Course:

Statistics for Business Analytics II

Professor:

Dimitris Karlis

Second Assignment:

Bank telemarketing phone calls
PART II

(Classification & Clustering)

Student:

Name: Maria Skoli

AM: p2822131

Mar 2022

Table of Contents

Introduction

In the previous assignment we were engaged in finding a model ideal to interpret our data. The aim of this assignment is a little different. We are going to use both supervised and unsupervised methods aiming in prediction. In the first part of this assignment, we are going to use supervised learning methods to identify to which category (Subscribed or non-subscribed) a new observation belongs. The second part of this report aims to find groups of observations with the property that within the group the observations look similar but within different groups are totally different.

In the first part, we will examine 3 different methods of Classification. Some of them are classified as hard classification methods meaning that we assign each observation to a class, no uncertainty around, and other are classified as soft classification because we have also available the probabilities of belonging to each class. Namely, we will use logistic regression and naïve bayes as soft classification methods and the decision trees method as hard. We will explain each of them in more detail below. Finally, we will compare them based on the accuracy and the predictive ability.

In the second part, we will examine if there are similarities across individuals, using the mathematical concept of distance. After the selection of the optimal number of clusters we will try to interpret the result.

PART 1: Classification

1.1 Which variables are useful?

The main aim of classification is to predict. Therefore, we have to select variables to use in our model wisely. The selected variables should contribute to predicting the outcome of the phone call. Too many not highly significant variables can add noise to our model.

Keeping in mind the model we have used to interpret our data; the duration of the call was one of the most important attributes. However, when we goal to prediction this attribute cannot be used as the duration of a phone call it is known at the end of the campaign, when the outcome is already defined.

Moreover, the month of the phone call should also be excluded from predictive models because in our data we have no sufficient information about all months of the year. More specifically, we have no observations for January and February. So, we will not be able to classify a new observation observed in those months.

Finally, Euribor3m is recorded on daily basis and therefore, it is not suitable to predict future phone calls campaigns.

To conclude, the variables that we used to start our analysis are the following:

- Age
- Job
- Marital
- Education
- Default
- Housing
- Loan
- Contact
- Day_of_week
- Campaign
- Pdays
- Previous
- Poutcome
- Emp. Var rate
- Cons price idx
- Cons conf idx

And the response variable “SUBSCRIBED”

In order to find out if there is noise in our data, we constructed two logistic regression models. The first one contains all aforementioned variables and the second one excludes from the analysis the following variables: marital, education, loan, housing, previous, default, pdays and day_of_week. We used the ROC curve metric to examine if those variables contribute to the predictive ability of the model.

