DECISION TREE ANALYSIS TO IMPROVE E-MAIL MARKETING CAMPAIGNS

Hamzah Qabbaah, George Sammour, Koen Vanhoof

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. Decision tree analysis 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 decision tree analysis such as CHAID , CART and QUIST.*

The methodology used in this study is Cross Industry Standard Process for Data Mining (CRISP – DM) methodology. The models are trained and tested using split sample validation. Furthermore, we used a classification measures to calculate the accuracy, precision, recall and F1 to evaluate the models. The models reported satisfactory results in predicting customer loyalty based on open rate, click through rate values and on customer demographic variables. The response rates also increase at the preferred moment at which e-mails should be send to customers in email campaigns.

Keywords: *E-Commerce, E-mail campaigns, CRISP-DM, Decision tree, open rates, click through rates.*

1. Introduction

The rapid development and popularization of internet use, has created an extraordinary growth of e-commerce and online shopping (Liu et al. 2015). We understand e-commerce to be the buying and selling of products or services through electronic media (Carmona et al. 2012). One important element of e-commerce practices is web advertising. Companies can address customers directly using different channels such as e-mail campaigns and contextual advertising (Xuerui Wang 2010).

The appeal of e-mail communication to talk directly to customers is double: cost effectiveness and time efficiency (Cases et al. 2010). In order to reach these objectives, companies wanting to use e-mail as a direct communication channel with their customers, need to thoroughly understand how e-mail campaigns may affect customers' attitudes and behaviour (Shan et al. 2016). This knowledge can lead to a more optimal design of their e-mail campaigns and thus be turned into a competitive advantage.

This subject has attracted much attention in recent years. E-mail marketing research studies have been conducted amongst others by online surveys, in-depth interviews, controlled experiments and by tracking customer behaviour. Click-through links and the visiting patterns have been used as descriptive variables but few of the studies have really investigated the effects of e-mail characteristics on customer attitudes and behavioural intentions (George Sammour 2009).

Loyalty and response rates have been an important focus of attention in the research on e-mail campaigns (i.e., more than one email sent by a company) (Cases et al. 2010). Building models to improve response rates by using individual preferences of customers to personalize e-mail newsletters is the major topic of this research stream. The reason of this is obvious : marketing campaigns and products can on the basis of the results of this research be customized to have a more effective appeal to groups of customers or to individuals (George Sammour 2009).

These data are in most companies abundantly available. They are a very valuable resource when efficient access to the data is created and salient information is extracted from them. Making efficient use of the information has thus become an urgent need. Data mining is an excellent tool to extract or detect hidden customer characteristics and behaviours from large databases (Ngai, Xiu, and Chau 2009; Chiu et al. 2009). It can be 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" (Michael J. A. Berry 2004). For many organizations, the goals of data mining essentially include improving marketing capabilities, detecting abnormal patterns, and predicting the future based on past experiences and current trends.

Web mining is the use of data mining techniques to automatically discover and extract knowledge from data available on the use of websites (Etzioni 1996). The importance of web mining is considering the behaviour and preferences of the user of websites(Cooley, Mobasher, and Srivastava 1999). 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) Analysing the obtained patterns(Liu 2006; Kosala and Blockeel 2000). This research analysis uses web usage mining, which is defined as "The process of applying data mining techniques to the discovery of usage patterns from Web data" (Srivastava et al. 2000).

2. Problem statement and Research questions

Customer loyalty is a complex phenomenon. Most research has focused on trying to look at it as a factor in purchase behaviour or as a final action by the customer. The link between customer loyalty and the steps in buying behaviour preceding the purchase, like e-mail behaviour have not been the focus of much attention. Our goal is to look at the link between this pair of variables, namely by researching into the impact of customer demographic and customer e-mail behaviour on customer loyalty and to see whether it can be used as a predictor of loyalty.

From this, we can derive our research questions. They are:

1. What is the effect of some descriptive variables of the e-mail behaviour of customers on customer loyalty?
2. Which model predicts customer loyalty best on the basis of both the customer demographic and customer e-mail behaviour variables ?

Next section we describe the methodology we used to answer the research questions.

3. Methodology

The data mining methodology used in this research was Cross Industry Standard Process for Data Mining (CRISP – DM) and the methodological steps were followed: business understanding, data understanding, data preparation, modeling, evaluation and deployment, as shown in figure 1.

- **Business understanding**

Email marketing for E-Commerce companies focusses on sending relevant emails to the right customers at the right time. It has thus become a “widely-used marketing tool for companies to communicate with customers, handle customer complaints, and cross-sell and up-sell products”. (Zhang 2015) The goal of an email marketing is consequently to generate profit by making customers become more active in purchases, taken into consideration that the customer is always entirely free to opt-out.

E-mail marketing is thus very closely linked to increasing customer loyalty. Loyalty has been used in a business context, to describe a customer’s willingness to continue buying from a company over the long term, preferably on an exclusive basis, and recommending the company’s products to friends and associates (Lovelock and Wirtz 2011). Reichheld and Jr suggest that "customer repeat purchase loyalty must be the yardstick of success" (F. Reichheld and Jr. W. E 1990). Customer loyalty is indeed widely seen as a key determinant of a firm’s profitability (Reinartz and Kumar 2002). The increased profit from loyalty comes from reduced marketing costs, increased sales, and marketing cost (T. and Shiang-Lih 2001). Loyalty reduces the need to incur customer acquisition costs (Reichheld and Teal 1996).

The common important efficiency metrics of e-mail marketing are thus : Delivery rate, Open rate and Click-through rate. Increased rates lead to an improve customer buying process, increased loyalty and increased firms profits.(Stokes 2011)

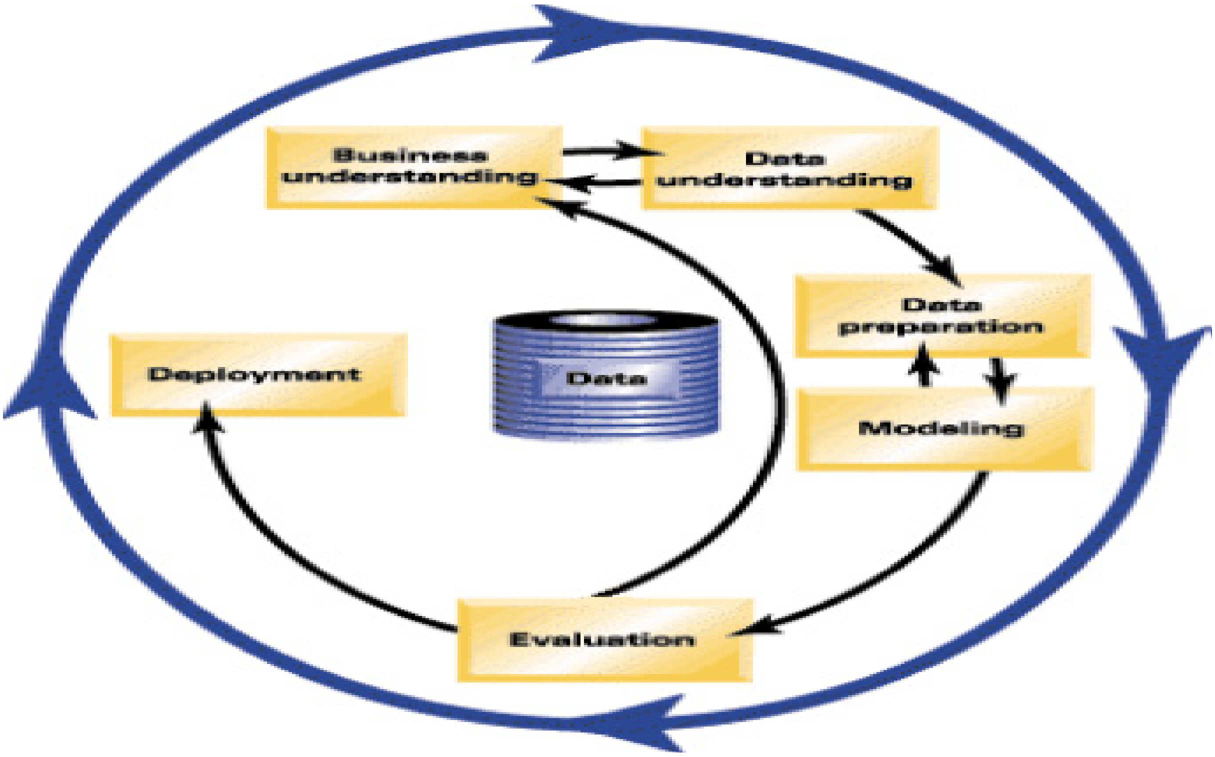

Figure 1. CRISP –DM life cycle.

- **Data understanding and Data Preparation**

Data have been obtained using the webmaster tool of Google Analytics from an E-commerce website. Cleaning, merging 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 , total e-mails received, total e-mails opened , total e-mails clicked, loyalty segment2 and the time of the campaign mail was opened. After that we calculated the click rate, open rate and the average time of opening the campaign for each customer. Table one shows the data type of the variables that are used in our study.

The company offers a loyalty program to their customers: loyal customers receive numerous additional benefits. Customers can buy a loyalty membership or get it through some consumption formula. For privacy reasons the company did not explain it, but they clarified that one of the important key factors affecting the loyalty program is collecting points by purchasing their products online using their website. Both the number of purchases and the value of the purchases play a role. Therefore, none of our independent variables are in the formula of the loyalty programme membership.

