Data Mining Techniques to Improve the Response Rate of E-mail campaigns and Customer Loyalty

Data Mining Techniques to Improve the Response Rate of E-mail campaigns and Customer Loyalty

Hamzah Qabbaah[1] , George Sammour[2] and Koen Vanhoof[1]
[1]Department of Business Informatics, Hasselt University, Diepenbeek, Belgium
Hamzah.qabbaah@uhasselt.be, koen.vanhoof@uhasselt.be
[2]Department of Management Information system, Princess Sumaya University for Technology, Amman, Jordan
George.sammour@psut.edu.jo

Abstract. The efficiency of e-mail campaigns is a big challenge for any e-commerce venture in terms of the response rate of e-mail campaigns and customer segmentation based on loyalty. Data mining techniques are useful tools to extract customer information related to response rate from e-mail campaigns data. This study aims at predicting customer loyalty and improving the response rate of e-mail campaigns, specifically open rate and click through rate, using data mining techniques such as logistic regression and clustering.
The models are trained using chi square and logistic regression techniques to detect the effect of customers' loyalty based on their demographic and behavioural characteristics. Furthermore, a clustering technique is used to segment customers based on their behavioural characteristics . The models reported satisfactory results in predicting customer loyalty based on open rate and click through rate values. In addition, the clustering of customers suggest that companies will have a better understanding of their customers in terms of their demographic and behavioural characteristics. The response rates also increase at the preferred moment at which e-mails should be send to customers in email campaigns.

Keywords: E-business, e-mail campaigns, logistic regression, chi square, open rates, click through rates.

1 Introduction

Due to the rapid development and popularization of internet use, e-commerce and online shopping have grown extraordinarily [1]. In this paper we understand e-commerce to be the buying and selling of products or services through electronic media, such as Internet and other computer networks [2].

One important element of e-commerce practices is web advertising, through which companies can deliver advertisements directly to customers using different channels such as e-mail campaigns and contextual advertising [3].

The appeal of e-mail communication as a direct channel to talk to customers is double: cost effectiveness and time efficiency [4]. However, if companies want to use

e-mail as a direct communication channel with their customers, they need to understand the process through which e-mail campaigns affect customers' attitudes and behavior thoroughly [5]. This knowledge can then be turned into a competitive advantage by optimizing the design of e-mail campaigns.

This subject has attracted much attention in recent years. But although e-mail marketing research studies have been conducted either by online surveys, by in-depth interviews, by controlled experiments or by tracking behavior patterns such as click-through links and the visiting patterns, few of these studies have really investigated the effects of e-mail characteristics on consumer attitudes and behavioral intentions [6].

Loyalty and response rates have been an important focus of attention of this research stream. Since e-mail campaigns (i.e., more than one email sent by a company) from one specific company are usually designed with the same architecture as the website of the company[4]. Other research wants to build models to improve response rates by using individual preferences of customers to personalize e-mail newsletters. Marketing campaigns and products can then be customized to appeal better to groups of customers or to individuals [6].

These data are in most companies abundantly available. They can be turned into a very valuable resource through efficient access to the data, sharing them, extracting information from them. Making efficient use of the information has become an urgent need. Data mining is the process of posting various queries and extracting useful information, patterns, and trends often previously unknown from such large quantities of data [7]. Essentially, for many organizations, the goals of data mining include improving marketing capabilities, detecting abnormal patterns, and predicting the future based on past experiences and current trends. The reason is that data mining techniques can extract or detect hidden customer characteristics and behaviors from large databases [7, 8]. Data mining also defined as "the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. It allows corporations to improve its marketing, sales, and customer support operations through a better understanding of its customers" [9]. It is thus not surprising that the result of data mining can be used for target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise and market analysis [10].

This research builds on the previously mentioned researches and uses on of the several types of models in data mining such as logistic regression and classification techniques. It aims at "predicting the loyalty of customers based on their response to email campaigns and linking that to their demographic and behavioral characteristics." The phases of data classification described are those associated with the extraction and pre-processing of the data, the extraction of knowledge and the analysis of results. Among the techniques employed to analyze the data set are chi square analysis, logistic regression and clustering techniques [2].

The rest of this paper is organized as follows. Section 2 specifies the methodology. Section 3 presents the data and describes the research hypotheses. In section 4 the analysis and the results of the study are presented. Section 5 shows the managerial

implications of the results . Finally, section 6 outlines the conclusions and the future research.

2 Methodology

In the previous section we have indicated that the main idea of this paper is to find a way of predicting the loyalty of customers based on their response to email campaigns and linking that to their demographic and behavioral characteristics.

We could have achieved this goal by experimenting with different campaigns and comparing the resulting customer behavior and linking these data to demographic characteristics of the respondents. This is however very difficult to realize since companies are reluctant to experiment with real life campaigns because they fear that the response rate will drop due to wrong experimenting , while our methodology uses data mining techniques of web data that allows to experiment with campaigns yielding a high potential for increased response rate levels .

Data mining of web data indeed allows to achieve the two major steps needed to achieve the required objective, namely: (1) analysing the impact of customers demographic and behavioural characteristics to customer's loyalty and (2) finding homogenous groups of customers based on customer demographic and behavioural characteristics, thus identifying groups with different loyalty.