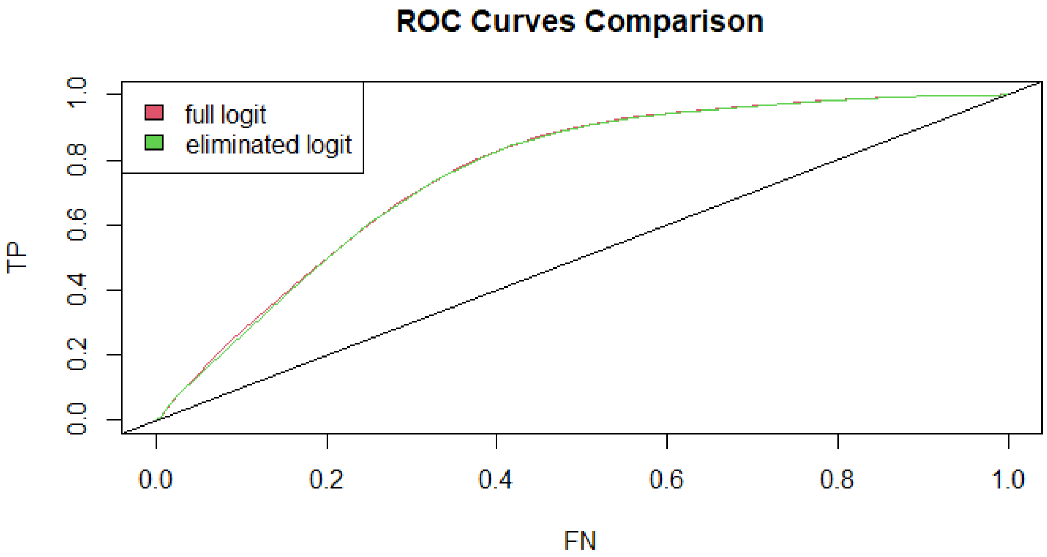


Figure 1: Roc curves of two logistic models

We found out that the full model and the eliminated model did not have a crucial difference and therefore, we can exclude these variables from the analysis. We will select the remaining variables to examine all methods of classification presented below.

1.2 Classification Methods

First, we must note here that our data are quite unbalanced as we have 90% observations related to non- subscribers and only 10% to subscribers. This indicates that the accuracy of a classification model should be higher than 90% to results in a better prediction than luck. In case of unbalanced data sometimes it useful the altering of the default threshold = 0,5.

Logistic Regression

The logistic regression models the probability that someone belongs to a particular category. In our case the categories are Subscribers and non-Subscribers. It is classified as soft classification.

The method used to fit the model is called maximum likelihood. The coefficients are chosen to maximize the likelihood function.

We run the model and we observed the followings about the predictive ability.

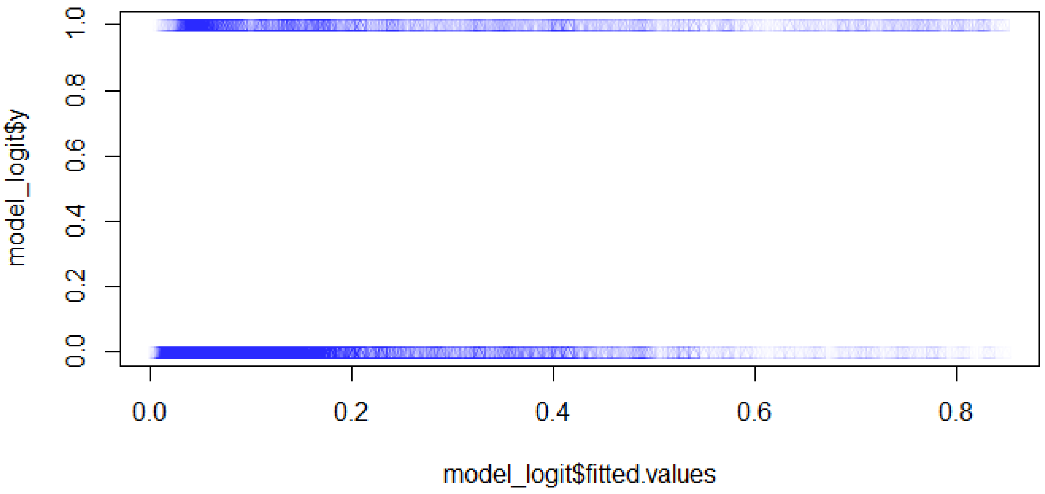


Figure 2: Expected outcome vs actual outcome

In the plot above we can assume that there are observations where our model assigns a low probability, but the actual outcome was client subscription. In the same way, there are observations where our model is almost sure of client subscription, but the actual outcome was that the client rejects the subscription. However, most of the observations appear at the lower left corner which is true as most observations are non-subscribers.

If our company wishes to be conservative in predicting individuals who are willing to subscribe, then we may choose to use a lower threshold than 0,5. In order to find the best threshold we can use a metric called ROC curve.