Since our goal is to predict customer loyalty, we analysed the impact of customer demographics and customer e-mail behaviour characteristics on customer loyalty. Consequently, the independent variables in our analysis are the customer demographics (determined by age, gender and country of residence) and the customer e-mail behaviour (determined by the parts of the day and the open and click through rates. The dependent variable is the customer loyalty segment.

Table 1. Study variables.

Variable	**Data type**
Customer ID	Continuous
Gender	Categorical
Age	Continuous
Country	Categorical
Open rate	Continuous
Click through rate(CTR)	Continuous
Loyalty segment2	Categorical
Part of the day(POD)	Categorical

- **Modeling**

Our research questions indicate that the main objective of this research is to find a way of predicting customer loyalty based on their response pattern to e-mail campaigns and linking that to their demographic and behavioural characteristics.

This goal can be achieved by experimenting with different campaigns and comparing the resulting customer e-mail behaviour to them. These data must then be linked to demographic characteristics of the respondents. This is however very difficult to realize since companies are reluctant to experiment with real life campaigns. They indeed fear that the response rates will drop due to the experiment itself.[6] This research is therefore based on the use of decision tree technique of the web data that allow to experiment with campaigns promising a high potential for increased response rate levels .

Decision tree analysis as a method of data mining techniques allows to achieve the major steps needed to achieve the required objectives of analysing and predicting models that measure the impact of customers demographic and behavioural characteristics to customer's loyalty .[6] Sections 4 and 5.1 will explain the decision tree algorithms used in this study and present the analysis and results.

- **Evaluation and Deployment**

The confusion matrix used in this study to evaluate the models by determining the classification measures: Accuracy, precision, recall and F1 of these models." (Carmona et al. 2012).

Table 2. shows the confusion matrix. Where TP is defined as normal behaviour that is correctly predicted, FP indicates normal behaviour 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 (Sumaiya Thaseen and Aswani Kumar ; Lopes and Roy 2015).

Table 2. 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 1 to 4 (Aziz Yarahmadi 2017).

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.

$$\text{Accuracy} = TP+TN /(TP+TN+FP+FN) \quad (1)$$

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

$$\text{Recall} = TP / (TP+FN) \quad (2)$$

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.

$$\text{Precision} = TP / (TP+FP) \quad (3)$$

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

$$F_1 = 2^* ((\text{Recall}^*\text{Precision})/(\text{Recall}+\text{Precision})) \quad (4)$$

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

Managerial Consequences and findings are shown in section 6. Next section we describe the Decision tree algorithms used in this study.

4. Decision tree analysis

Decision tree analysis is one of data mining techniques that can encompass numerous algorithms to predict a dependent variable. These predictions are determined by the influence of independent predictor variables. It is a structure that can be used to divide a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another. Decision trees thus use a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. (Linoff and Berry 2011).

Each procedure goes as follows. After the population is split into *n* nodes, the process is repeated on each node until the data is completely partitioned out or until a stopping rule is met. This means that there are no longer enough cases to partition out, or only one case remains for the last node. The entire classification of data is then published through a graphical tree, which explains the interaction of *x* (*x1, x2, …, xm*) and *y*. An ideal decision tree is one that has no limitations; this tree would correctly classify each individual case. Since this is impossible, the classification procedures will attempt to provide as much separation as possible in regards to classifying data into correct cases (Long et al. 1993). Graphically, decision trees will produce a tree *T* which is comprised of a root node and child nodes (De'ath 2000) (Ripley and Hjort 1995). The tree is essentially a connected graphic that is inverted; however, the tree display is user specific and can be displayed horizontally (Loh and Shih 1997; L. and Jr 2012)

There are numerous decision tree classification procedures. The algorithms used in this paper are CHAID, CART and QUEST. A brief description of these techniques is added.

4.1 The Chi-Squared Automatic Interaction Detection (CHAID) algorithm

CHAID is the most popular classification procedure because it can manage both continuous and categorical variables (Díaz-Pérez and Bethencourt-Cejas

2016). The algorithm has the ability to choose different statistical techniques for splitting, determined by the level of measurement for the dependent variable. Since the algorithm can handle continuous or categorical dependents, multiple statistical techniques are available. If *the* dependant variable is continuous, an F-Test will be selected for splitting. In contrast, if *it* is either ordinal or nominal then a maximum likelihood estimate is selected. The user has the ability to select Pearson-Chi when *the variable* is nominal (L. and Jr 2012).

In regards to splitting, CHIAD is not bound to binary splits and allows for multiple level splits at each node. The algorithm searches for the best possible split for the different values of the independent variable then determines if the data should be merged to form a node or split in the data. Essentially, the best predictor is selected to begin splitting the data. During this process the CHAID is selecting the best predictor based on comparisons of the adjusted p-value for each predictor (L. and Jr 2012). This is a three step process for determining splitting. The splitting and merging procedures for CHAID are independent of one another. The algorithm uses as sequential process where splitting and merging occur at the same time. The basis behind this aspect of the algorithm is rooted in the computation time of *T*. CHAID has difficulty determining when and where to stop (Loh and Shih 1997). Stopping is also a three step process determined by the chi-square or F-Test statistic. As such, stopping occurs when the critical level of the selected test statistic (chi-square or F) fails to meet the test.

The Independent variable statistics in the tree display of CHIAD algorithm shown are *F* value (for scale dependent variables) or chi-square value (for categorical dependent variables) as well as significance value and degrees of freedom.(IBM 2016)

4.2 Classification and regression trees (CART/C&RT)

This algorithm is a decision tree program very well introduced in the statistic community (Long et al. 1993). It produces a binary decision tree by restricting the partitions at each node to two (G. T. Denison, K. Mallick, and F. M. Smith

1998). CART algorithm uses an extensive and exhaustive search of all possible univariate splits to determine how the data have to be split for the classification tree (Breiman et al. 1984). Partitioning will continue until the algorithm is unable to produce mutually exclusive and homogenous groups further (De'ath 2000; G. T. Denison, K. Mallick, and F. M. Smith 1998). At this point, the node is considered a terminal node.

CART can account for missing values by estimating a prediction based on other independent factors or "surrogates" (Breiman et al. 1984). The CART algorithm thus allows for the inclusion of partial data. Therefore missing values or cases are not removed and bias is reduced(De'ath 2000). CART handles these missing values for predictor variables in two ways. First, if the predictor variable has missing values, it can be restricted from moving farther down the tree. Second, it can send them farther down the tree (De'ath 2000). If the predictor variable is stopped, the cases with missing values will be recorded with response values as the rest of the cases in that node. If the predictor variable with missing cases is sent further down the tree, its missing values will be replaced by responses from another independent factor with non-missing cases. CART will determine which method produces the best agreement with the splitting variable. Since CART is a forward stepwise method, data that may not be influential in predicting the dependant variable can however be included in the tree leading to extensive and lengthy trees since the algorithm uses an exhaustive search to determine the splits in the data (L. and Jr 2012).

The Independent variable statistic in the tree display of CART algorithm shown is improvement value . The minimum decrease in impurity required to split a node is 0.0001. Higher values tend to produce trees with fewer nodes, this called Minimum change in improvement. To measure impurity and the minimum decrease in impurity required to split nodes. In this for categorical (nominal, ordinal) dependent variables, the impurity measure selected is: Gini. Splits are found that maximize the homogeneity of child nodes with respect to the value of the dependent variable. Gini is based on squared probabilities of membership for each category of the dependent variable. It reaches its minimum (zero) when all cases in a node fall into a single category. This is the default measure.(IBM 2016)

4.3 Quick unbiased efficient statistical trees (QUEST)

QUEST is a binary splitting algorithm for building decision trees. The algorithm is essentially a complex and technical version of CART. While the two algorithms share commonalities, the major difference is the speed in which they are produced (L. and Jr 2012).

The most significant advantage in choosing QUEST over another algorithm lies in the speed of computation, QUEST uses to reduce the processing time required for large CART Tree analyses. However, while there is a significant increase in speed, accuracy in prediction is not reduced, Sometimes QUEST is better and other times other algorithms is better (Loh and Shih 1997).The algorithm contains no identified bias in variable selection for the splitting. The lack of bias in selection is controlled by the algorithm because it handles the variation of levels in *the independent variables x*. Therefore, QUEST can control for the inclusions of $x_1, x_2, \ldots, x_m$ when variables have variations in levels.

There are two steps within the splitting aspect of the QUEST algorithm. These steps are dependent upon the significance test to split each node (Loh and Shih 1997). The first step of the algorithm examines the association of each predictor variable (X_i) with respect to the dependent variable y. Then the most significant variable is selected. After this step is finished, the algorithm conducts an exhaustive search for set *S*, which is the subset of values taken by the predictor variable in the node above. QUEST does not conduct exhaustive searches unless the variable is unordered and has a small number of values. QUEST is significantly better when searching for variables because selection bias and computational cost were not present, but the accuracy is identical to the other algorithms (Loh and Shih 1997; L. and Jr 2012).

The Independent variable statistics in the tree display of QUEST algorithm shown are F,

significance value, and degrees of freedom for scale and ordinal independent variables; and for nominal independent variables, chi-square, significance value, and degrees of freedom are shown.(IBM 2016)

The analysis and results of the use of these methodologies will be presented in section 5.

5. Analysis and Results

The results of the decision tree analysis are presented in section 5.1 which we present three models according to which we have tried to predict customer loyalty using the three decision tree algorithms , Next on section 5.2 we present the evaluation of the models using classification measures tests.

