SCHOOL OF BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ACADEMIC YEAR OF 2021 – 2022

STATISTICS FOR BA II – ASSIGNMENT 2

NINAS KONSTANTINOS

f2822108

SUPERVISING PROFESSOR: KARLIS DIMITRIOS

Table of Contents

1. Introduction

We are provided a dataset that contains the call history of a telemarketing company that is promoting a new product on behalf of a retail bank. The company's agents make daily phone calls to lists of the company's existing customers to promote and sell a new product. At the same time, clients may call the company's center for other reasons, during which the agents that serve them also try to promote the company's new product. It should be mentioned that many customers were contacted more than once, to persuade them to finally buy the product.

The dataset contains data regarding almost 40K phone calls that were conducted from May 2008 to June 2010. The dataset includes data that describe the bank's client (such as age, job, marital status, education and more), the date and the duration of the last call with each client, attributes that concern the telemarketing company's campaigns (for example number of calls per client and more), social and economic attributes, like consumer price monthly indexes and finally, the client's final decision on buying the product. The main purposes of this analysis are to produce classification predictive models that can efficiently predict whether a customer will end up buying the bank's product and to identify customers with common attributes through clustering.

2. Data Cleaning

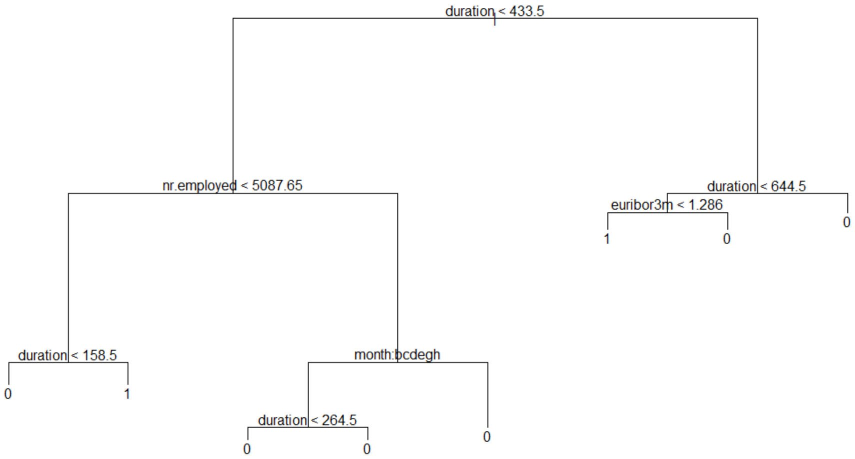
The first step, prior to the classification models and the identification of customers' clusters, is the thorough clean process of the data. Specifically, the client's job as well as their marital status, education, credit default index, housing, subscription and personal loan indexes were updated to categorical variables. The same was applied for the communication type ('cellular', 'telephone'), the month and the weekday of the last contact with each client and outcome of the previous campaign with them. Additionally, clients that were not contacted in the telemarketing company's previous campaigns received the value '999' as a placeholder in the variable 'pdays', which index the number of days number that passed by after the client was last contacted from a previous campaign. In fact, it was found that only 1.000 out of almost 40.000 customers in the dataset had been contacted at least once in previous campaigns. As a result, the rest 39.000 customers would have blank values in this specific variable. To resolve this issue, it made more sense to create an index of whether each customer was contacted at least once in any previous campaign. Although, it was found that the variable 'poutcome' indexed whether a customer was contacted in the previous campaign, so it was not needed to create a new variable. As a result, the variable 'pdays' was dropped completely from the dataset. Also, it is noticed that all the values in the consumer confidence index variable are negative, which does not make any sense, so they were all updated with their corresponding positive numbers.

3. Classification Models

Following, five classification methods will be implemented to predict whether a customer will buy the bank's product based on the rest of his attributes (such as age, job etc.). In order to identify different effects of the data, each model will be trained and tested (in a 80-20 manner) on 5 different folds of the data using cross validation to capture their different effects. Finally, the best method will be also evaluated on a validation subset of the data.

Each method's effectiveness will be evaluated based on its accuracy, its precision, its recall and its f1-score. A model's accuracy can be described as the number of correct predictions divided by the number of all of its predictions. Precision can be described by the number of the correctly reported positive predictions (a customer buys the product) divided by all of its positive predictions. Recall (or sensitivity) can be described by the number of the correctly reported positive predictions (a customer buys the product) divided by the number of all actual buyers in the predictions. Finally, F1-score can be described as a weighted average of the precision and the recall measures. The complete results of each method can be observed in detail in the tables 1 and 2.