The ROC curve tells us how well our classifier is classifying between term deposit subscriptions (True Positives) and non-term deposit subscriptions. The X-axis is represented by False positive rates and the Y-axis is represented by the True Positive Rate. As the line moves the threshold of the classification changes giving us different values. The closer is the line to our top left corner the better is our model separating both classes. We found out that the threshold that refers to the minimum distance of the curve from the point (0,1) is 0,09.

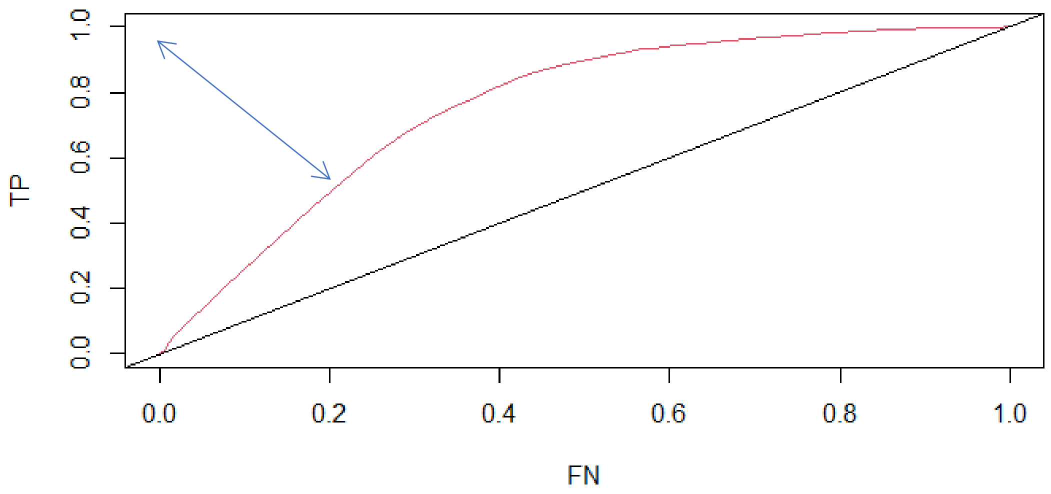
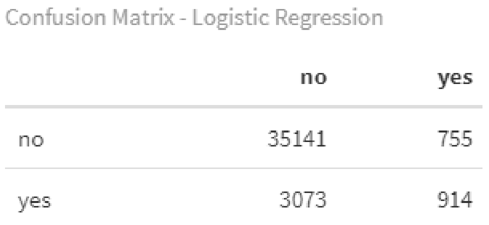


Figure 3: Identify the smallest distance from (0,1) using Roc Curve

After that, important metrics of our model is the accuracy of the model, namely the number of correct predictions, the recall which is how many “Yes” labels do our model detect, and the precision metric which stands for the certainty in the prediction of our model that the actual label is a “Yes”. This is very important because the higher the precision the more likely the model is to miss instances that are actually a “Yes”.

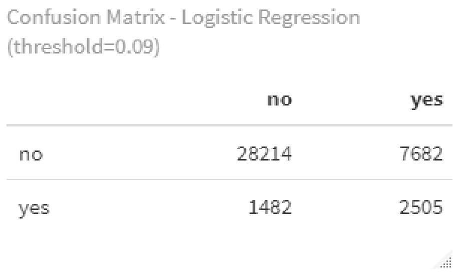
With the confusion matrix below we can determine how many observations were correctly or incorrectly classified. The required **Accuracy is 90,65%** which is quite better than luck.