Web mining is the use of data mining techniques to automatically discover and extract knowledge from data available on the use of websites [11]. The importance of web mining is considering the behavior and preferences of the user of websites [12]. Several authors subdivide web mining in several stages: (1) Finding resources, (2) Selecting information and pre-processing of the data, (3) Discovering knowledge and finally (4) Analyzing the obtained patterns [13] [14]. This paper uses web usage mining, which is defined as "The process of applying data mining techniques to the discovery of usage patterns from Web data" [15].

In our research data are represented as a collection of e-mails campaigns opened and websites visited both by customers. These data can be employed to understand the main features of the visitors' behaviors in order to find ways to improve the response rate to the e-mail campaigns, more specifically increasing their open and click through rates [2].

The data mining techniques we used in our research are descriptive in nature and threefold: K-means clustering, chi-square and logistical regression techniques , below a brief description of these techniques.

Cluster analysis seeks to separate data elements into groups or clusters with similar characteristics, such that both homogeneity of elements within clusters and the heterogeneity between clusters are maximized [16]. It has been applied in a wide variety of fields, ranging from engineering, computer sciences (web mining, spatial database analysis, and segmentation), life and medical sciences, to earth sciences, social sciences and economics (in marketing, business analysis and CRM management) [6]. Cluster analysis is based on heuristics that try to maximize the similarity between in-

cluster elements and the dissimilarity between inter-cluster elements [17]. We have performed this task through the k-Means algorithm. The main objective of this algorithm is to partition the data set into k clusters in which each instance belongs to the cluster with the nearest mean [2]. K-Means algorithm starts from k central point's chosen randomly and every instance is assigned to the closest central point. Next, the heuristic performs a reassignment of the central points. The algorithm is finally deemed to have converged when the assignments of the individual instances no longer change.

Chi-square analysis uses a numerical test that measures deviation from the expected distribution considering that the featured event is independent of the class value. The chi-square statistic is also used to test the hypothesis of no association between two or more variables or criteria [18].

Calculating the chi-square statistic for the A and B variables requires creating observed and expected tables as follows[19] .

Table 1. Observed table for (A;B)

	B	$\overline{B}$
A	n P(A \ B)	n P(A \ $\overline{B}$)
$\overline{A}$	n P($\overline{A}$ \ B)	n P($\overline{A}$ \ $\overline{B}$)

Table 2. Expected table for (A;B)

	B	$\overline{B}$
A	n P(A) P(B)	n P(A) (1 - P(B))
$\overline{A}$	n (1- P(A)) P(B)	n (1 - P(A)) (1 - P(B))

The Chi-square equation is :

$$X^2 = \sum ((\text{observed} - \text{expected})^2 / \text{expected}) \quad (1)$$

X^2 represents a summed normalized square deviation of the observed values from the corresponding expected values.

Logistical regression is regularly used when there are two categories of the independent and dependent variables. In instances where the independent variables are categorical, or a mix of continuous and categorical variables, and the dependent variable is a categorical variable for instance, logistic regression is necessary. The goal is to correctly predict the category of outcome for individual cases using the proposed model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. There are two main uses of logistic regression: (1) To predict group memberships and (2) to provide additional knowledge on the relationship between the variables. In this sense logistic regression calculates the probability of success over the probability of failure and indicates the results of the analysis in the form of an odds ratio and gives an indication on the strength of the relationship (e.g the customers who have high click rate have a higher probability to buy the product).[20]

The Logistic regression equation is :

$$P= 1/(1 + \exp (–(b_0 + x1*b1 + x2*b2 + \ldots +xp*bp))) \quad (2)$$

3 Data and research hypothesis

Data have been obtained using the webmaster tool of Google Analytics from an E-commerce website. Cleaning and pre-processing of the data has been applied in order to obtain the final data set. As such we filtered out customers who received all 32 campaigns, resulting in a sample of 1428 customers (n=1428). For each customer we collected information such as Id, gender, age, country , language, total e-mails received, total e-mails opened and total e-mails clicked. After that we calculated the click rate, open rate and the average time of opening the e-mail for each customer.

As we wanted to analyze the impact of customers' demographic and behavioral characteristics to customer's loyalty and find homogenous groups of customers based on customer demographic and behavioral characteristics, we first had to identify the effect of "Loyalty segment" based on gender, age, parts of day and country as descriptive variables. We used these as independent variable in our study to define this segment using chi-square analysis.

We then proposed the following research hypothesis:

"Open rates and Click rates are a significant predictor of loyalty" (or "belonging to the loyalty segment"). We tested this hypothesis using a logistic regression technique. Fig 1 shows this conceptual model.