The first classification method is named Decision Tree. In such methods, the trees grow by selecting iteratively the variable that provides the most information until there are no more variables, the training examples associated with the leaf nodes all have the same target value or when the new leaf nodes are not statistically significant (Plot 1). The method had a mean accuracy of 90%, a mean precision of 51%, a mean recall of 32% and a f1-score of 33% in its in-sample predictions (table 1).



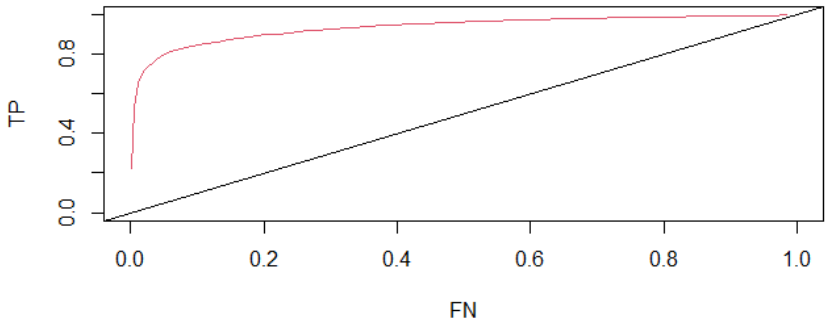
Plot 1. Decision Tree Classifier

The second classification method is named Support Vector Machines (SVM). The simple idea behind SVMs is that the two classes in the dataset (buyers of the product/ non buyers) can be separated through a hyperplane. The method had a mean accuracy of 92%, a mean precision of almost 66%, a mean recall of 29% and a f1-score of almost 40% in its in-sample predictions (table 1)

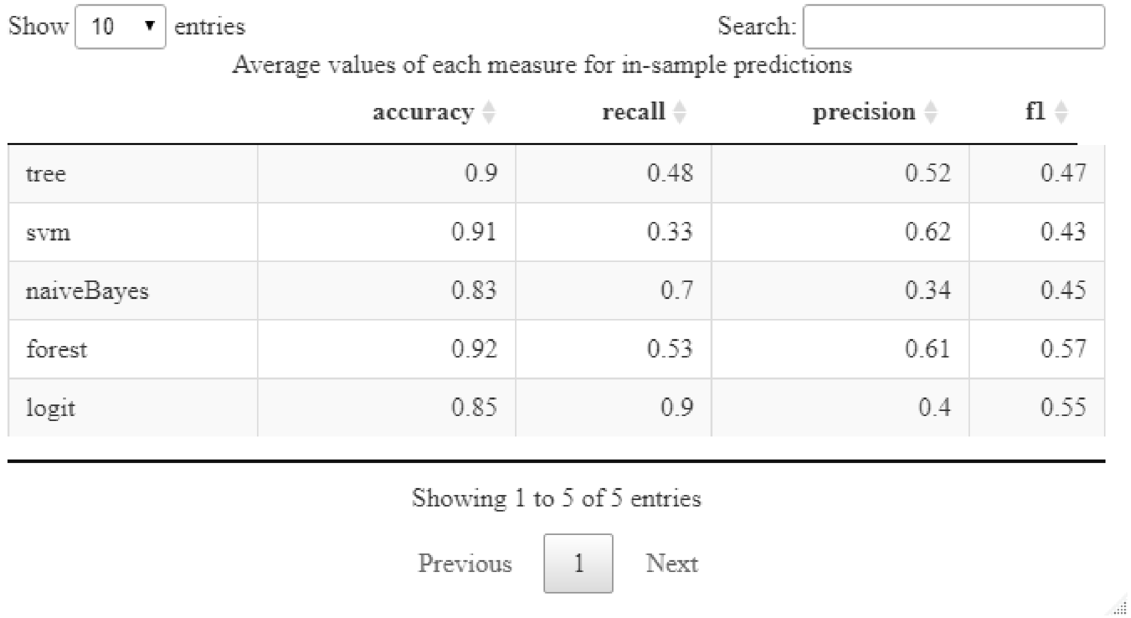
The third classification method is named Naive-Bayes Classifier. This method calculates the probability of each feature being assigned to a class for new observations. Upon calculating the probabilities for each feature, it multiplies them and selects to assign the new observation to the class with the greatest value. The method had a mean accuracy of 83%, a mean precision of 33%, a mean recall of 70% and a f1-score of almost 45% in its in-sample predictions (table 1).

