rules" for setting up mailings to customers. However, the classification results indicate that it is advisable to solve the problem of sample imbalance since the results are close to random classification by most methods (response to feedback is distributed almost 50%/50% between classes). At the moment, the share of customers who respond to the E-mail is 14.5%, which creates a significant imbalance between those who will respond and those who will not respond to the advertising message. Using the oversampling algorithm, that is, increasing the share of a certain class, we will balance the sample: considering the current share of the response = 1 class, it is advisable to increase it by 5 times to obtain balanced results (46% for class 1 and 54% for class 0). Let's repeat the clustering according to the algorithms mentioned above for balanced samples, and compare the results in Table 2.

According to the classification results, Table 2 shows that IBk, J48 and Random Forest have the best accuracy (92.9-97.2%). Among them, the Random Forest approach showed the highest results both in terms of overall accuracy and in terms of minimizing the cost of inefficient mailings. The Random Forest algorithm is recommended for implementation in marketing planning and marketing activities from the point of view of optimizing costs for advertising activity.

The next step in improving the classification results is the application of the Cost-Sensitive Classifier for all methods, taking into account the cost matrix presented in Table 1. According to the classification results from Table 2 and Table 3, there is a conclusion that JBk, J48 and Random Forest maintain the best performance in terms of accuracy (83.2-95.8%) and potential revenue. Among them, the Random Forest approach showed the highest results without taking into account the Cost-Sensitive classification approach. The Random Forest algorithm in combination with J48 and IBk is recommended for implementation in telecommunication companies in order to optimize costs for advertising activity, which is one of the key areas of marketing activity. Thus, based on the applied data classification methods, the company can tailor mailings to those customers who are more likely to respond to the message. As a result, costs will be minimized, and revenues will increase.

To maximize marketing efficiency, it is advised to implement machine learning technologies into the regular management of consumer behavior. The effective concept of such implementation of modeling is a cyclic process, which accumulates next stages:

- Obtaining historical data on the influencing factors and variables that describe the factors;
- Updating the models, evaluating the effectiveness of previous decisions and current results;
- Formation of recommendations for marketing activities and work with the consumer base;
- Implementation of recommended solutions.

In the case of regular support of machine learning models for the company, we may determine business tasks depending on different time intervals (monthly and quarterly). The main tasks for weekly time intervals are the realization of business monitoring, checking the efficiency of marketing solutions, and evaluating the quality of constructed models and their accuracy. The main tasks on a

quarterly basis are model updates and formation of recommendations for future marketing decisions, evaluation of the effectiveness of previous solutions.

Table 2

The by different methods for a balanced sample

Method	Accuracy	AUC	Confusion Matrix			Return, UAH
PART	84.6%	0.94	a	b	←classified as	15 430
			8991	769	a=1	
			2552	9030	b=0	
OneR	76.2%	0.77	a	b	←classified as	12 887
			8200	1560	a=1	
			3513	8039	b=0	
JRip	86.6%	0.91	c	b	←classified as	15 612
			8710	1050	a=1	
			1808	9744	b=0	
Decision Table	82.3%	0.92	a	b	←classified as	15 077
			9097	663	a=1	
			3117	8435	b=0	
IBk	95.8%	0.94	a	b	←classified as	18 616
			9760	0	a=1	
			904	10648	b=0	
SMO	79.3%	0.79	a	b	←classified as	12 583
			7229	2531	a=1	
			1875	9677	b=0	
Naïve Bayes	77.9%	0.88	a	b	←classified as	10 931
			5880	3880	a=1	
			829	10723	b=0	
J48	92.9%	0.96	a	b	←classified as	17 847
			9592	168	a=1	
			1337	10215	b=0	
Random Forest	97.2%	0.999	a	b	←classified as	18 918
			9760	0	a=1	
			602	10950	b=0	
Logistic	79.1%	0.87	a	b	←classified as	12 373
			7063	2697	a=1	
			1753	9799	b=0	
AdaBoostM1	79.5%	0.90	a	b	←classified as	13 961
			8579	1181	a=1	
			3197	8355	b=0	

5. Conclusions

Customer segmentation and marketing management are crucial tasks of the marketing system of a telecommunications company, which is developing in conditions of oversaturation of the market and the task of increasing the quality of communication with a client become an area of potential optimization, which will minimize unnecessary costs and generate additional growth of revenue. To ensure the effective functioning of companies on the market, it is necessary to implement approaches, methods and models of machine learning for customer data.