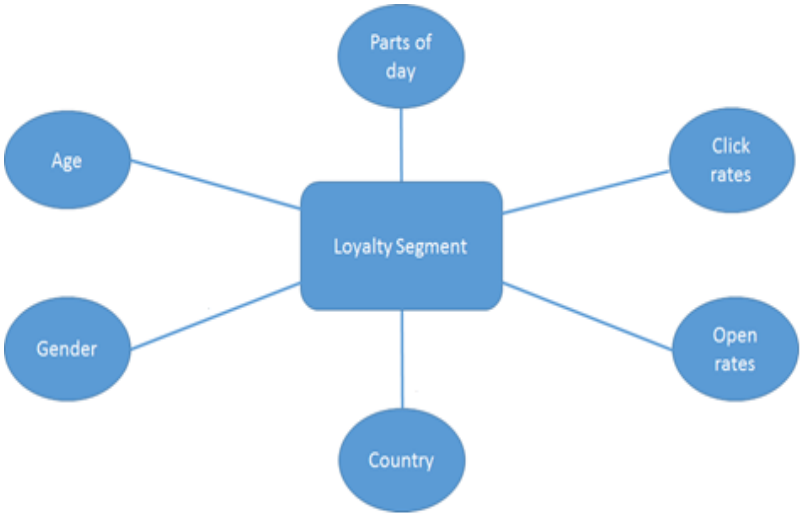

Fig. 1. The conceptual model

4 Analysis and results

The results of the both the chi-square analysis and the logistic regression are presented in the first part of section (4.1.). Next we present three models according to which we have tried to predict customer loyalty (4.2.). Finally, we have used cluster analysis on the behavioral characteristics of the observed case data. The results are in the last part of this section (4.3.)

4.1 Chi square analysis and logistic regression results

Chi-square analysis results. Chi-square test has been used to identify the effect of "Loyalty segment" based on gender, age, parts of day and country as descriptive variables. Table 3 shows the results of the description of the loyalty segment according to the demographic variables using the chi square test.

Table 3 : Chi square results

	Independent variable	Dependent variable	Chi square result	P value
Var 1	Gender (Male , female)	Loyalty segment	X2 (2,N=1428)=46.6	P=0.000
Var 2	Age (younger than 18, 18-29, 30-45, 46-59 , older than 59)	Loyalty segment	X2 (8,N=1428)=779	P=0.000
Var 3	Parts of day (6-9, 9-12, 12-15, 15-18, 18-21 , 21-24, 00-03, 3-6)	Loyalty segment	X2 (14,N=1428)=539	P=0.000
Var 4	Country (Belgium, Netherland)	Loyalty segment	X2 (2,N=1428)=1.93	P =0.380

*P-value<0.001

It's clear that gender, age and parts of day are a significant parameters in describing customer loyalty since P-value < 0.001 , while country is not a significant parameter in describing customer loyalty since P-value = 0.380.

Logistic regression analysis. Logistic regression has been used to assess if open and click rates are a significant predictor of customer loyalty. The results of the research hypothesis are presented in Table 4. We can observe that for both open and click rates the p-value is <0.001. The pseudo R^2 =0.4 and the X^2 model is (526.4, df=2, p-value<0.001). We can conclude that open and click rates are thus a significant predictor of customer loyalty.

Table 4. Logistic regression results (dependent variable is loyalty segment)

Independent variable	B	S.E.	Wald	df	P Value	Odds ratio	95% C.I
Click Rate	.084	.009	80.538	1	.000	1.088	1.068-1.108
Open Rate	.055	.007	65.076	1	.000	1.056	1.042-1.071
Model X^2	526.4 , df=2 , P value<0.001						
Pseudo R^2	0.4						
N	1428						
Hosmer and Lemeshow Test	45.6 , df=8 , P<0.001						

4.2 Prediction models and classification measures tests

We have created three models to predict customer loyalty based on demographic and behavioral variables using logistic regression technique . They can be described as follows:

Model 1: Customer loyalty based on customer demographic attributes.

Model 2: Customer loyalty based on customer behavioral attributes.

Model 3: Customer loyalty based on customer age, gender, part of day, open and click rates.

To evaluate the quality of these models three classification measures, accuracy, recall and precision have been used to see which of the three models predicts customer loyalty best.

Table 5 shows the confusion matrix. Where TP is defined as normal behavior that is correctly predicted, FP indicates normal behavior wrongly assumed whereas abnormal TN specifies the normal performance that is detected as correct and FN indicates the abnormal performance that is misidentified as normal [18, 21].

Table 5. Confusion matrix

		Observed
Predicted	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Based on this confusion matrix we can measure accuracy, recall , precision and F-score as defined in equations 4 to 7 [22].

Accuracy is used to measure the effectiveness of proposed models based on the variables in the equation , it is defined as the ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = TP+TN /(TP+TN+FP+FN) \quad (4)$$

Recall is the ratio of correctly predicted positive observations to the all observations, recall used to measure whether the selected variables that are relevant.

$$\text{Recall} = TP / (TP+FN) \quad (5)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, precision detects the fraction of all recommended variables that are relevant.

$$\text{Precision} = TP / (TP+FP) \quad (6)$$

F-measure combined precision and recall which is the harmonic mean of precision and recall.

$$F_1= 2* ((Recall*Precision)/(Recall+Precision)) \quad (7)$$

These metrics are focused on the performance of the models , all the three derived metrics should be close to 1 for a good model.

Fig 2 shows the classification analysis results of the created models.