The fourth classification method is named Random Forest. This method produces many trees that contain less features than the initial dataset and uses them to conduct its predictions. Once all the trees have been produced, the mode of their prediction is kept as the final prediction for each new observation. The method had a mean accuracy of almost 92%, a mean precision of 60%, a mean recall of almost 51% and a f1-score of almost 55% in its in-sample predictions (table 1). Additionally, the random-forest method was also evaluated on a validation subset of the initial dataset. The out-of-sample predictions had an accuracy of 92%, a precision of 59%, a recall of 56% and a f1-score of 57% (table 2). In other words, for the most part, the out-of-sample predictions of this method were slightly better in comparison to those of the in-sample predictions based on the below measures.

The final classification method is named Logistic Regression. This method calculates the probability that one may buy the product based on the rest of his/her attributes and based on a fixed threshold, classifies him/her as a buyer or non-buyer. The threshold of the above probability was calculated with the assistance of the ROC Curve (plot 2). Specifically, the ROC Curve plots the True Positive rate against the False Positive rate for every threshold from 0 to 1. The method had a mean accuracy of 85%, a mean precision of 40%, a mean recall of 90% and a f1-score of 55%.



Plot 2. ROC Curve

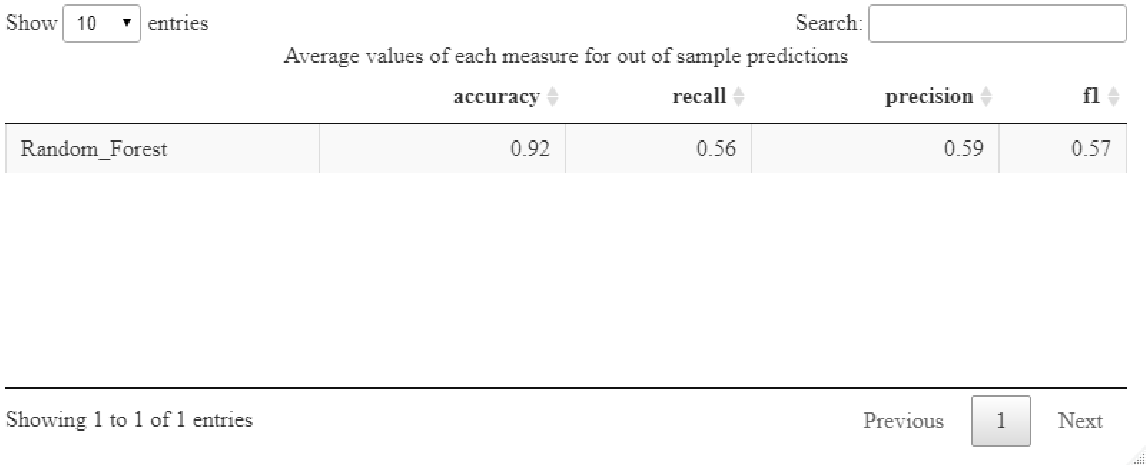


Average values of each measure for in-sample predictions

	accuracy	recall	precision	f1
tree	0.9	0.48	0.52	0.47
svm	0.91	0.33	0.62	0.43
naiveBayes	0.83	0.7	0.34	0.45
forest	0.92	0.53	0.61	0.57
logit	0.85	0.9	0.4	0.55