The implementation of machine learning technologies provided a qualitative result of the research. The results of constructing the SOM and k-means clustering for customer segmentation for one of the leading Ukrainian telecommunication companies become a basis for the development of client profiles based on their demographics and data about their consumption of the company's services in terms of frequency, duration of service use, as well as the level of expenses. Such an approach will help to optimize marketing activities, in particular via assessment and determination of the most profitable customer segments and identification of the key differences between target groups. These differences form a basis for future development of marketing activities aimed at certain groups of customers (for example, personalized communications and promotional activities); development of new tariff plans; optimization of costs for addressed SMS/E-mail mailing; minimizing the outflow of customers to competitors.

Table 3

The results of Cost-Sensitive classification by different methods for a balanced sample

Method	Accuracy	AUC	Confusion Matrix			Return, UAH
ZeroR	45.8%	0.50	a	b	←classified as	7 968
			9760	0	a=1	
			11552	0	b=0	
PART	81.7%	0.83	a	b	←classified as	15 627
			9759	1	a=1	
			3891	7661	b=0	
JRip	82.7%	0.83	c	b	←classified as	14 925
			8848	912	a=1	
			2771	8781	b=0	
Decision Table	80.8%	0.82	a	b	←classified as	15 122
			9455	305	a=1	
			3788	7764	b=0	
IBk	95.8%	0.96	a	b	←classified as	18 616
			9760	0	a=1	
			904	10648	b=0	
Naïve Bayes	75.5%	0.77	a	b	←classified as	13 520
			8709	1051	a=1	
			4168	7384	b=0	
J48	92.9%	0.93	a	b	←classified as	17 883
			9632	128	a=1	
			1381	10171	b=0	
Random Forest	83.2%	0.85	a	b	←classified as	15 949
			9760	0	a=1	
			3571	7981	b=0	
AdaBoostM1	81.2%	0.82	a	b	←classified as	15 109
			9363	397	a=1	
			3617	7935	b=0	

Considering the goal of choosing the best classification model, there was an investigation of the following methods: ZeroR, PART, OneR, JRip, Decision Table, IBk, SMO, Naïve Bayes, J48(C4.5), Random Forest, Logistic regression and AdaBoostM1 with additionally using the Cost-Sensitive option. As a result, the Random Forest algorithm in combination with J48 and IBk showed the best performance and the best economic effect and is recommended for implementation in telecommunication companies in order to optimize costs for advertising activity, which is one of the key areas of marketing activity. Thanks to the applied classification, the company can tailor mailings to those customers who have the highest probability of positively responding to the advertising SMS / E-mail. These decisions will help to minimize advertising costs and increase revenue by more than 137% vs random mailing for all consumer base. The estimation of accuracy reached over 80%, which indicates the possibility and feasibility of using models in the further classification of customer responses to determine the most effective consumer segments.

To maximize marketing efficiency, it is advised to implement machine learning technologies into the regular management of consumer behavior and customer relationship management. The effective concept of such implementation of modeling is a cyclic process for maintaining the actuality of constructed models for the current business environment and consumer behavior and preferences.

The results of the research, constructed models and the proposed concept of the research can be applied in real business practice to optimize marketing activities for both Ukrainian and international companies in the telecommunications market by making effective data-driven decisions and to improve the mathematical methodology of consumer segmentation and optimization of advertising mailings. Marketing strategy optimization based on data-based decisions and finding hidden insights in data has a significant influence on business efficiency due to the high quality and great validity of the decision-making process in very dynamically developing conditions on the market. As an area of future research, it is relevant to focus on overcoming the limitations of current research (in particular, the collection of different indicators about consumer characteristics and their service preferences), and the periodic support of constructed models in different market conditions due to possible changes in consumer behavior. It is necessary to identify new potential factors in a timely manner, which will lead to enhancing marketing decisions. Therefore, it is advisable to

conduct research on a regular basis, which can be effectively implemented in future marketing activities.

