ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY

MSc BUSINESS ANALYTICS

“Assignment 2 in course: Statistics for Business Analytics II”

Full Name: STAMATIOS SIDERIS

Register Number: f2822113

ATHENS, 2022

Contents

1. Introduction

The dataset refers to telemarketing phone calls to sell long-term deposits. It consists of 39883 observations and 20 variables. The purpose is to create a predictive model to classify whether a client will buy or not the new product as well as use specific variables to cluster the clients and to characterize the clusters.

2. Data Cleaning

The dataset has 9 numeric and 11 character variables. The dataset has zero blank values and 12520 unknown values. Character variables are changed to type factor and separated to their respected levels and integer variables are changed to numeric.

3. Classification

The purpose is to create a predictive model to classify whether a client will buy or not the new product.

3.1. Training & Testing Datasets

Firstly, we need to separate our dataset to 2 subsets for training and testing. We will use 5-Fold Cross Validation to train and test our model. In K-Fold Cross Validation, the dataset is split to k subsets and each time the model is trained it uses k-1 subsets while the remaining 1 subset is used for the testing validation. As our dataset consists of 39883 observations, a 5-Fold split is enough as to have a balanced number of observations across each fold.

3.2. Classification Methods

The first method to be used is the **Naive Bayes**. The method uses Bayes theorem of conditional probability to classify the observation on the event with the highest probability given some of its characteristics. The training dataset is used to train the model and the testing dataset is used on our trained model to predict whether a client will buy or not the new product. As Naive Bayes implies a very complicated multivariate distribution for a large number of variables given, we assume that the variables are conditionally independent to the class we check if they should be classified to and so our final estimation is not that accurate but naive.

The second method to be used is the **Decision Tree**. In this method, we try to create a tree that is consisted of a Rooting Node, branches and leaves by assigning variables to discrete classes. The root node includes the variable that explains most of the variance of our target variable for classification. The rooting node is connected to other nodes via branches which show us the way to follow in the decision tree until we reach nodes with no branches that are called leaves. Each node performs a classification based on our target variable until the tree reaches to a leaf where the classification of most observations agrees. The variables to be used as well as their thresholds (for continuous variables) are determined with the use of Gini index. Gini index is equal to 1 minus the sum of probabilities to appear for each class of the variable. The variable with the lowest Gini index is chosen for the Node each time. Again, the training dataset is used to train the model and the testing dataset is used on our trained model to predict whether a client will buy or not the new product. As Decision Trees try to maximize the reduction of impurity in a Node, they are exposed to overfitting with no bias and potentially large variance and so a method called pruning is available for use where branches of the tree are left off creating less accurate trees (with bias) but with better results on testing datasets (lower variance).

The third method to be used is the **Support Vector Machine.** The method tries to create a line (for 2-dimensional data), plane (for 3-dimensional data) or hyperplane (for more dimensions) so that the margins created between it and the closest to it observations are the minimum possible. If the separation is difficult to be performed as the classes are mixed together, then a constant is developed that indicates the error we allow for misclassification. The algorithm performs cross validation to find the best constant that minimizes the allowed misclassification error. The method is very sensitive to outliers as an observation of one class close to another class but far away to its own class could lead to false minimal margins and false classifications. Again, the training dataset is used to train the model and the testing dataset is used on our trained model to predict whether a client will buy or not the new product.

3.3. Accuracy Evaluation

To evaluate the accuracy between our methods, we will calculate the ratio of accuracy by finding the number of matches between the predictions and the initial testing dataset per method and dividing it by the number of observations. As we have used 5-Fold CV, the procedure will be performed for all the 5 different trained models and the final predictive accuracy per model will be calculated as the average accuracy of the 5 models per method. The method with the highest average ratio of accuracy is the Support Vector Machine equal to 91% against that of Decision Tree (90%) and Naive Bayes (85%), as depicted in the median of the boxplots in **Figure 1**.