Showing 1 to 5 of 5 entries

Table 1. All metrics for in-sample predictions for each method



Average values of each measure for out of sample predictions

	accuracy	recall	precision	f1
Random_Forest	0.92	0.56	0.59	0.57

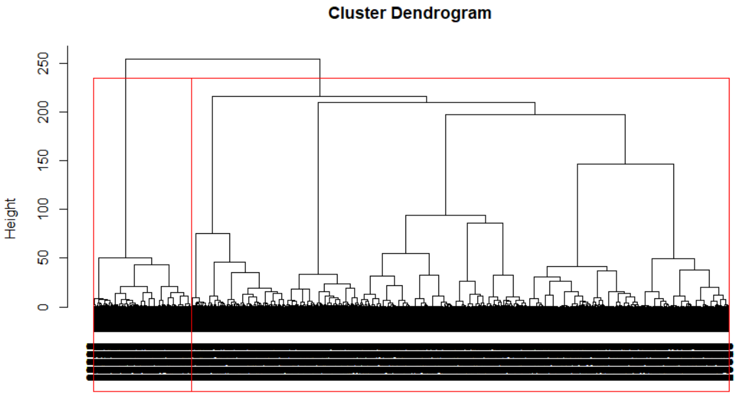
Showing 1 to 1 of 1 entries

Table 2. All metrics for out of sample predictions – Random Forest

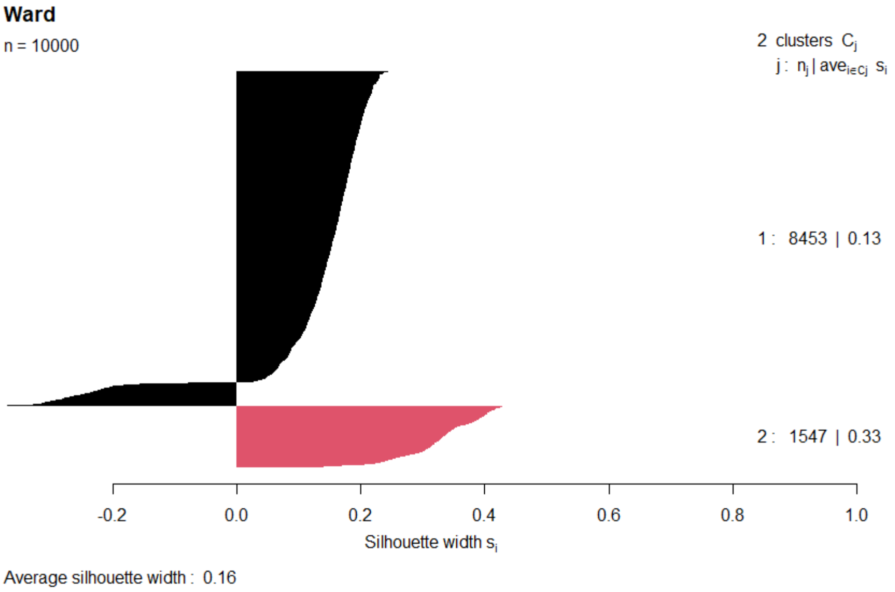
4. Identification of Customer Clusters

Following, the identification of customers segments that share common attributes in the sample will be attempted by implementing a clustering method. In detail, through the clustering method it will be possible to identify groups of customers that are similar, while at the same time differ from customers that belong to different groups. The attributes that will be employed to identify these clusters will include the customers' age, job, marital status, education level, and the indexes that show whether they have credit in default, a housing and a personal loan. In addition to the above features, the attributes of the campaign that will be employed in this method are the number of contacts during the current and the previous campaign and their decision in the previous marketing campaign. In other words, the variables that described the social and economic context attributes, the date and the duration of the last contact with each customer were disregarded since they do not offer any additional information about the customers. To identify the number of customers' clusters, a random sample of 10.000 customers from the initial dataset will be utilized.

A Hierarchical Clustering method was implemented utilizing a Ward link to conduct the above analysis. Hierarchical clustering methods seek to build hierarchies of clusters that can be later on be divided to many smaller clusters. The results of such methods can be presented and evaluated using dendrograms (plot 3) and silhouette plots (plot 4). Silhouette plots can show how good have the customers been clustered in the different clusters that have been set either by observing the plot or by examining the average silhouette width. Clusters with average silhouette width close to 1 are considered as well separated, while the opposite can be inferred for those with an average width close to 0. Through the exploration of different clustering methods, it was discovered that the customers can be divided into 2 customer groups/clusters given the above attributes. However, the two customer groups are not separated very good, since the average silhouette width of the groups is equal to 0.16. The two customer groups that were formed were not of equal sizes, as the first one consists of almost 8.500 customers out of the 10.000 in the sample (plot 4).