5.1 Prediction models of customer loyalty using decision tree analysis

In this section we create a three models using three decision tree algorithms (CHAID, CART and QUEST) to predict customer loyalty based on our independent variables. They can be described as follows:

Model 1: Customer loyalty based on customer demographic variables.

Model 2: Customer loyalty based on customer e-mail behaviour variables

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

For each decision tree we have used the **split sample validation**, the data were split into two sets: a training set of 80%, and a testing set of 20 %. The model trained will be tested on the test data to verify whether they give similar results. The total number of customers present in each node is based on the loyalty variable (category "Not loyal" or category "Loyal”). Then each time, the rules extracted from the tree are mentioned.

Since three decision trees were developed for three different models, in total nine analysis sets are presented.

1. Decision trees for Model 1 (predicting loyalty based on customer demographic variables)

The decision tree using the CHAID method is presented in Figure 2.

The rules explaining the nodes in Figure 2. are as follows :

Node 1:If (Age <=45) then class-loyal is (81.7%) and class-non loyal is 18.3%.

Node 3:If (Age>45) and (gender is male) then class-loyal is 20.7% and class non loyal is 79.3%.

Node 4: If (Age>45) and (gender is female) then class loyal is 15.3 % and class non loyal is 84.7%

It is clear that according to the CHAID model the country of residence does not play an important role in predicting loyalty as it does not figure in the decision rules at the different nodes. The strongest independent variable predicts loyalty segment is age since X^2=385 (df=1, p-value<0.01).

The decision tree using the CART method is presented in Figure 3.

The nodes in Figure 3 can be explained by the following rules :

Node 1: If (Age<=45.5) then class loyal is 88.6 % and class non loyal is 11.4%

Node 4:If (Age>45.5) and (gender is female) then class loyal is 16.9% and class non loyal is 83.1%

Node 6 : If (Age>63.5) and (gender is male) then class loyal is 33.3% and class non loyal is 66.7%.

Node 7: If (45.5<Age<55.5) and (gender is male) then class loyal is 33.3% and class non loyal is 66.7%.

Node 8: if (63.5>age >55.5) and (gender is male) then class loyal is 24.2% and class non loyal is 75.8%.

Again it appears as if the country of residence plays a less important role than the other demographic variables in explaining the nodes of this CART decision tree. The strongest independent variable predicts loyalty segment is also age since the improvement value is 0.163 .

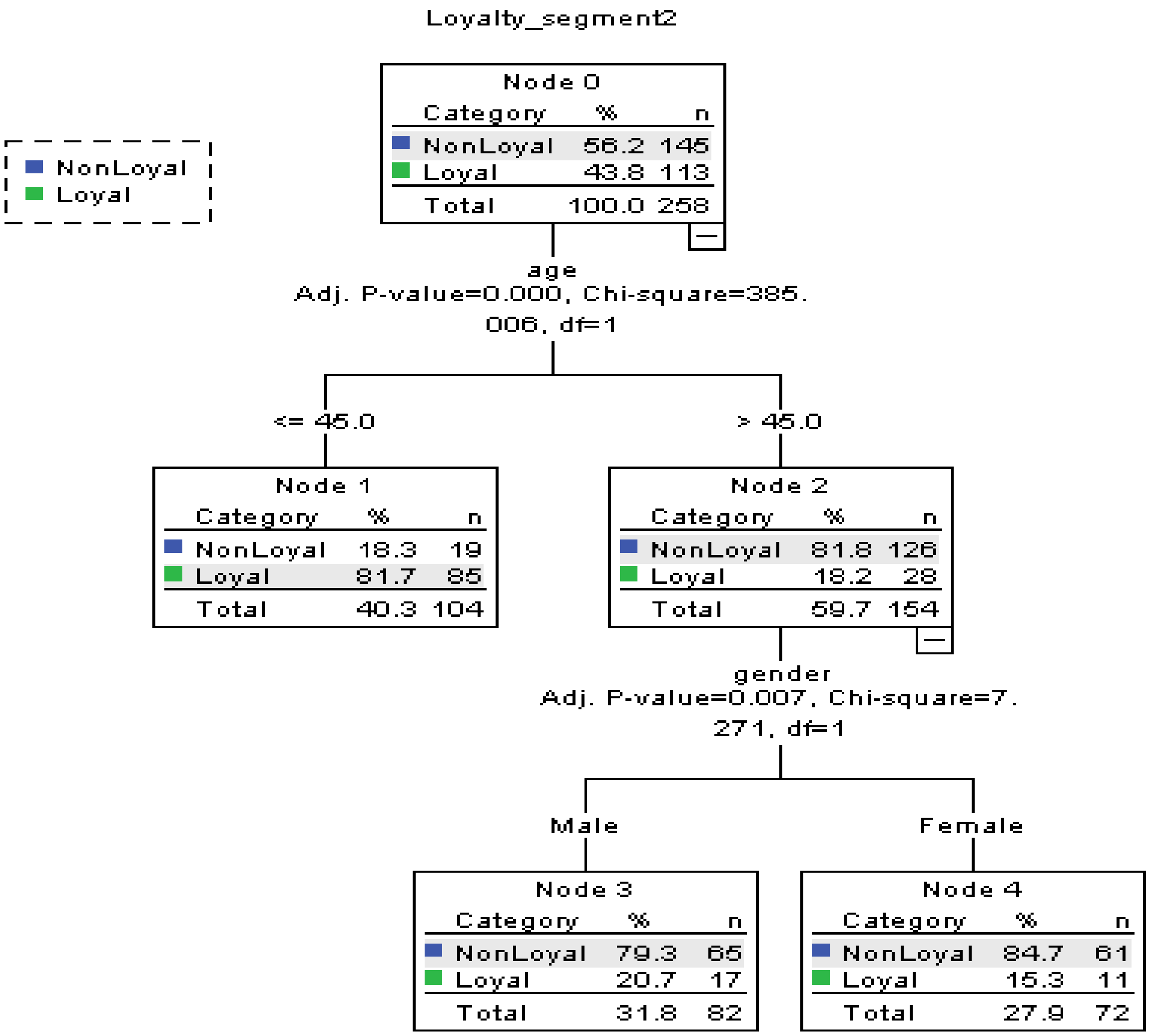

Figure 2. CHAID decision tree for Model 1 (using customer demographic variables)

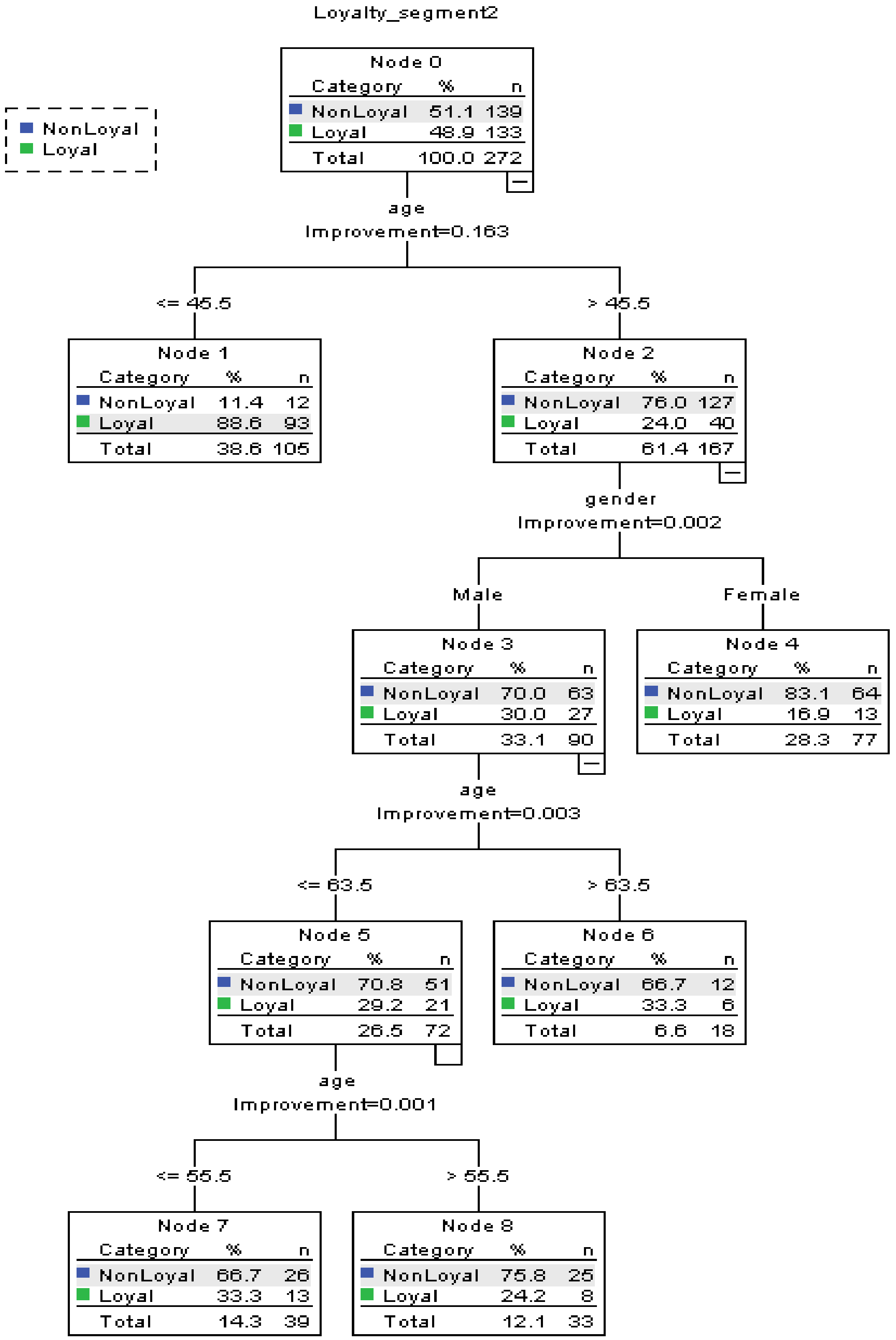

Figure 3. CART decision tree for Model 1 (using customer demographic variables)

The decision tree using the QUEST model is presented in Figure 4. The rules explaining the decision tree in Figure 4 are fairly simple. They read as follows :

Node 1: if (age<=43.5) then class loyal = 86.2% and class non loyal = 13.8%.