6. References

- [1] S. Verma, R. Sharma, S. Deb, D. Maitra, Artificial intelligence in marketing: Systematic review and future research direction, *International Journal of Information Management Data Insights* (2021), Vol. 1, Issue 1. DOI: <https://doi.org/10.1016/j.ijime.2020.100002>.
- [2] J. Saidali, H. Rahich, Y. Tabaa, A. Medouri, The combination between Big Data and Marketing Strategies to gain valuable Business Insights for better Production Success, *Procedia Manufacturing* (2019), Vol. 32, pp. 1017-1023. DOI: [10.1016/j.promfg.2019.02.316](https://doi.org/10.1016/j.promfg.2019.02.316).
- [3] W. H. Susilo, An Impact of Behavioral Segmentation to Increase Consumer Loyalty: Empirical Study in Higher Education of Postgraduate Institutions at Jakarta, *Procedia - Social and Behavioral Sciences* (2016), Vol. 229, pp. 183-195. DOI: <https://doi.org/10.1016/j.sbspro.2016.07.128>.
- [4] R. Amir, D. Machowska, M. Troege, Advertising patterns in a dynamic oligopolistic growing market with decay, *Journal of Economic Dynamics and Control* (2021), Vol. 131. DOI: <https://doi.org/10.1016/j.jedc.2021.104229>.
- [5] A. S. Abrahams, E. Coupey, E. X. Zhong, R. Barkhi, P. S. Manasantivongs, Audience targeting by B-to-B advertisement classification: A neural network approach, *Expert Systems with Applications* (2013), Vol. 40, Issue 8, pp. 2777-2791. DOI: [10.1016/j.eswa.2012.10.068](https://doi.org/10.1016/j.eswa.2012.10.068).
- [6] P.-T. Chen, H.-P. Hsieh, Personalized mobile advertising: Its key attributes, trends, and social impact, *Technological Forecasting and Social Change* (2012), Vol. 79, Issue 3, pp. 543-557. DOI: <https://doi.org/10.1016/j.techfore.2011.08.011>.
- [7] K. Y. Kim, B. G. Lee, Marketing insights for mobile advertising and consumer segmentation in the cloud era: A Q-R hybrid methodology and practices, *Technological Forecasting and Social Change* (2015), Vol. 91, pp. 78-92. DOI: <https://doi.org/10.1016/j.techfore.2014.01.011>.
- [8] J. Zhou, L. Zhai, A. A. Pantelous, Market segmentation using high-dimensional sparse consumers data, *Expert Systems with Applications* (2020), Vol. 145. DOI: [10.1016/j.eswa.2019.113136](https://doi.org/10.1016/j.eswa.2019.113136).
- [9] D. Arunachalam, N. Kumar, Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making, *Expert Systems with Applications* (2018), Vol. 111, pp. 11-34. DOI: <https://doi.org/10.1016/j.eswa.2018.03.007>.
- [10] R. Pukala, Use of Neural Networks in Risk Assessment and Optimization of Insurance Cover in Innovative Enterprises, *Engineering Management in Production and Services* (2016), Vol. 8, No. 3, pp. 43-56. DOI: <https://doi.org/10.1515/emj-2016-0023>.
- [11] E. Pesikov, O. Zaikin, E. Kozlova, Conducting market segmentation and diagnostics of the consumer printed products by using methods of multivariate statistical analysis and artificial intelligence, *IFAC Proceedings Volumes* (2013), Vol. 46, Issue 9, pp. 2116-2121. DOI: <https://doi.org/10.3182/20130619-3-RU-3018.00642>.
- [12] J. Zethmayr, R. S. Makhija, Six unique load shapes: A segmentation analysis of Illinois residential electricity consumers, *The Electricity Journal* (2019), Vol. 32, Issue 9. DOI: <https://doi.org/10.1016/j.tej.2019.106643>.
- [13] G. Di Vita, R. Zanchini, G. Falcone, M. D'Amico, F. Brun, G. Gulisano, Local, organic or protected? Detecting the role of different quality signals among Italian olive oil consumers through a hierarchical cluster analysis, *Journal of Cleaner Production* (2021), Vol. 290. DOI: <https://doi.org/10.1016/j.jclepro.2021.125795>.
- [14] A. Ortiz, C. Díaz-Caro, D. Tejerina, M. Escribano, E. Crespo, P. Gaspar, Consumption of fresh Iberian pork: Two-stage cluster for the identification of segments of consumers according to their habits and lifestyles, *Meat Science* (2021), Vol. 173. DOI: [10.1016/j.meatsci.2020.108373](https://doi.org/10.1016/j.meatsci.2020.108373).
- [15] A. Higuchi, R. Maehara, A factor-cluster analysis profile of consumers, *Journal of Business Research* (2021), Vol. 123, pp. 70-78. DOI: <https://doi.org/10.1016/j.jbusres.2020.09.030>.
- [16] E. Geldiev, N. Nenkov, M. Petrova, Exercise of Machine Learning Using Some Python Tools and Techniques. *CBU International conference proceedings 2018: Innovations in Science and Education*, 21.-23.03.2018 (2018), pp.1062-1070. DOI: <https://doi.org/10.12955/cbup.v6.1295>.
- [17] S. Ramazanov, V. Babenko, O. Honcharenko, N. Moisieieva, V. Dykan, Integrated Intelligent Information and Analytical System of Management of a Life Cycle of Products of Transport Companies. *Journal of Information Technology Management* (2020), 12(3), pp. 26-33. DOI: <https://doi.org/10.22059/JITM.2020.76291>.