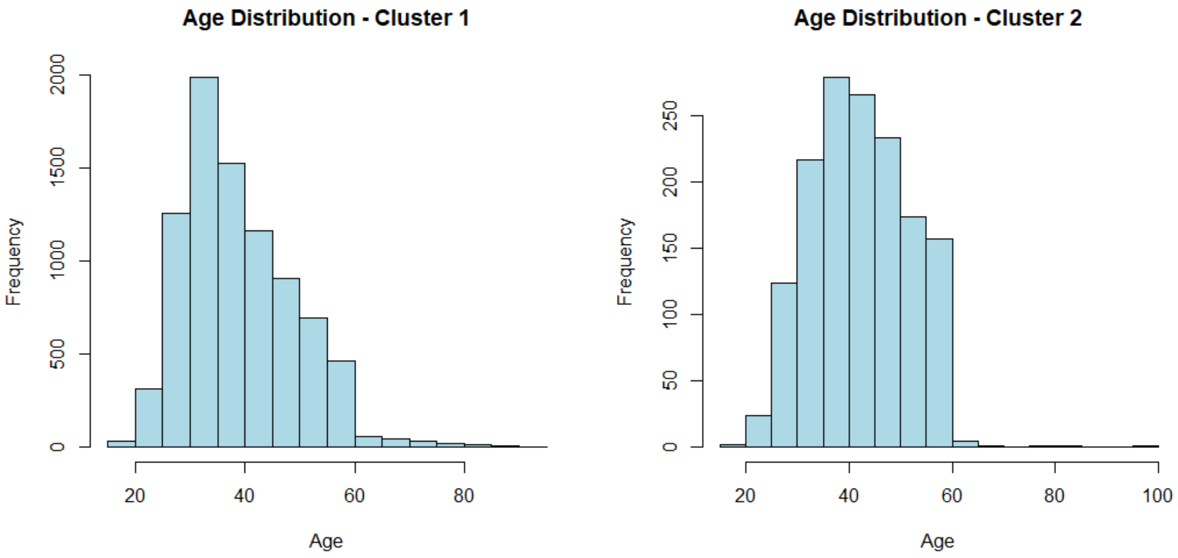


Plot 3. Cluster Dendrogram

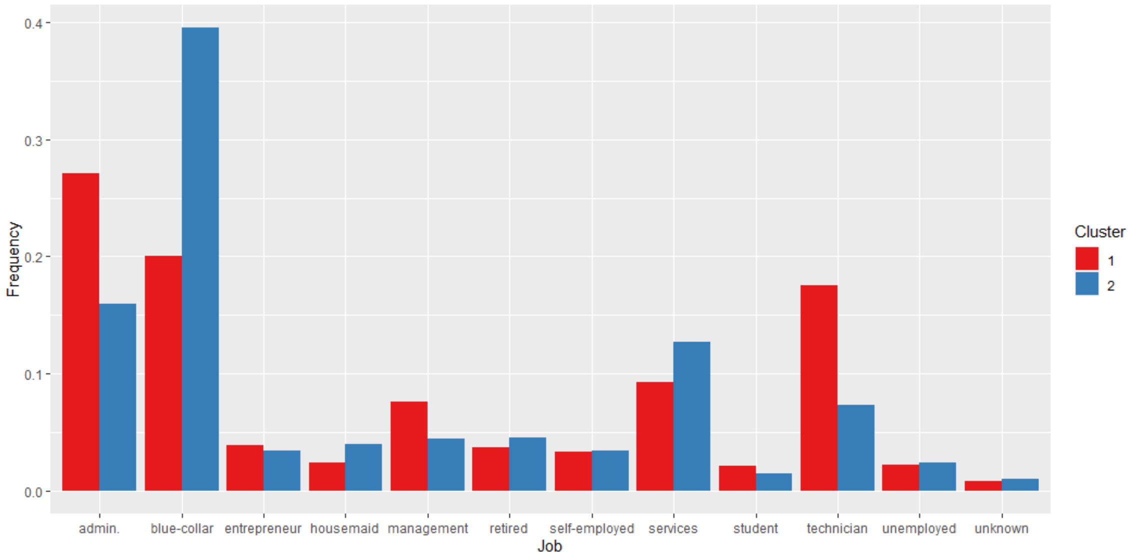


Plot 4. Silhouette plot

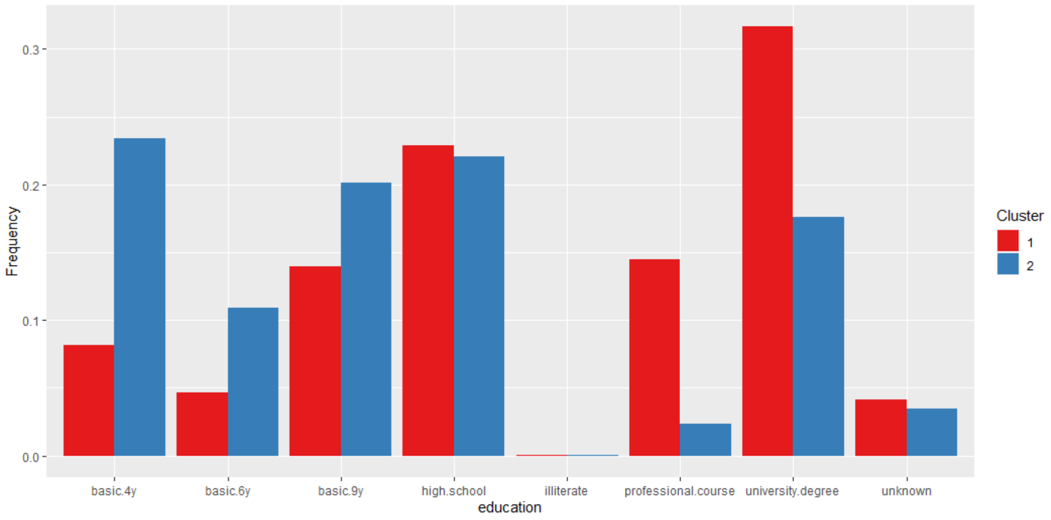
The customers that belong in each group, as expected differ in some of the attributes that are examined. The first key difference that can be spotted is in their ages (plot 5). It can be observed that the majority of the customers in the first group are between 30-40 years old, while only a few of them are older than 40 years old. On the other hand, most of the customers in the second group tend to be older than those of the first group, given that most of them are approximately 40 years old, while a significant number of customers in the group are up to 60 years old. The second key difference is spotted in their job occupations (plot 6). The highest percentage of the customers in the first group seem to be administrators, while the by-far highest percentage of customers in the second group seem to be occupied in blue-collar jobs. The third key difference is observed in their education levels (plot 7). It can be observed that the vast majority of the customers that belong in the first group have a university degree, while most of the customers in the other group have either a basic 4-year or 9-year education or a high-school degree. In addition, the clusters were compared with the subscription indexes of each customer in the sample by using the Adjusted Random Index (ARI). This action was conducted to determine whether these two are related. The Adjusted Random Index can usually reveal such relationships since it examines how similar are the two ways that the data have been partitioned. To be specific, ARI values that are close to 1 may show that there is a similarity in their partition, while ARI values close to 0 show that there is no similarity whatsoever. It was revealed that the clustering was not related with the subscription index, since they had an ARI value approximately equal to zero.



Plot 5. Age distribution per clusters



Plot 6. Job occupation of the customers per cluster