Node 2: if (age >43.5) then class loyal = 26.5% and class non loyal = 73.5%

Using this model both gender and country of residence do not seem to play an important role in prediction loyalty of the customers. While Age is the only independent variables effects loyalty segment based on QUEST method in this model, since F=471.2 and P-value<0.01.

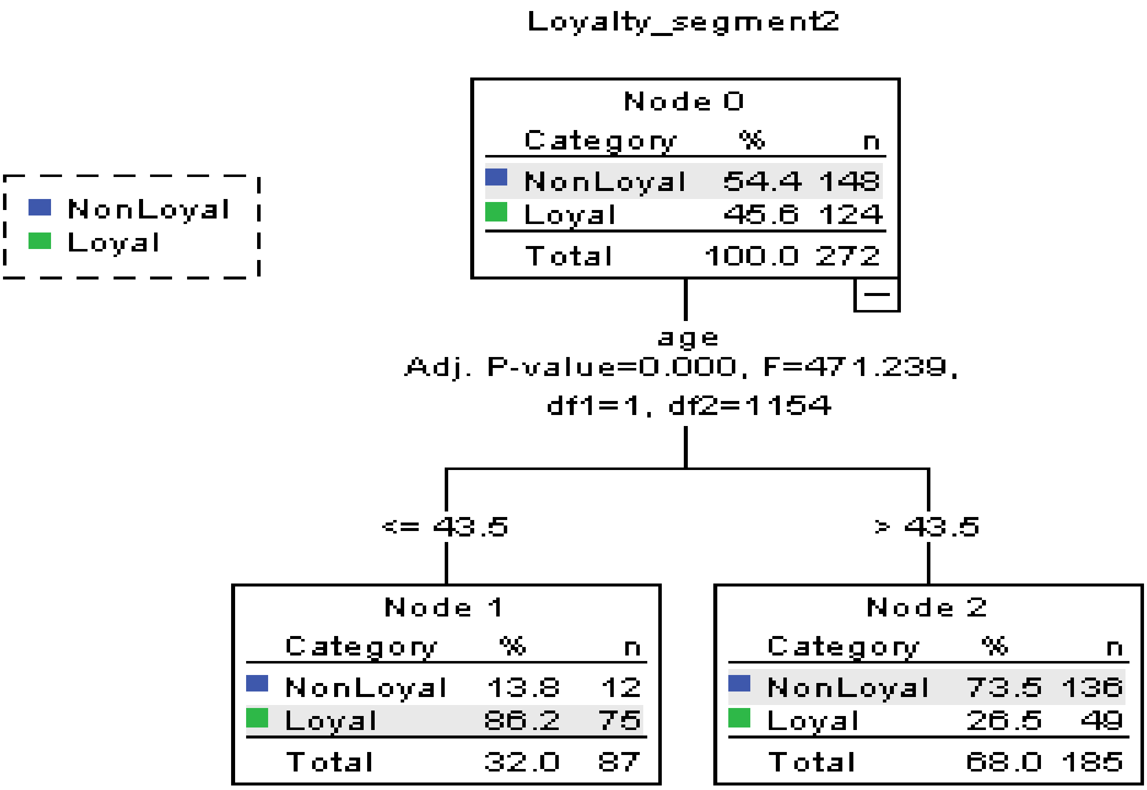

Figure 4. QUEST decision tree for Model 1 (using customer demographic variables)

2. Decision trees for Model 2 (predicting loyalty based on customer e-mail behaviour variables)

The decision tree using the CHAID method is presented in Figure 5.

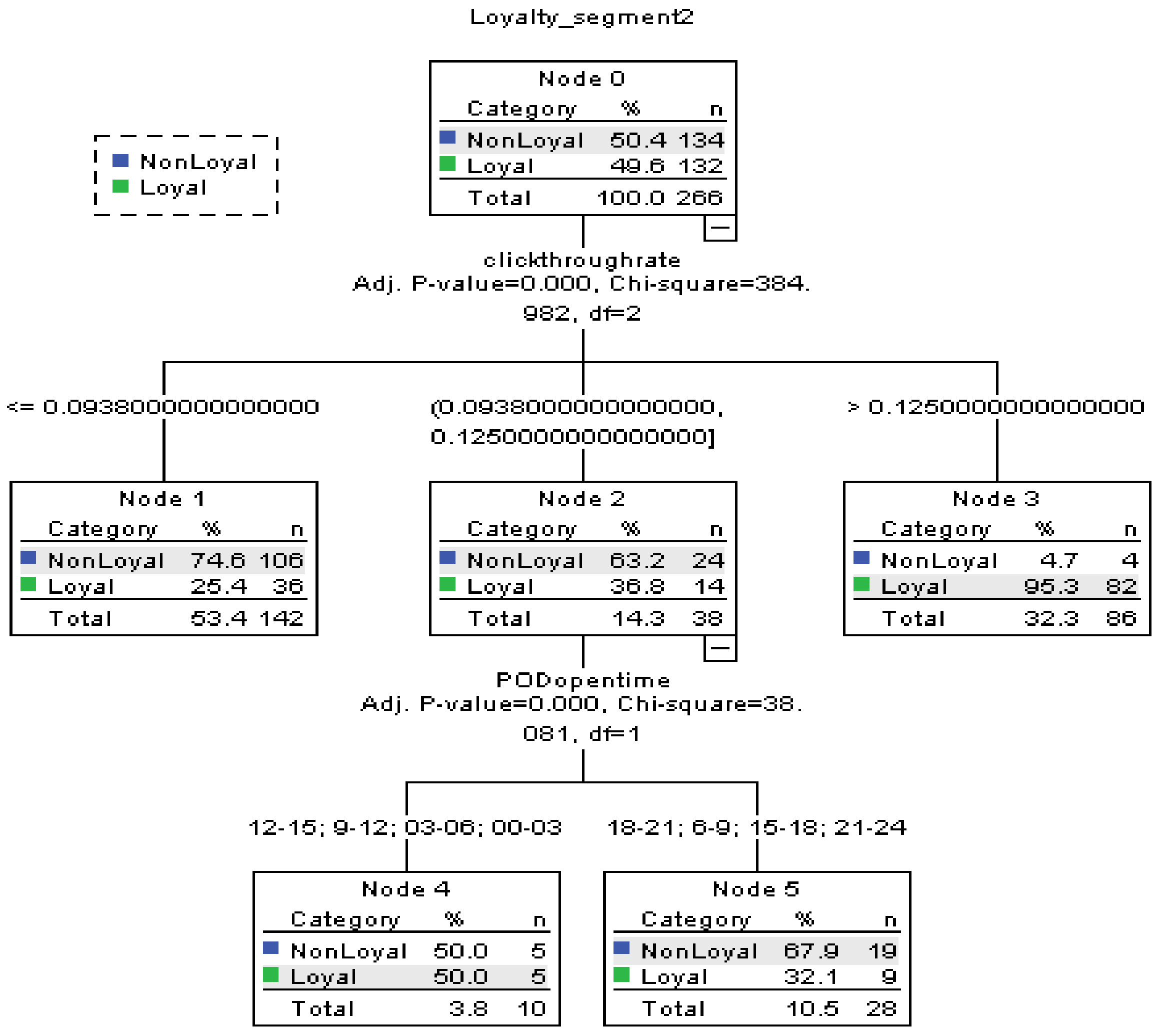

Figure 5. CHAID decision tree for Model 2 (using customer e-mail behaviour variables)

The CHAID method is somewhat more complicated for this model when the rules of the different nodes are read completely. They are :

Node 1 : if (click-through rate (CTR) < 0.094) then class loyal = 25.4% and class non loyal = 74.6%.

Node 3: if (CTR > 0.125) then class loyal = 95.3% and class non loyal = 4.7%.

Node 4: if (0.094< CTR < 0.125) And (POD open time on [12-15] OR[9-12] OR[03-06] OR [00-03]) then class loyal =50% and class non loyal = 50%.

Node 5: if (0.094< CTR < 0.125) And (POD open time on [18-21] OR[6-9] OR[15-18] OR [21-24]) then class loyal =32.1% and class non loyal = 67.9%.

According to CHAID method in this model the strongest independent variable predicts loyalty segment is CTR since X^2=384 (df=2, p-value<0.01).

The decision tree using the CART method is presented in Figure 6.

The rules that guide the decision tree presented in Figure 6 are as follows :

Node 4: If (CTR<=0.14) and (open rate>0.55) then class loyal = 55.6% and class non loyal = 44.4%.