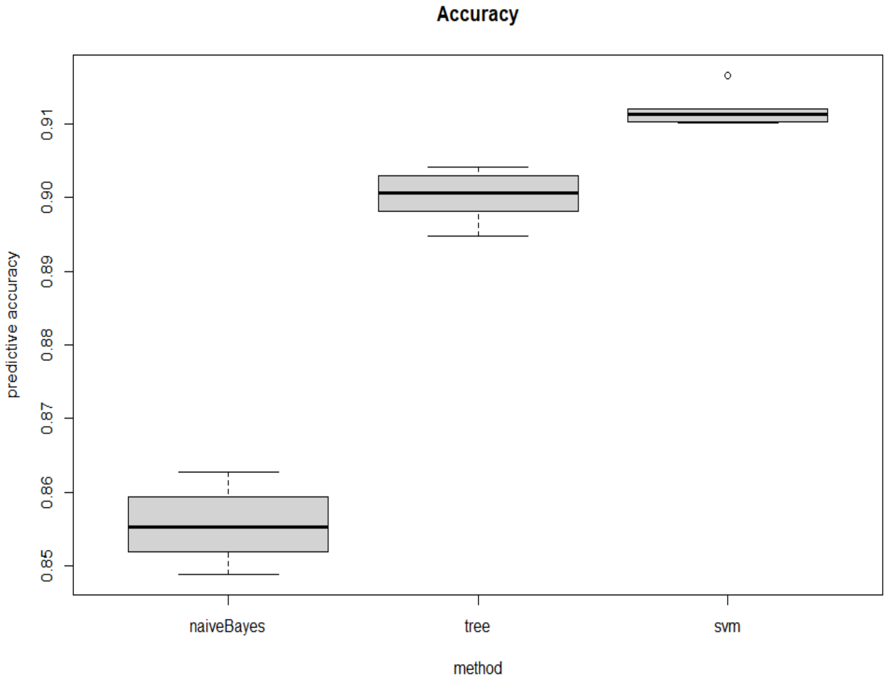

Figure 1 Predictive Accuracy per Method

4. Clustering

The purpose is to use specific variables to cluster clients and characterize the clusters.

4.1. Variable Selection

The number of variables in the dataset is high, so we need to reduce the number of variables for reasons of computational power and speed. As our purpose is to separate clients into clusters, we need variables that characterize clients the most, so we keep only the variables that refer to clients' personal data. These variables are:

1 - age (numeric)
2 - job : type of job (categorical)

3 - marital : marital status (categorical)
4 - education (categorical)
5 - default: has credit in default? (categorical)
6 - housing: has housing loan? (categorical)
7 - loan: has personal loan? (categorical)
8 - campaign: number of contacts performed during this campaign and for this client (numeric)
9 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric)
10 - previous: number of contacts performed before this campaign and for this client (numeric)
11 - poutcome: outcome of the previous marketing campaign (categorical)

4.2. Distance Matrix

Our dataset consists of a large amount of data and so it is difficult for the computational power and memory of our Hardware to create a distance matrix of that many dimensions. We will proceed with a random sample of 10000 observations. Our dataset consists of numeric and categorical variables (mixed data) and as a result we will use the Gower's Distance to calculate the distances between the observations. Gower's distance is computed as the average of partial dissimilarities across individuals. The distance is always a number between 0 (identical) and 1 (maximally dissimilar).

4.3. Hierarchical Clustering

We use the Hierarchical method to perform clustering on our data. According to this method, each observation is assigned to each own cluster. We calculate the Euclidean distance between each observation and merge the ones with the lowest distance into the same cluster. We continue the same procedure until 1 or a given number of clusters has been constructed. To calculate the distance between clusters and not points we will choose the complete linkage and compute the Euclidean distance between the farthest points of the two clusters.

As we are not sure about the appropriate number of clusters our algorithm needs to stop, we will use the method for 2 and 3 clusters and compare the results. In **Figure 2,** it is presented the Dendrogram of the algorithm and with red light are noted the clusters for k=2 and k=3 number of clusters.