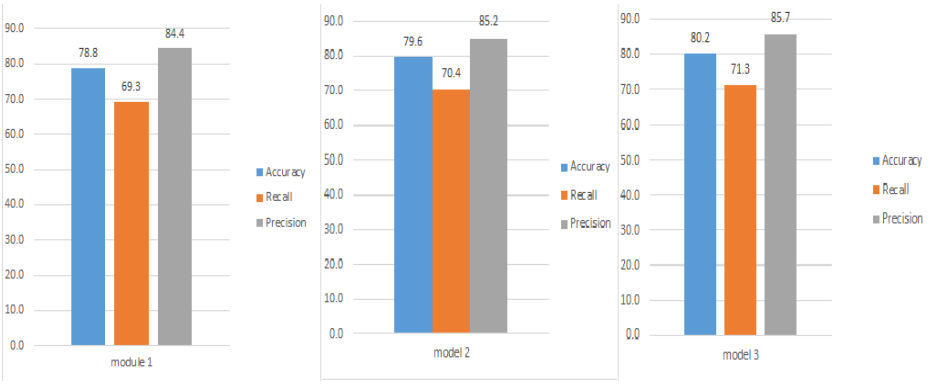

Fig.2 . Classification Results

The accuracy of models 1 and 2 respectively are 78.8% , 79.6% whereas in the model 3 it is 80.2% , The highest recall and precision values are achieved also by model 3. The F-score of models 1,2 and 3 respectively are 76.1%,77.1% and 77.8%. We thus conclude that model 3 can predict customer loyalty somewhat better than model 1 and model 2.

The results also show that the company should focus on the attributes that have a significant impact on customer loyalty when they send the e-mails to customers in order to increase open and click rates of these customers.

4.3 K-means cluster analysis

In this section we presented the results of the application of the K-means cluster algorithm using SPSS software on the data set in order to find homogenous groups of customers based on their demographic and behavioral characteristics. To increase customer loyalty we used the attributes of model 3 as a parameters of K-means algorithm since they are more likely to predict the loyalty accurately. K-Means algorithm obtains the number of clusters indicated by customers interests and variables.

Three separate cluster models were created to find groups of customers sharing the same interests in order to increase customer loyalty as follows . We have experiment the analysis for 2, 3 and 4 clusters and have chosen the best cluster fit in each model.

Cluster Model A. The variables used in cluster model A are open rate , click rate , parts of the day and the loyalty segment. Fig 3 shows the cluster results of the cluster A model. Fig 3A uses a 2 cluster solution, Fig 3B a three cluster solution and Fig 3C a four cluster solution.

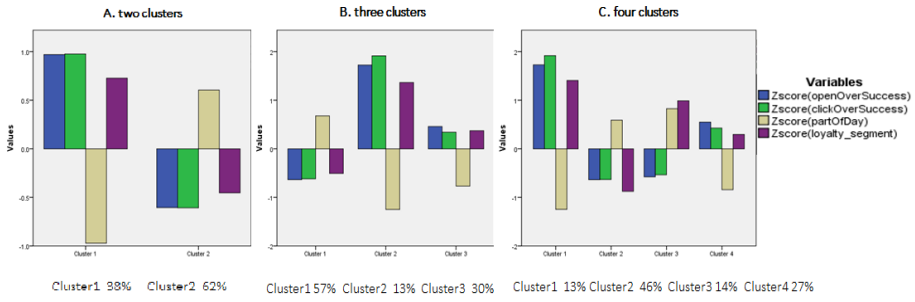

Fig.3. Cluster A model results

Fig 3A shows that the customers in cluster 1 have a higher percentage of open and click rates and they are more loyal than the customers in cluster 2 in the two cluster solution. In Fig 3B the customers in cluster 2 of the three cluster solution seem to have a higher value of open and click rates and they also are more loyal than customers in cluster 1 and cluster 3. Finally in Fig 3C the customers in cluster 1 have the highest percentage of click and open rates and they are more loyal than in the other 3 clusters of the four cluster solution.

Post hoc test has been used to compare the means of the variables in each cluster with the other clusters to validate if the variables are significantly different between the clusters groups. Fig 4 shows a sample of post hoc test table for the 3C (four cluster solution) model in cluster A.

The results show that open rates in cluster 2 compared to cluster 3 are not significantly different since the p-value = 1, and that the click rates in cluster 2 compared to cluster 3 are not significantly either since the p-value = 0.083. This indicates that the 4 cluster solution of clustering model A is not a good model for clustering the variables in the A model. All variables in the three cluster solution model (3B) are significantly different based on the post hoc test as shown in Fig 5. Thus this 3 cluster solution is capable of grouping customers in low, medium and high open and click rates and loyalty. This solution also indicates at which time of the day companies should preferably send e-mail campaigns.

Dependent Variable	(I) Cluster Number of Case	(J) Cluster Number of Case	J)	Std. Error	Sig.	Lower Bound	Upper Bound
Zscore(openOverSuccess)	1	2	2.36664696*	.04464684	.000	2.2486923	2.4846016
		3	2.30866778*	.05477663	.000	2.1639507	2.4533848
		4	1.17994771*	.04811851	.000	1.0528211	1.3070744
	2	1	-2.36664696*	.04464684	.000	-2.4846016	-2.2486923
		3	-.05797917	.04350939	1.000	-.1729288	.0569704
		4	-1.18669924*	.03475600	.000	-1.2785228	-1.0948757
	3	1	-2.30866778*	.05477663	.000	-2.4533848	-2.1639507
		2	.05797917	.04350939	1.000	-.0569704	.1729288
		4	-1.12872007*	.04706504	.000	-1.2530635	-1.0043766
	4	1	-1.17994771*	.04811851	.000	-1.3070744	-1.0528211
		2	1.18669924*	.03475600	.000	1.0948757	1.2785228
		3	1.12872007*	.04706504	.000	1.0043766	1.2530635
Zscore(clickOverSuccess)	1	2	2.55271909*	.04058746	.000	2.4454891	2.6599491
		3	2.45529096*	.04979623	.000	2.3237319	2.5868501
		4	1.49157022*	.04374348	.000	1.3760022	1.6071383
	2	1	-2.55271909*	.04058746	.000	-2.6599491	-2.4454891
		3	-.09742813	.03955343	.083	-.2019263	.0070700
		4	-1.06114887*	.03159591	.000	-1.1446237	-.9776741
	3	1	-2.45529096*	.04979623	.000	-2.5868501	-2.3237319
		2	.09742813	.03955343	.083	-.0070700	.2019263