Node 5: If (CTR>0.14) and (POD open time between [9-12] OR [18-21] OR [21-24]) then class loyal = 93.2% and class non loyal = 6.8%.

Node 6: If (CTR>0.14) and (POD open time between [12-15] OR [15-18] OR [6-9] OR [00-03] OR [03-06]) then class loyal = 87.2% and class non loyal = 12.8%.

Node 8: if (0.14>CTR> 0.11) and (open rate<= 0.55) then class loyal = 31.6% and class non loyal = 68.4%

Node 10: if (CTR< 0.11) and (open rate<= 0.55) and (POD open time between [15-18] or [21-24] or [6-9] or [00-03] or [03-06]) then class loyal = 27% and class non loyal = 73%.

Node 11: if (CTR< 0.11) and (open rate<= 0.55) and (POD open time between [12-15] or [9-12]) then class loyal = 16.7% and class non loyal = 83.3%.

Node 12: if (CTR< 0.11) and (open rate<= 0.55) and (POD open time between [18-21]) then class loyal = 25% and class non loyal = 75%. Contrary to the CHAID method and the CART method clearly also uses the CTR (click-through rate) in developing the tree model. since the improvement value of the first split was CTR =0.157 .

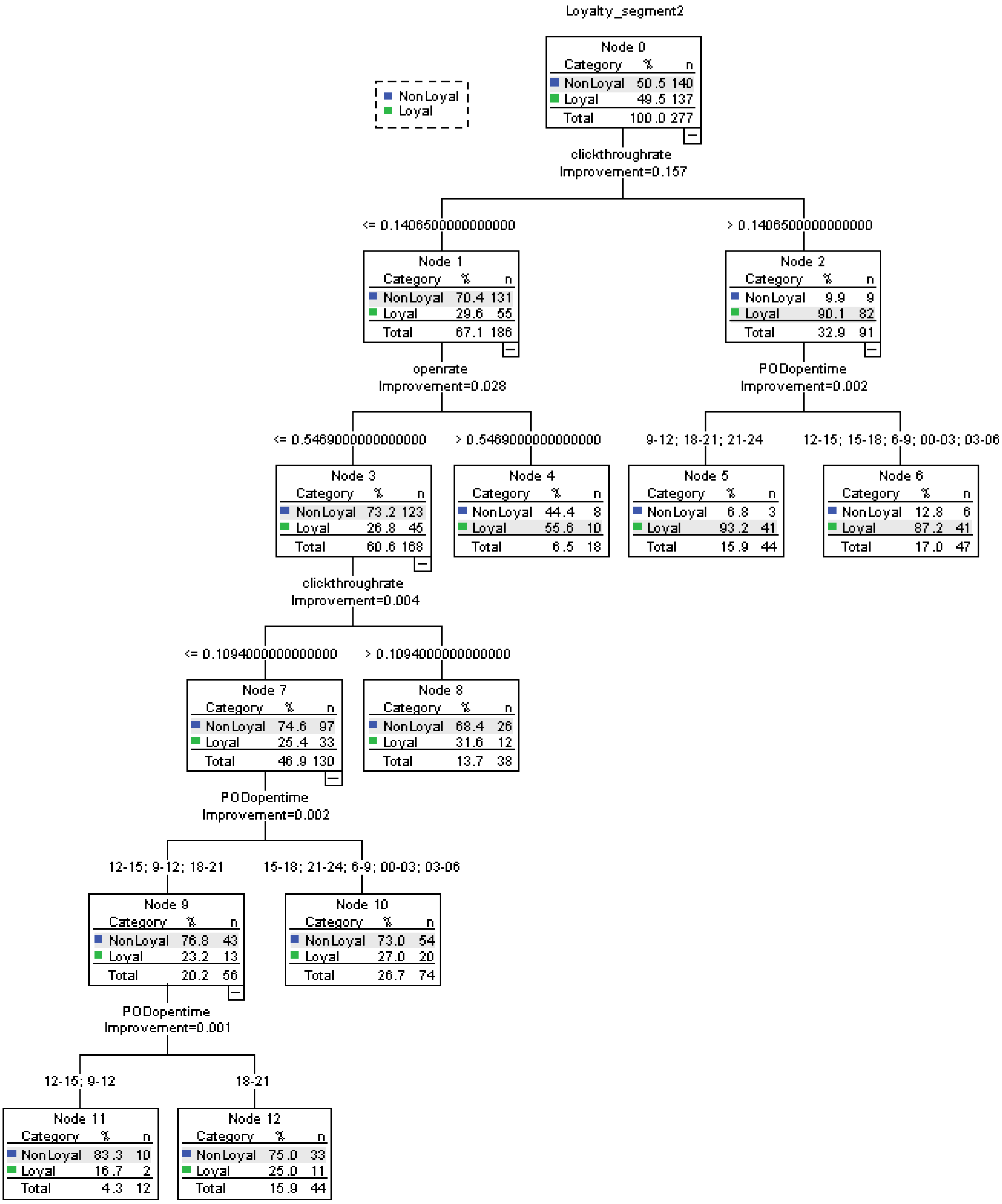

Figure 6. CART decision tree for Model 2 (using customer e-mail behaviour variables)

The decision tree using the QUEST method is presented in Figure 7.

The rules explaining the nodes of this QUEST method analysis are simple, Since only CTR effects loyalty segment based on QUEST method in this model, the F=455.8 and P-value <0.01. The rules are:

Node 1: if (CTR <= 0.155) then class loyal = 31.9% and class non loyal = 68.1%.

Node 2: if (CTR>0.155) then class loyal is 84.7% and class non loyal = 15.3%.

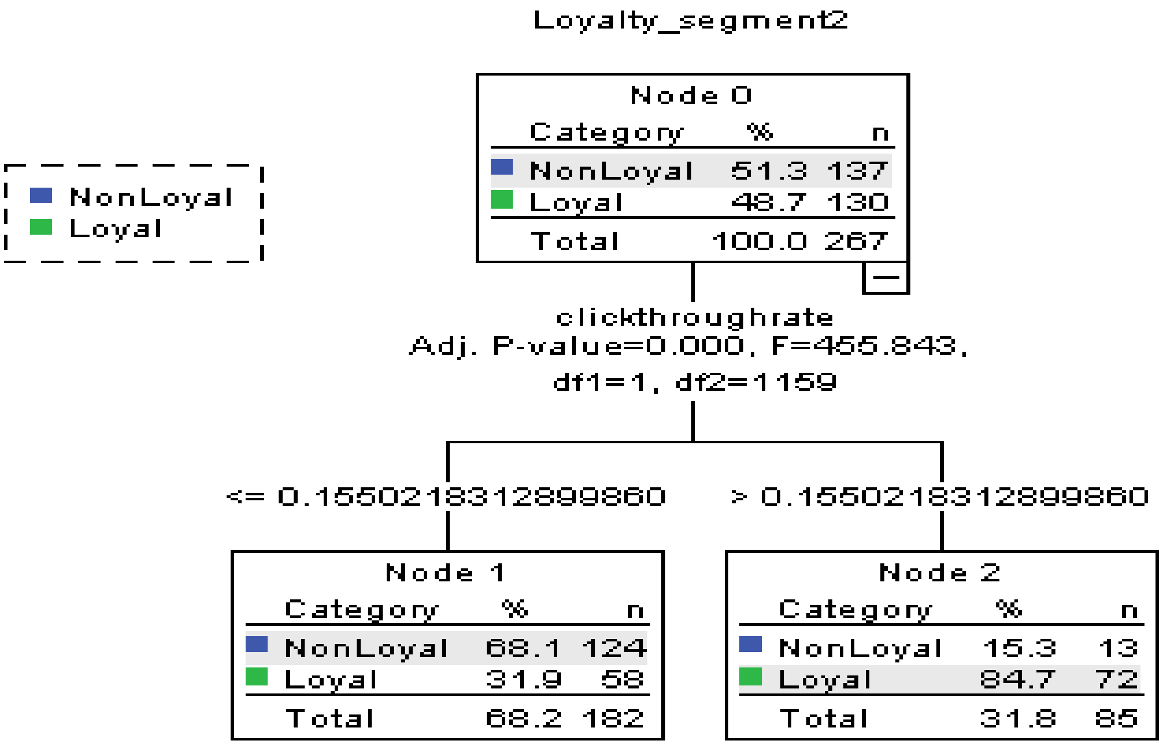

Figure 7. QUEST decision tree for Model 2 (using customer e-mail behaviour variables)

3. Decision trees for Model 3 (predicting loyalty based on age, gender, open rate, click-through rate and Parts of the day)

The decision tree using the CHAID method is presented in Figure 8.

The rules explaining the nodes in Figure 8 are as follows :

Node 3: If (age <=44) and (CTR<=0.22) then class loyal = 78.8% and class non loyal = 21.2%.

Node 4: If (age<=44) and (CTR>0.22) then class loyal = 93.2% and class non loyal = 6.8%.

Node 7: If (44<age<=57) and (POD open time between [12-15] or [18-21] or [21-24]) then class loyal = 23.2% and class non loyal = 76.8%.

Node 8: If (age >57) and (POD open time between [12-15] or [18-21] or [21-24]) then class loyal = 24.2% and class non loyal = 75.8%.

Node 9: If (POD open time between [9-12] or [6-9] or [15-18] or [03-06] or [00-03]) and (open rate<=0.44) and (age >44) then class loyal = 25.6% and class non loyal = 74.4%.