- [18] M. C. Onwezen, M. J. Reinders, I. A. van der Lans, S. J. Sijtsema, A. Jasiulewicz, M. D. Guardia, Luis Guerrero, A cross-national consumer segmentation based on food benefits: The link with consumption situations and food perceptions, *Food Quality and Preference* (2012), Vol. 24, Issue 2, pp. 276-286. DOI: <https://doi.org/10.1016/j.foodqual.2011.11.002>.
- [19] M. C.D. Verain, S. J. Sijtsema, G. Antonides, Consumer segmentation based on food-category attribute importance: The relation with healthiness and sustainability perceptions, *Food Quality and Preference* (2016), Vol. 48, Part A, pp. 99-106. DOI: [10.1016/j.foodqual.2015.08.012](https://doi.org/10.1016/j.foodqual.2015.08.012).
- [20] V. Cariou, T. F. Wilderjans, Consumer segmentation in multi-attribute product evaluation by means of non-negatively constrained CLV3W, *Food Quality and Preference* (2018), Vol. 67, pp. 18-26. DOI: <https://doi.org/10.1016/j.foodqual.2017.01.006>.
- [21] E. Vigneau, M. Charles, M. Chen, External preference segmentation with additional information on consumers: A case study on apples, *Food Quality and Preference* (2014), Vol. 32, Part A, pp. 83-92, <https://doi.org/10.1016/j.foodqual.2013.05.007>.
- [22] M. Delley, T. A. Brunner, A segmentation of Swiss fluid milk consumers and suggestions for target product concepts, *Journal of Dairy Science* (2020), Vol. 103, Issue 4, pp. 3095-3106. DOI: <https://doi.org/10.3168/jds.2019-17325>.
- [23] A. Díaz, M. Gómez, A. Molina, J. Santos, A segmentation study of cinema consumers based on values and lifestyle, *Journal of Retailing and Consumer Services* (2018), Vol. 41, pp. 79-89. DOI: <https://doi.org/10.1016/j.jretconser.2017.12.001>.
- [24] Y. U. Cha, M. J. Park, Consumer preference and market segmentation strategy in the fast moving consumer goods industry: The case of women's disposable sanitary pads, *Sustainable Production and Consumption* (2019), Vol. 19, pp. 130-140. DOI: <https://doi.org/10.1016/j.spc.2019.04.002>.
- [25] A. Albert, M. Maasoumy, Predictive segmentation of energy consumers, *Applied Energy* (2016), Vol. 177, pp. 435-448. DOI: <https://doi.org/10.1016/j.apenergy.2016.05.128>.
- [26] C. Lutz, G. Newlands, Consumer segmentation within the sharing economy: The case of Airbnb, *Journal of Business Research* (2018), Vol. 88, pp. 187-196. DOI: [10.1016/j.jbusres.2018.03.019](https://doi.org/10.1016/j.jbusres.2018.03.019).
- [27] Y. Yang, P. Zhai, Click-through rate prediction in online advertising: A literature review, *Information Processing & Management* (2022), Vol. 59, Issue 2. DOI: <https://doi.org/10.1016/j.ipm.2021.102853>.