*. The mean difference is significant at the 0.05 level.

Fig. 4. Post hoc test results of the means of the variables for the four cluster solution in clustering model A , thus related to Fig 3C

Dependent Variable	(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Zscore(openOverSuccess)	1	2	-2.35788308*	.04444594	.000	-2.4644115	-2.2513546
		3	-1.09470410*	.03341741	.000	-1.1747993	-1.0146089
	2	1	2.35788308*	.04444594	.000	2.2513546	2.4644115
		3	1.26317898*	.04841558	.000	1.1471360	1.3792219
	3	1	1.09470410*	.03341741	.000	1.0146089	1.1747993
		2	-1.26317898*	.04841558	.000	-1.3792219	-1.1471360
Zscore(clickOverSuccess)	1	2	-2.53544971*	.04019363	.000	-2.6317862	-2.4391132
		3	-.96146187*	.03022025	.000	-1.0338941	-.8890297
	2	1	2.53544971*	.04019363	.000	2.4391132	2.6317862
		3	1.57398784*	.04378349	.000	1.4690471	1.6789285
	3	1	.96146187*	.03022025	.000	.8890297	1.0338941
		2	-1.57398784*	.04378349	.000	-1.6789285	-1.4690471
Zscore(partOfDay)	1	2	1.92710950*	.04776771	.000	1.8126194	2.0415996
		3	1.44489896*	.03591494	.000	1.3588177	1.5309802
	2	1	-1.92710950*	.04776771	.000	-2.0415996	-1.8126194
		3	-.48221054*	.05203403	.000	-.6069262	-.3574948
	3	1	-1.44489896*	.03591494	.000	-1.5309802	-1.3588177
		2	.48221054*	.05203403	.000	.3574948	.6069262
Zscore(loyalty_segment)	1	2	-1.87235888*	.06008197	.000	-2.0163640	-1.7283538
		3	-.87849970*	.04517362	.000	-.9867723	-.7702271
	2	1	1.87235888*	.06008197	.000	1.7283538	2.0163640
		3	.99385918*	.06544813	.000	.8369924	1.1507259
	3	1	.87849970*	.04517362	.000	.7702271	.9867723
		2	-.99385918*	.06544813	.000	-1.1507259	-.8369924

*. The mean difference is significant at the 0.05 level.

Fig. 5. Post hoc test results of the means of the variables for the three cluster solution in clustering model A , thus related to Fig 3B.

Cluster Model B. The variables used in cluster model B are customer age , gender, open rate , click rate and loyalty segment. Fig 6 shows the cluster results of this model, again in a two, three and four cluster solution option.

Fig. 6. Cluster Model B results

Fig 6A shows that the customers in cluster 2 have a higher percentages of open and click rates when compared to the customers in cluster 1 in a two cluster solution. In Fig 6B the customers in cluster 3 seem to have a higher percentages of open and click rates and are also more loyal than customers in the other clusters of the three cluster solution. Finally in Fig 6C, which shows the four cluster solution , the customers in cluster 3 have the highest percentages of click and open rates compared to the other clusters. Fig 7 shows a sample of a post hoc test table for the four cluster solution in cluster model B (Fig 6C).

Dependent Variable	(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Zscore(openOverSuccess)	1	2	-1.19208408*	.03574195	.000	-1.2865125	-1.0976557
		3	-2.33735989*	.04441375	.000	-2.4546987	-2.2200210
		4	-.11387515	.04586025	.079	-.2350356	.0072853
	2	1	1.19208408*	.03574195	.000	1.0976557	1.2865125
		3	-1.14527581*	.04903111	.000	-1.2748135	-1.0157381
		4	1.07820893*	.05034513	.000	.9451997	1.2112182
	3	1	2.33735989*	.04441375	.000	2.2200210	2.4546987
		2	1.14527581*	.04903111	.000	1.0157381	1.2748135
Zscore(gender)	1	2	.71526691*	.05670657	.000	.5654510	.8650828
		3	.78862610*	.07046485	.000	.6024616	.9747907
		4	-.76923835*	.07275981	.000	-.9614661	-.5770106
	2	1	-.71526691*	.05670657	.000	-.8650828	-.5654510
		3	.07335919	.07779056	1.000	-.1321595	.2788779
		4	-1.48450526*	.07987531	.000	-1.6955318	-1.2734787
	3	1	-.78862610*	.07046485	.000	-.9747907	-.6024616
		2	-.07335919	.07779056	1.000	-.2788779	.1321595
		4	-1.55786445*	.09016499	.000	-1.7960758	-1.3196531

*. The mean difference is significant at the 0.05 level.