Node 10: If (POD open time between [9-12] or [6-9] or [15-18] or [03-06] or [00-03]) and (open rate>0.44) and (age>44) then class loyal = 33.3% and class non loyal = 66.7%.

Age is the strongest independent variable effects loyalty segment in this model since the first split of the tree is based on it , the X^2 is 374 (df=1, p-value<0.01). CTR, open rate and POD open time variables also play an important role in predicting loyalty in this model.

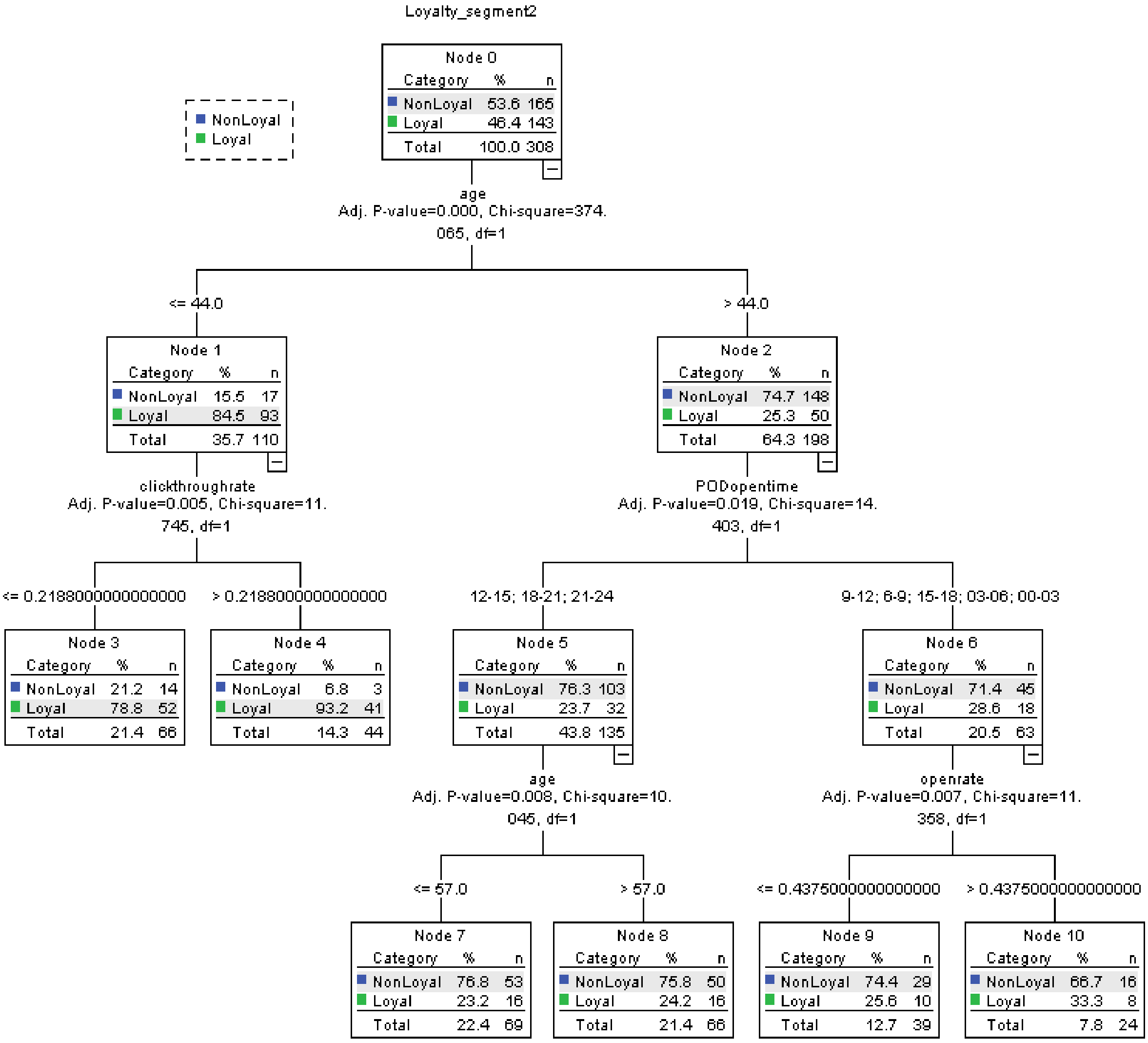

Figure 8. CHAID decision tree for Model 3

The decision tree using the CART method is presented in Figure 9.

The rules explaining the nodes of this CART method are very simple comparing to the model 1 and model 2, the nodes are based only on the age variable, and the improvement value was 0.166 .

The rules are:

Node 1: If (age<45.5) then class loyal = 86% and class non loyal = 14%.

Node 2: If (age >=45.5) then class loyal =23.4% and class non loyal = 76.6%.

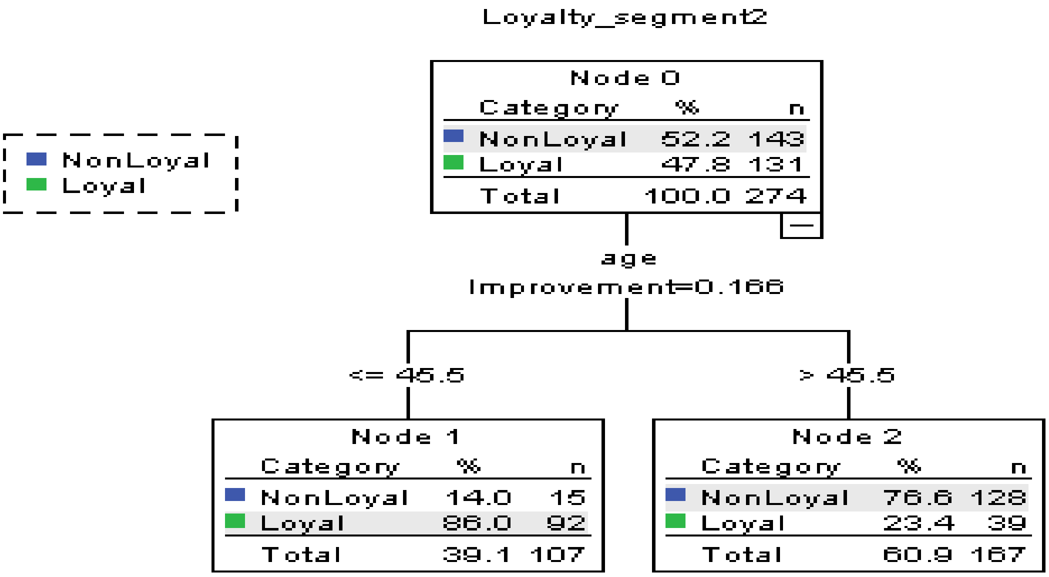

Figure 9. CART decision tree for Model 3

The decision tree using the QUEST method is presented in Figure 10.

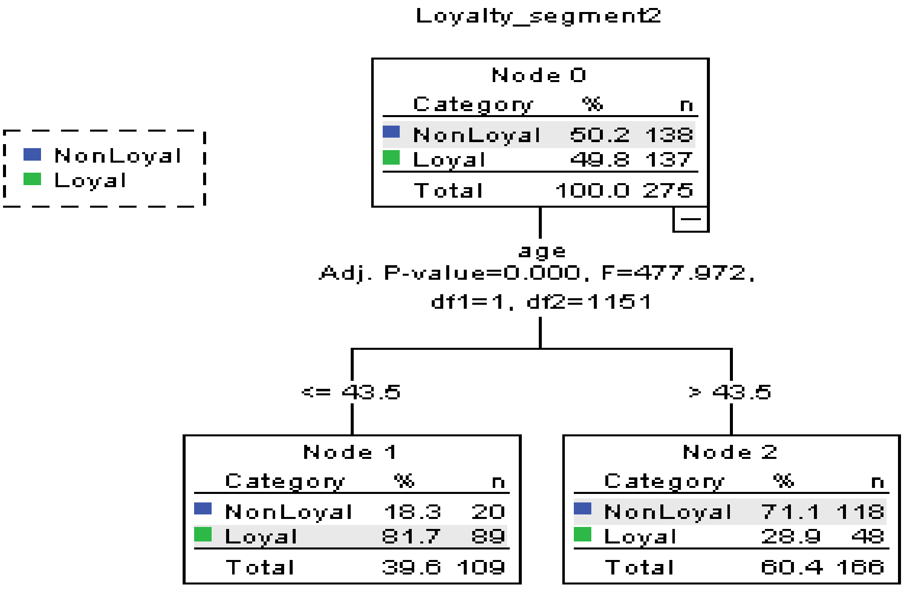

Figure 10. QUEST decision tree for Model 3

The rules explaining the nodes of this QUEST method are simple as the rules extracted from the model 1 and model 2, the nodes are based only on the age variable, since F=478, and P-value<0.01. The rules are:

Node 1: If (age<43.5) then class loyal = 81.7% and class non loyal = 18.3%.

Node 2: If (age >=43.5) then class loyal =28.9% and class non loyal = 71.1%.

Looking at all these models, we can make some initial conclusions. First, The CART method is clearly more complex than the other the two methods we have tested. The resulting decision trees have more nodes and rules of subdivision. Second, age and click-through rate play the most important role in subdividing the population according to customer loyalty, with the open rate to a lesser degree. The other variables are less important in the decision trees resulting from these methods.