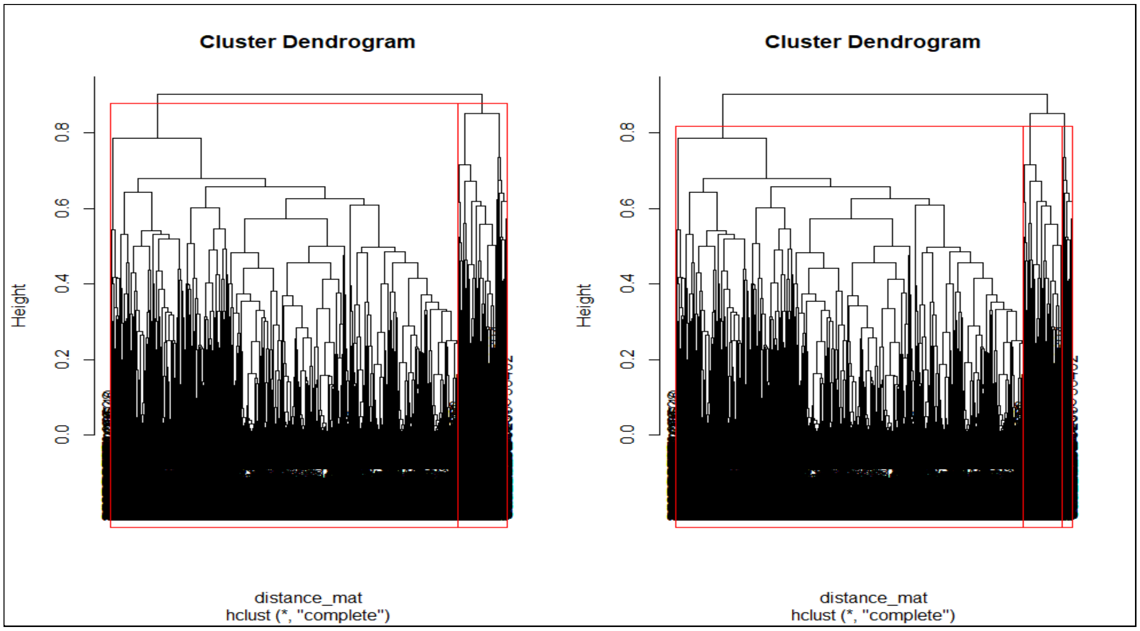

Figure 2 Cluster Dendrogram for 2 (on the left) and 3 (on the right) number of Clusters

To evaluate the 2 cases, we will plot the Silhouette Diagrams of the 2 cases, in **Figure 3**. In a Silhouette Diagram, it is observed the number of clusters and the number of observations each cluster includes as well as the average width per cluster and the total average width. Observations with close to 1 width are well classified, for close to 0 width could belong to any of the clusters and for negative width are falsely classified. A small number of observations is falsely classified and specifically 28 for 2 clusters and 31 for 3 clusters. The overall average width is almost identical for both cases as it is 0.27 for k=2 and 0.25 for k=3. At the same time, all the clusters include a significant number of observations. As the overall average width is slightly lower for k=3 we will proceed with 3 clusters.

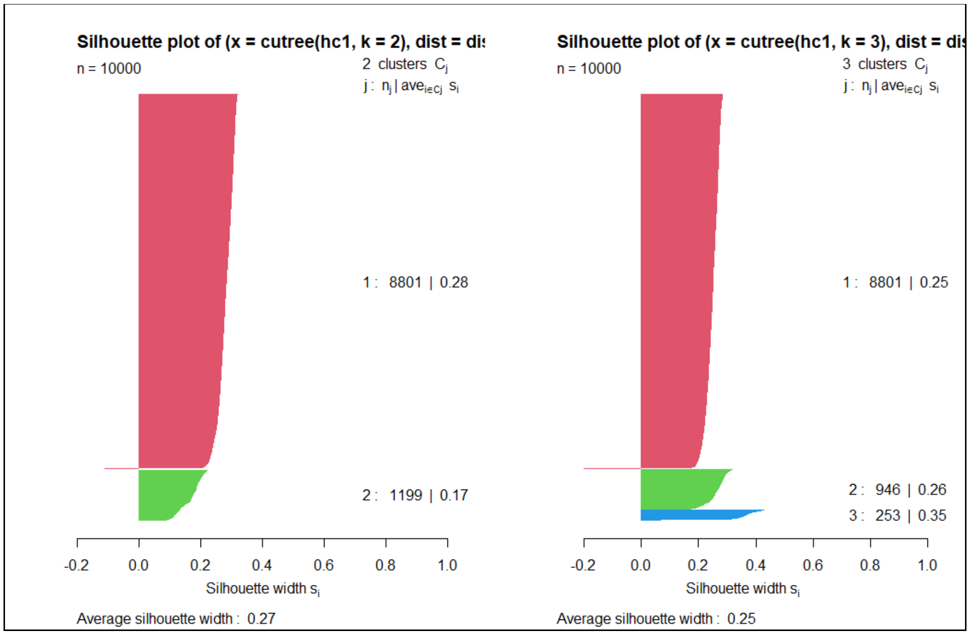

Figure 3 The Silhouettes Diagrams for k=2 (on the left) and k=3 (on the right) number of Clusters

To evaluate if the clustering has any correlation to if the client subscribed for the product or not, we will use the Adjusted Rand Index and compare the clusters we found to the hypothetical 2 clusters created by the variable SUBSCRIBED (if subscribed or not). The ARI is equal to 0.12 meaning that a very low correlation exists between the clusters we found and the variable SUBSCRIBED and so if someone subscribed or not does not have any importance for our clusters.

5. Conclusion

In conclusion, the most accurate classification method to be used to predict if a client will buy the new product or not is the support vector machine with a ratio of 91%. The best number of clusters for our clients, as proposed by the hierarchical method based on variables that indicate clients' personal data, is 3 but with small differences to the clustering by 2 clusters and so it is on the company's hand to decide the use of the information based on its needs. The subscription of a client for the new product does not characterize him enough to separate him from other clients based on available clients' personal data.