Fig. 7. Post hoc test results of the four solution model in cluster model B (related to Fig 6C)

The results show that open rates in cluster 1 compared to cluster 4 are not significantly different which is indicated by a p-value = 0.079. Moreover gender in cluster 2 compared to cluster 3 is not significantly different either with a p-value = 1. This again indicates that a four cluster solution does not offer a good model fit for the variables in cluster model B. All variables in the three cluster solution on the contrary are significantly different indicates P-value < 0.05 based on the post hoc test as shown

in Fig 8. This solution that can group the customers in low, medium and high open, click rates and loyalty with a good fit .

Dependent Variable	(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Zscore(openOverSuccess)	1	2	1.16140578*	.03324821	.000	1.0817161	1.2410954
		3	-1.19069655*	.04780685	.000	-1.3052805	-1.0761126
	2	1	-1.16140578*	.03324821	.000	-1.2410954	-1.0817161
		3	-2.35210233*	.04305650	.000	-2.4553006	-2.2489041
	3	1	1.19069655*	.04780685	.000	1.0761126	1.3052805
		2	2.35210233*	.04305650	.000	2.2489041	2.4553006
Zscore(clickOverSuccess)	1	2	1.05801241*	.02971141	.000	.9867998	1.1292250
		3	-1.46869639*	.04272136	.000	-1.5710914	-1.3663014
	2	1	-1.05801241*	.02971141	.000	-1.1292250	-.9867998
		3	-2.52670880*	.03847634	.000	-2.6189293	-2.4344883
	3	1	1.46869639*	.04272136	.000	1.3663014	1.5710914
		2	2.52670880*	.03847634	.000	2.4344883	2.6189293
Zscore(loyalty_segment)	1	2	.80975130*	.04778659	.000	.6952159	.9242867
		3	-.99625330*	.06871126	.000	-1.1609411	-.8315655
	2	1	-.80975130*	.04778659	.000	-.9242867	-.6952159
		3	-1.80600460*	.06188374	.000	-1.9543282	-1.6576810
	3	1	.99625330*	.06871126	.000	.8315655	1.1609411
		2	1.80600460*	.06188374	.000	1.6576810	1.9543282
Zscore(age)	1	2	-2.07647166*	.02503886	.000	-2.1364851	-2.0164583
		3	-.82790359*	.03600281	.000	-.9141955	-.7416117
	2	1	2.07647166*	.02503886	.000	2.0164583	2.1364851
		3	1.24856807*	.03242538	.000	1.1708506	1.3262855
	3	1	.82790359*	.03600281	.000	.7416117	.9141955
		2	-1.24856807*	.03242538	.000	-1.3262855	-1.1708506
Zscore(gender)	1	2	-.76609303*	.05632512	.000	-.9010937	-.6310924
		3	.19772882*	.08098862	.044	.0036145	.3918432
	2	1	.76609303*	.05632512	.000	.6310924	.9010937
		3	.96382185*	.07294116	.000	.7889957	1.1386480
	3	1	-.19772882*	.08098862	.044	-.3918432	-.0036145
		2	-.96382185*	.07294116	.000	-1.1386480	-.7889957

*. The mean difference is significant at the 0.05 level.

Fig. 8. Post hoc test results of the three solution model in cluster model B (related to Fig 6B)

Cluster Model C. The variables used in cluster C are customer age , gender, open rate, click rate, loyalty and part of day. Fig 9 shows the cluster results of this model in a two, three and four cluster solution.

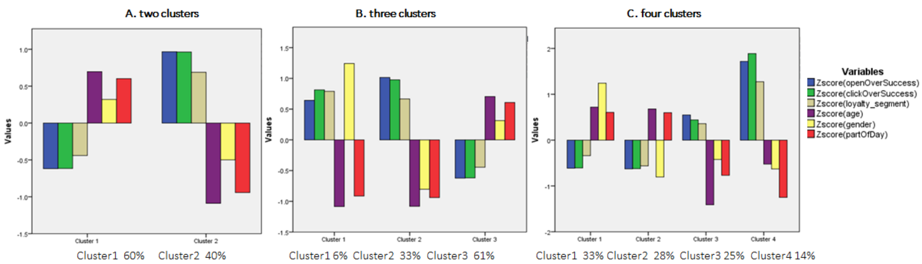

Fig. 9. Cluster Model C results

Fig 9A shows that the customers in cluster 2 of the two cluster solution have a higher percentages of open and click rates compared to the customers in cluster 1. In Fig 9B the customers in cluster 2 have the highest percentages of open and click rates compared to the two other clusters of this three cluster solution, whereas the size of cluster 1 is only 6% which indicates that the three solution model is in this case a weak model. Post hoc test also shows that. Fig 9C finally indicates that the customers in cluster 4 have the highest percentage of click and open rates compared to the other

clusters of the four cluster solution. Fig 10 shows a sample of post hoc tests table for the three solution model in cluster model C.