5.2 Evaluation the models using classification measures tests

Using To evaluate the quality of these models four classification measures, accuracy, recall, precision and F1- score as mentioned in the methodology section have been used to see which of the three models predicts customer loyalty best.

Figure 11 Shows the accuracy of the three models according to the decision tree analysis.

Given these results we can conclude that the accuracies achieved by CHAID and CART methods are higher than the accuracies achieved by QUEST method.

The precision measures for the decision tree analysis are presented in Figure 12.

From figure 12 we can conclude that the precision of the model 2 achieved by CHAID method using only the behavioural variables is higher than the precision of the other models using any of the three decision tree models.

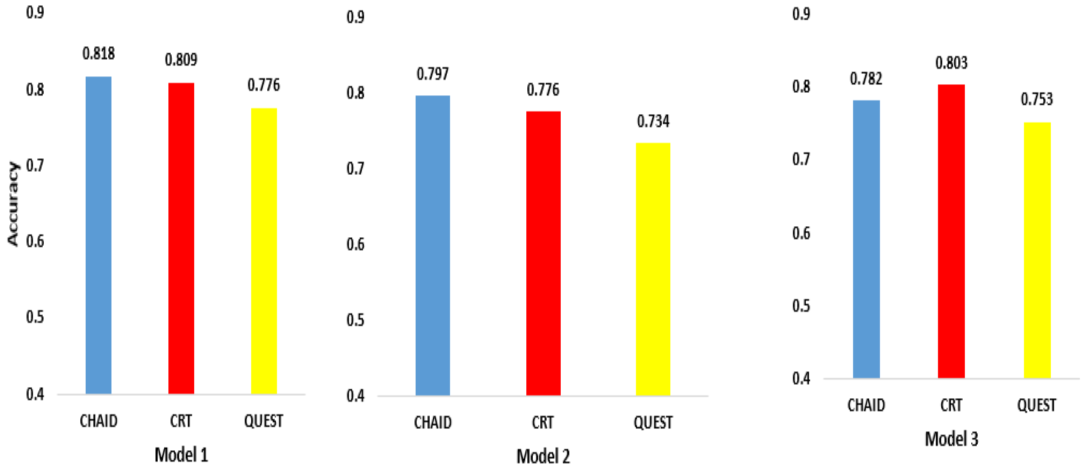

Figure 11. Accuracy test of the results of the decision tree models

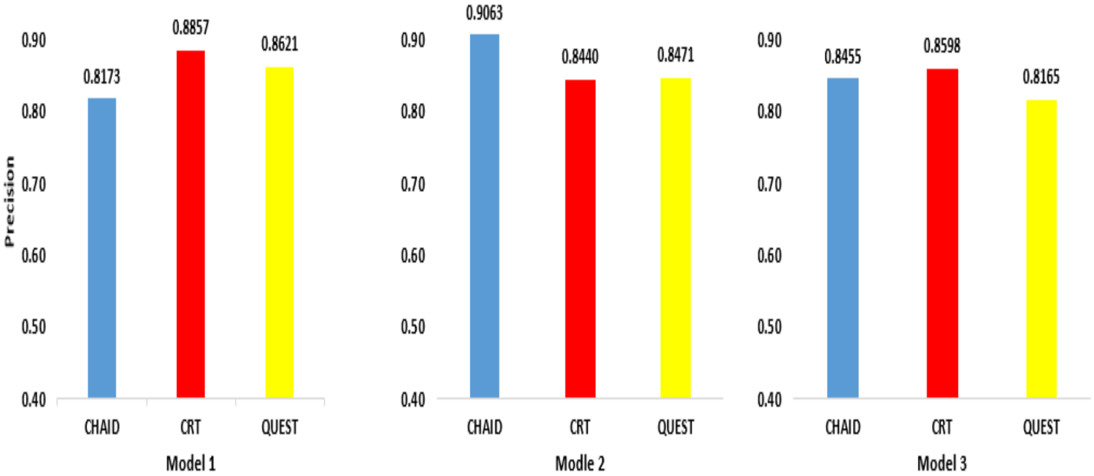

Figure 12. Precision test for the three decision tree models

Figure 13 shows the results of the recall test on the different decision tree models presented before.

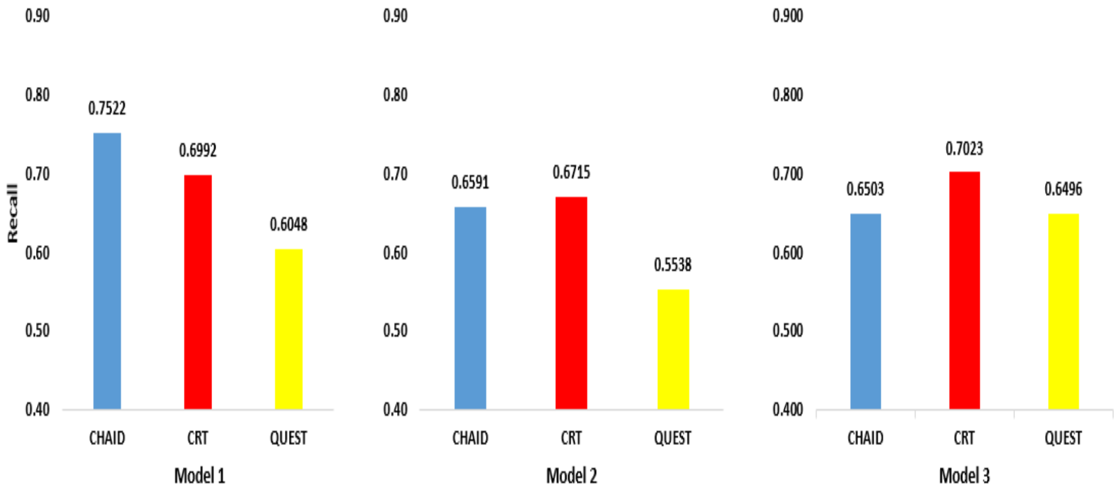

Figure 13. Recall measures of the different decision tree models.

In figure 13 we can observe that the recall measures for the first model achieved by CHAID method are higher than the recall results of the other models.

Figure 14 represents the results the F1-scores for those decision tree models.

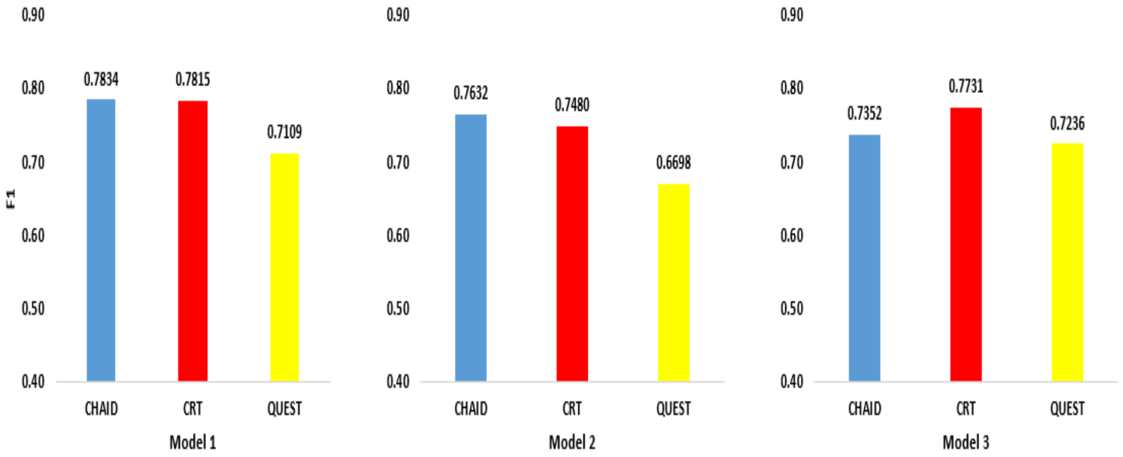

Figure 14. F1-score for the decision tree analysis models.

Figure 14 shows the F-scores for the first model are higher than for the other models.

Figure 15 shows there is no overfitting risk when we compare between the accuracy of the decision tree models for the training and test data.

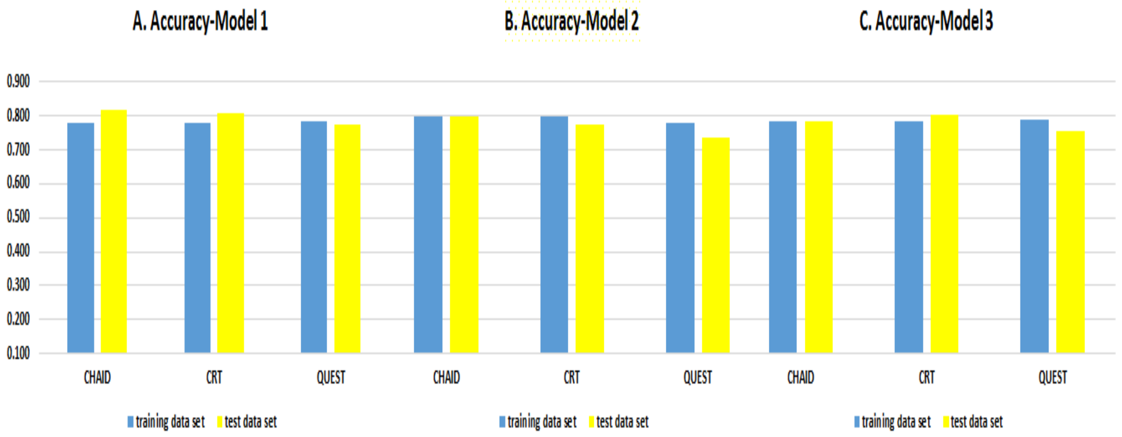

Figure 15. Accuracy of the decision tree models for the training and test data.