- [28] N. Deshpande, S. Ahmed, A. Khode, Web based Targeted Advertising: A Study based on Patent Information, *Procedia Economics and Finance* (2014), Vol. 11, pp. 522-535. DOI: [https://doi.org/10.1016/S2212-5671\(14\)00218-4](https://doi.org/10.1016/S2212-5671(14)00218-4).
- [29] J.-A. Choi, K. Lim, Identifying machine learning techniques for classification of target advertising, *ICT Express* (2020), Vol. 6, Issue 3, pp. 175-180. DOI: <https://doi.org/10.1016/j.icte.2020.04.012>.
- [30] K. Coussement, P. Harrigan, D. F. Benoit, Improving direct mail targeting through customer response modeling, *Expert Systems with Applications* (2015), Vol. 42, Issue 22, pp. 8403-8412. DOI: <https://doi.org/10.1016/j.eswa.2015.06.054>.
- [31] N. Reynaldo, Goenawan, W. Chanrico, D. Suhartono, F. Purnomo, Gender Demography Classification on Instagram based on User's Comments Section, *Procedia Computer Science* (2019), Vol. 157, pp. 64-71. DOI: <https://doi.org/10.1016/j.procs.2019.08.142>.
- [32] F. Kaefer, C. M. Heilman, S. D. Ramenofsky, A neural network application to consumer classification to improve the timing of direct marketing activities, *Computers & Operations Research* (2005), Vol. 32, Issue 10, pp. 2595-2615. DOI: [10.1016/j.cor.2004.06.021](https://doi.org/10.1016/j.cor.2004.06.021).
- [33] A. Sharma, Y. K. Dwivedi, V. Arya, M. Q. Siddiqui, Does SMS advertising still have relevance to increase consumer purchase intention? A hybrid PLS-SEM-neural network modelling approach, *Computers in Human Behavior* (2021), Vol. 124. DOI: [10.1016/j.chb.2021.106919](https://doi.org/10.1016/j.chb.2021.106919).
- [34] A. Muk, C. Chung, Applying the technology acceptance model in a two-country study of SMS advertising, *Journal of Business Research* (2015), Vol. 68, Issue 1, pp. 1-6. DOI: <https://doi.org/10.1016/j.jbusres.2014.06.001>.
- [35] Á. J. Lorente-Páramo, J. Chaparro-Peláez, Á. Hernández-García, How to improve e-mail click-through rates – A national culture approach, *Technological Forecasting and Social Change* (2020), Vol. 161. DOI: <https://doi.org/10.1016/j.techfore.2020.120283>.
- [36] S. Ma, L. Hou, W. Yao, B. Lee, A nonhomogeneous hidden Markov model of response dynamics and mailing optimization in direct marketing, *European Journal of Operational Research* (2016), Vol. 253, Issue 2, pp. 514-523. DOI: <https://doi.org/10.1016/j.ejor.2016.02.055>.
- [37] M. Paulo, V. L. Miguéis, I. Pereira, Leveraging email marketing: Using the subject line to anticipate the open rate, *Expert Systems with Applications* (2022), Vol. 207. DOI: <https://doi.org/10.1016/j.eswa.2022.117974>.