Dependent Variable	(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Zscore(openOverSuccess)	1	2	-.37282689*	.07281911	.000	-.5473605	-.1982933
		3	1.26521511*	.07016920	.000	1.0970329	1.4333973
	2	1	.37282689*	.07281911	.000	.1982933	.5473605
		3	1.63804200*	.03602205	.000	1.5517040	1.7243800
	3	1	-1.26521511*	.07016920	.000	-1.4333973	-1.0970329
		2	-1.63804200*	.03602205	.000	-1.7243800	-1.5517040
Zscore(loyalty_segment)	1	2	.12366425	.09627640	.598	-.1070920	.3544205
		3	1.23497676*	.09277286	.000	1.0126178	1.4573357
	2	1	-.12366425	.09627640	.598	-.3544205	.1070920
		3	1.11131251*	.04762586	.000	.9971624	1.2254627
	3	1	-1.23497676*	.09277286	.000	-1.4573357	-1.0126178
		2	-1.11131251*	.04762586	.000	-1.2254627	-.9971624
Zscore(age)	1	2	-.00367517	.05637447	1.000	-.1387941	.1314437

*. The mean difference is significant at the 0.05 level.

Fig. 10. Post hoc test results of the three cluster solution of cluster model C (Fig 9B)

The results show that loyalty in cluster 1 compared to cluster 2 is not statistically significantly different since the p-value = 0.598, whereas age in cluster 1 compared to cluster 2 are not significantly either at a p-value = 1. This indicates that the three cluster solution in Model C does not offer a good model for the variables of the model. Post hoc tests have been applied on the four solution model (9C) as well and the results show that open and click rates in cluster 1 are not significantly different compared to cluster 2, Fig 11 shows a sample of post hoc test of four cluster solution . Thus this model is not a good model fit solution either.

						95% Confidence Interval	
Dependent Variable	(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Zscore(openOverSuccess)	1	2	.01454489	.03598987	1.000	-.0805385	.1096283
		3	-1.16102197*	.03691244	.000	-1.2585428	-1.0635012
		4	-2.33156658*	.04458243	.000	-2.4493511	-2.2137821
	2	1	-.01454489	.03598987	1.000	-.1096283	.0805385
		3	-1.17556686*	.03856928	.000	-1.2774649	-1.0736688
		4	-2.34611147*	.04596361	.000	-2.4675450	-2.2246779
	3	1	1.16102197*	.03691244	.000	1.0635012	1.2585428
		2	1.17556686*	.03856928	.000	1.0736688	1.2774649
		4	-1.17054461*	.04668952	.000	-1.2938960	-1.0471933
	4	1	2.33156658*	.04458243	.000	2.2137821	2.4493511
		2	2.34611147*	.04596361	.000	2.2246779	2.4675450
		3	1.17054461*	.04668952	.000	1.0471933	1.2938960
Zscore(clickOverSuccess)	1	2	.01455230	.03233799	1.000	-.0708830	.0999876
		3	-1.04772202*	.03316696	.000	-1.1353474	-.9600966
		4	-2.49430684*	.04005867	.000	-2.6001398	-2.3884739
	2	1	-.01455230	.03233799	1.000	-.0999876	.0708830
		3	-1.06227432*	.03465567	.000	-1.1538328	-.9707158
		4	-2.50885914*	.04129970	.000	-2.6179709	-2.3997474
	3	1	1.04772202*	.03316696	.000	.9600966	1.1353474
		2	1.06227432*	.03465567	.000	.9707158	1.1538328
		4	-1.44658482*	.04195196	.000	-1.5574198	-1.3357499

*. The mean difference is significant at the 0.05 level.

Fig. 11. Post hoc test results of the four cluster solution of cluster model C (Fig 9C)

4.4. Major conclusions of the clustering analysis

When comparing the three cluster models in their three different forms (two, three and four cluster solution) with one another, we can observe that although all three models have a certain predictive value, the best fit with the demographic and behavioral characteristics of the respondents is obtained by model B with a three cluster solution. The variables used in this model are customer age, gender, open rate, click rate and loyalty. All variables in the three cluster solution of this model are significantly different as their p-values are lower than 0.05, whereas most of the other models and different cluster number solutions have higher p-values for the variables. Thus it is the best one in finding homogenous groups of customers with different loyalty based on customer demographic and behavioral characteristics.

5 Managerial implications of the results

The results of the logistic regression have already learned us that any e-commerce company should focus on the attributes that have a significant impact on customer loyalty when wanting to increase open and click rates of these customers.

If the company intends to increase the response rates of the customers that have low click rates, it should send relevant advertisement to the customers based on their demographic characteristics to avoid that customers unsubscribe from their website . Thus the results of K-mean cluster analysis recommend the web master team of the e-commerce company to use the variables used cluster model B (three cluster solution) more than the ones used in the other cluster models as this model promises the best cluster fit. This solution groups the customers into three clusters based on their demographic variables and interests. The sending process of the e-mails should consider the different cluster groups and at which time the customers usually open the e-mail campaigns in their e-mail. Fig 12 explains the different steps of the e-mail campaigns that we recommend the company to use.