Section 6 discusses the conclusion , Managerial Consequences and findings.

6. Conclusion and future work

The major results of our different analyses enable us to answer the research questions that we have formulated.

The first research question read as follows :

What is the effect of some descriptive variables of the e-mail behaviour of customers on customer loyalty?

The effect of customer demographic variables and customer e-mail behavioural characteristics on customer loyalty was measured by decision tree analysis. Results indicated that nearly all those variables had a significant impact on customer loyalty. The only variable not to have a significant impact proved to be the country of residence of the customer.

The second research question was :

Which model predicts customer loyalty best on the basis of both the customer demographic and customer e-mail behaviour variables ?

This question was answered by using decision tree analysis. The results of the different analysis algorithms and models were evaluated using the classification measures of accuracy, precision, recall and F-score.

The decision tree analysis used three different methods or procedures: CHAID, CART and QUEST. In all cases the decision tree analysis of a model based on demographic variables alone proved to be a little bit more accurate, precise and better on recall as the other models. The best results were mostly obtained using the CHAID method. Since the accuracy of all models (both separate and the combined models of variables) is mostly close to 80 % all models have a good predictive value.

Managerial Consequences and findings

In general, this paper proves that it is useful to analyse and examine the loyalty of customers based on their behaviour when receiving e-mail advertisements as a part of an e-mail marketing campaign. Results of our research based on the decision tree analysis 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.

It is further important for all companies that use e-mail advertisements for their business to send these e-mails at the right time to the customers. Our study shows that the younger generation, female customers and customers who open the advertisements in the morning are more likely to be loyal customers. The Click through rate plays an important role in predicting customer loyalty, since the CTR and the high customer loyalty segment have a positive relationship .

So, If the company intends to increase the profit by increasing the response rates of the customers that have low click rates, it should send relevant advertisements to these customers based on their demographic characteristics to avoid customers unsubscribing from their website.

We recommend that the process of sending e-mails should consider the contents of the advertisement as well as the time customers usually open the e-mail campaigns in their mail. Figure 16 explains the different steps we recommend the company to use.

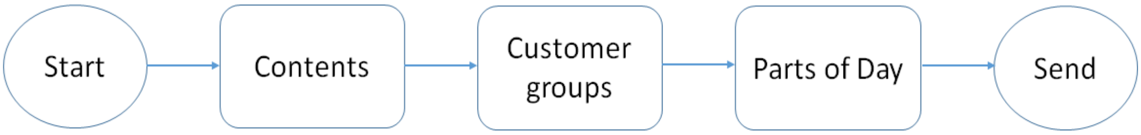

Figure. 16. The steps of the E-mail campaign process

The first step in this process is to know which contents of the e-mail campaign are relevant to the customer based on their demographic characteristics (age and gender) . Finally the web master team should consider several parts of day to send the e-mail to customers based on the right time at which customers normally open their e-mail advertisings.

Future work will have to address customers buying behaviour so that a company wanting to attract customers depending on their buying history by sending e-mails can be advised in more detail and depending on the products which customers like to buy.

Bibliography

Aziz Yarahmadi, Mathijs Creemers, Hamzah Qabbaah, Koen Vanhoof. 2017. 'UNRAVLING BI-LINGUAL MULTI-FEATURE BASED TEXT CLASSIFICATION: A CASE STUDY', International Journal “Information Theories and Applications, 24.

Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen. 1984. Classification and Regression Trees (Taylor & Francis).

Carmona, C. J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. García. 2012. 'Web usage mining to improve the design of an e-commerce website: OrOliveSur.com', Expert Systems with Applications, 39: 11243-49.

Cases, Anne-Sophie, Christophe Fournier, Pierre-Louis Dubois, and John F. Tanner Jr. 2010. 'Web Site spill over to email campaigns: The role of privacy, trust and shoppers' attitudes', Journal of Business Research, 63: 993-99.

Chiu, Chui-Yu, Yi-Feng Chen, I. Ting Kuo, and He Chun Ku. 2009. 'An intelligent market segmentation system using k-means and particle swarm optimization', Expert Systems with Applications, 36: 4558-65.

Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. 1999. 'Data Preparation for Mining World Wide Web Browsing Patterns', Knowledge and Information Systems, 1: 5-32.

De'ath, Glenn. 2000. 'Classification and regression trees: a powerful yet simple technique for ecological data analysis', Ecology, v. 81: pp. 3178-92-2000 v.81 no.11.

Díaz-Pérez, Flora M., and M. Bethencourt-Cejas. 2016. 'CHAID algorithm as an appropriate analytical method for tourism market segmentation', Journal of Destination Marketing & Management, 5: 275-82.

Etzioni, Oren. 1996. 'The World-Wide Web: quagmire or gold mine?', Commun. ACM, 39: 65-68.

F. Reichheld, Frederick, and Sasser Jr. W. E. 1990. Zero Defections: Quality Comes to Services.

G. T. Denison, D., B. K. Mallick, and A. F. M. Smith. 1998. A Bayesian CART algorithm.

George Sammour , Benoît Depaire , Koen Vanhoof and Geert Wets. 2009. "Identiying homogenous customer segments for risk email marketing experements." In 11th International Conference on Enterprise Information Systems, 89-94. milan , italy.

IBM. 2016. 'IBM SPSS Decision Trees 24.' in.

Kosala, Raymond, and Hendrik Blockeel. 2000. 'Web mining research: a survey', SIGKDD Explor. Newsl., 2: 1-15.

L., Steven, and Brewer Jr. 2012. 'AN EMPIRICAL COMPARISON OF LOGISTIC REGRESSION TO DECISION TREE INDUCTION IN THE PREDICTION OF INTIMATE PARTNER VIOLENCE REASSAULT', Indiana University of Pennsylvania.

Linoff, G.S., and M.J.A. Berry. 2011. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (Wiley).

Liu, Bing. 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications) (Springer-Verlag New York, Inc.).

Liu, Yanbin, Ping Yuan, Wei Liu, and Xingsen Li. 2015. 'What Drives Click-Through Rates of Tourism Product Advertisements on Group Buying Websites?', Procedia Computer Science, 55: 221-30.

Loh, Wei-Yin, and Yu-Shan Shih. 1997. 'SPLIT SELECTION METHODS FOR CLASSIFICATION TREES', Statistica Sinica, 7: 815-40.

Long, William J., John L. Griffith, Harry P. Selker, and Ralph B. D'Agostino. 1993. 'A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain', Computers and Biomedical Research, 26: 74-97.

Lopes, Prajyoti, and Bidisha Roy. 2015. 'Dynamic Recommendation System Using Web Usage Mining for E-commerce Users', Procedia Computer Science, 45: 60-69.

Lovelock, C.H., and J. Wirtz. 2011. Services Marketing: People, Technology, Strategy (Prentice Hall).

Michael J. A. Berry, Gordon S. Linoff. 2004. Data mining techniques second edition – for marketing, sales, and customer relationship management (wiley: canada).

Ngai, E. W. T., Li Xiu, and D. C. K. Chau. 2009. 'Application of data mining techniques in customer relationship management: A literature review and classification', Expert Systems with Applications, 36: 2592-602.

Reichheld, F.F., and T. Teal. 1996. The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value (Harvard Business School Press).

Reinartz, Werner, and V. Kumar. 2002. The Mismanagement of Customer Loyalty.

Ripley, Brian D., and N. L. Hjort. 1995. Pattern Recognition and Neural Networks (Cambridge University Press).

Shan, Lili, Lei Lin, Chengjie Sun, and Xiaolong Wang. 2016. 'Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization', Electronic Commerce Research and Applications, 16: 30-42.

Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. 2000. 'Web usage mining: discovery and applications of usage patterns from Web data', SIGKDD Explor. Newsl., 1: 12-23.

Stokes, R. 2011. Emarketing: The Essential Guide to Digital Marketing (Porcupine Press).

Sumaiya Thaseen, Ikram, and Cherukuri 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.

T., Bowen John, and Chen Shiang-Lih. 2001. 'The relationship between customer loyalty and customer satisfaction', International Journal of Contemporary Hospitality Management, 13: 213-17.

Xuerui Wang, Wei Li, Ying Cui, Ruofei (Bruce) Zhang, Jianchang Mao. 2010. 'Click-Through Rate Estimation for Rare Events in Online Advertising, In : online multimedia advertising ', Techniques and Technologies: 1-12.

Zhang, Xi. 2015. 'Managing a Profitable Interactive Email Marketing Program: Modeling and Analysis', Georgia State University.

Authors' Information

Hamzah Qabbaah *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B2a, Campus Diepenbeek, B-3590 Diepenbeek, Limburg,Belgium*

e-mail: hamzah.qabbaah@uhasselt.be

Major Fields of Scientific Research: Data mining, digital marketing, Social Network Analysis, Knowledge Management

George Sammour *- Director of Quality Assurance and Accreditation Centre, Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan.*

e-mail: George.Sammour@psut.edu.jo

Major Fields of Scientific Research: Data mining, Digital learning, lifelong learning, professional development

Koen Vanhoof *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek, Limburg,Belgium*

e-mail:koen.vanhoof@uhasselt.be

Major Fields of Scientific Research: data mining, knowledge retrieval