Plot 7. Education levels per cluster

5. Conclusions

Summarizing, it is observed that the most accurate and the best one in terms of F1-score classification model was the Random Forest, the most precise one was the Support Vector Machines model and the best in terms of sensitivity was the Logistic Regression model. Additionally, it is observed that even though most of the classification method had an accuracy close to 90% in their predictions, most of the rest of the metrics evaluated had a much worse performance. This proves that by measuring only one metric (such as accuracy) is not sufficient to fully evaluate a method’s performance. Furthermore, the Random Forest method was selected to be evaluated on out-of-sample data. The method proved to be slightly better in its out-of-sample predictions compared to its in-sample predictions. However, it is mandatory to state that it is possible, through further tuning of the existing models or through the implementation of different models that the predictions of the customers’ decisions can be even better than those of the existing models.

In addition, through the Hierarchical clustering method implemented 2 different customer groups were identified. The first group mostly consists of customers that are between 30 to 40 years old, are occupied as administrators and have a university degree. The second group mostly consists of people that are approximately 40 years old, are occupied in blue-collar jobs and have a basic 4-year or a high school level education. The two customer groups cannot be divided perfectly, which results to many customers having common attributes regardless of the group they belong to. Also, it was discovered that the two customer groups are not associated with the subscription decision of each customer. Again, it is possible through further tuning of existing methods or through the exploration of new clustering methods that the customers can be divided into more dissimilar groups and that their partitions be associated with their subscription decision.