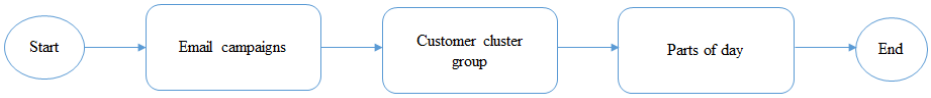

Fig. 12. The steps of the E-mail campaign process

The first step in this process is to know which content of the e-mail campaigns are relevant to the customer clusters, then the company will choose the best cluster to send the e-mail based on its demographic characteristics (age and gender) . Finally the web master team they should consider parts of day to send the e-mail to customers which should be the appropriate time that the customers normally open their e-mail campaigns.

This process based on our results can decrease the percentage of customers unsubscribing to the website since they will be receiving relevant e-mails campaigns. Using this model can also increase open and click rates when the company considers the sending time and the content of the e-mail campaigns carefully. Moreover it can help the company in increasing customer loyalty.

6 Conclusions and further research

In this paper, we have analyzed and examined the loyalty of receiving daily e-mail advertisements as a part of an e-mail marketing campaigns. The aim of this study was predicting customer loyalty and improving the response rate of e-mail campaigns, specifically open rate and click through rate, using data mining techniques.

First we have analyzed the impact of customers demographic and behavioral characteristics on customer loyalty using chi square and logistic regression techniques. Second we applied the k-means cluster algorithm to the data to find homogenous groups of customers based on their demographic and behavioral characteristics.

Results of our logistic regression research show that it is indeed possible to predict customer loyalty based on response rates of e-mail campaigns thus allowing web master teams of an e-commerce company to group customers in different segments. Clustering results show that identification of the profiles of these groups is possible on the basis of demographic and behavioral characteristics linked to click rates. Clustering results indeed show a good fit of one of the models with these characteristics. A process describing how to go about stepwise was depicted as well.

The future work will be applying predictive and descriptive models using decision tree analysis as a segmentation technique for each cluster we have obtained based on the customer demographic and behavioral characteristics. On the other hand customers buying behavioral should be addressed , that will help the company to attract the customers depending on their buying history by sending e-mails depending on the products which they like to buy.

References

1. Liu, Y., et al., What Drives Click-Through Rates of Tourism Product Advertisements on Group Buying Websites? Procedia Computer Science, 2015. 55: p. 221-230.
2. Carmona, C.J., et al., Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Systems with Applications, 2012. 39(12): p. 11243-11249.
3. Xuerui Wang, W.L., Ying Cui, Ruofei (Bruce) Zhang, Jianchang Mao, Click-Through Rate Estimation for Rare Events in Online Advertising, In : online multimedia advertising Techniques and Technologies, 2010: p. 1-12.
4. Cases, A.-S., et al., Web Site spill over to email campaigns: The role of privacy, trust and shoppers' attitudes. Journal of Business Research, 2010. 63(9–10): p. 993-999.
5. Shan, L., et al., Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. Electronic Commerce Research and Applications, 2016. 16: p. 30-42.

6. George Sammour , Benoit Depaire., Koen Vanhoof and Geert Wets, Identifying homogenous customer segments for risk email marketing experiments, in 11th International Conference on Enterprise Information Systems. 2009: milan , italy. p. 89-94.
7. Chiu, C.-Y., et al., An intelligent market segmentation system using k-means and particle swarm optimization. Expert Systems with Applications, 2009. 36(3, Part 1): p. 4558-4565.
8. Ngai, E.W.T., L. Xiu, and D.C.K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 2009. 36(2, Part 2): p. 2592-2602.
9. Michael J. A. Berry, G.S.L., Data mining techniques second edition – for marketing, sales, and customer relationship management. 2004, canada: wiley.
10. Devi, B.N., et al., Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce. Procedia Engineering, 2012. 30: p. 20-27.
11. Etzioni, O., The World-Wide Web: quagmire or gold mine? Commun. ACM, 1996. 39(11): p. 65-68.
12. Cooley, R., B. Mobasher, and J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1999. 1(1): p. 5-32.
13. Liu, B., Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). 2006: Springer-Verlag New York, Inc.
14. Kosala, R. and H. Blockeel, Web mining research: a survey. SIGKDD Explor. Newsl., 2000. 2(1): p. 1-15.
15. Srivastava, J., et al., Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 2000. 1(2): p. 12-23.
16. Joseph F. Hair, J., et al., Multivariate data analysis (4th ed.): with readings. 1995: Prentice-Hall, Inc. 745.
17. Fraley, C. and A.E. Raftery, Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 2002. 97(458): p. 611-631.
18. Sumaiya Thaseen, I. and C. Aswani Kumar, Intrusion detection model using fusion of chi-square feature selection and multi class SVM. Journal of King Saud University - Computer and Information Sciences.
19. Alvarez, S.A., Chi-squared computation for association rules: preliminary results 2003, Computer Science Department, Boston College: Boston
20. Robert P Burns , R.B., Business Research Methods and Statistics Using SPSS. 2008, london: SAGE Publications.
21. Lopes, P. and B. Roy, Dynamic Recommendation System Using Web Usage Mining for E-commerce Users. Procedia Computer Science, 2015. 45: p. 60-69.
22. Aziz Yarahmadi, Mathijs Creemers, Hamzah Qabbaah, Koen Vanhoof, UNRAVLING BILINGUAL MULTI-FEATURE BASED TEXT CLASSIFICATION: A CASE STUDY. International Journal “Information Theories and Applications, 2017. 24.