Confusion Matrix - Logistic Regression

	no	yes
no	35141	755
yes	3073	914

Figure 5: Confusion Table of logistic regression with threshold = 0,5



Confusion Matrix - Logistic Regression (threshold=0.09)

	no	yes
no	28214	7682
yes	1482	2505

Figure 4: Confusion Table of logistic regression with threshold =0,09

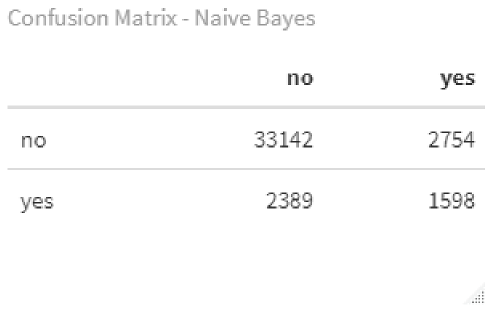
We observed that in the logistic regression with threshold =0,5 we have better accuracy and precision. In other words, when it predicts a subscriber, it is correct 54% of the time. However, the recall is low as it correctly identifies only the 23% of subscribers.

On the other hand, the model with the threshold = 0,09 the recall is ameliorated as it correctly identifies the 63% of subscribers but when it predicts a subscriber it is correct only 25% of the time. It is a business decision which view we prefer to follow.

Finally, the logistic function will always produce an S-shaped curve and so, regardless of the values of the attributes we will always obtain a sensible prediction.

Naïve Bayes

Naïve Bayes assumes independence of the features. However, this is a strong assumption hard to verify. It is a soft classification method as it assigns posterior probabilities. The way they get these probabilities is by using Bayes' Theorem, which describes the probability of an individual, based on prior knowledge of conditions that might be related to him/her.



Confusion Matrix - Naive Bayes

	no	yes
no	33142	2754
yes	2389	1598

Figure 6: Confusion Matrix for Naive Bayes Method

From the confusion matrix we can see that the accuracy of the model is **87,10%**. Moreover, the precision of the model is 36% and the recall is 40%.

It is definitely a strict model.

Decision Trees

Decision trees can provide a clear indication of which fields are most important for prediction or classification. It is classified as hard classification. We discretize the variable age in 3 age groups and exclude from the model the employee rate. In this way we can achieve a better accuracy of the model. The final tree we extracted is the follow:

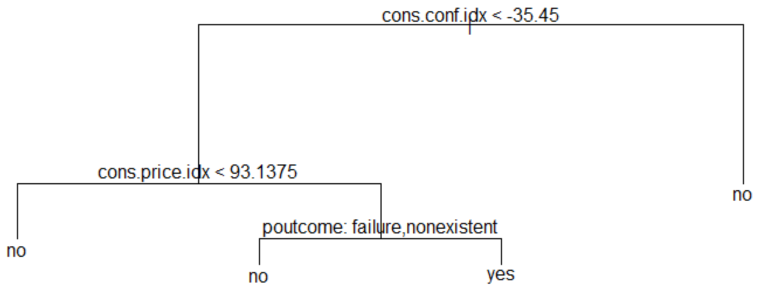
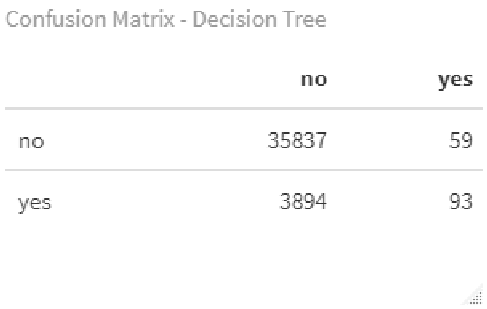


Figure 7: Desicion Tree graphical representation

The above graph is easy to be interpreted and implemented for prediction. More specifically, if the monthly consumer confidence index is lower than -35,45 then the final answer of a client is classified as “No – subscription”. Otherwise, we examine the consumer price index. If it is higher than 93,137 then we classify the answer of the client as “No- subscription”.Otherwise, the final parameter to examine is the outcome of the previous campaign. It may sound strange but if a previous campaign was successful this means that the client will not subscribe again.

On the other hand, if the previous campaign had failed or we had approached a new client it is more possible to gain a subscriber. This can be explained thinking the most observations we examine refers to clients with no previous outcome result.

Finally, we will examine the accuracy of this model.



Confusion Matrix - Decision Tree

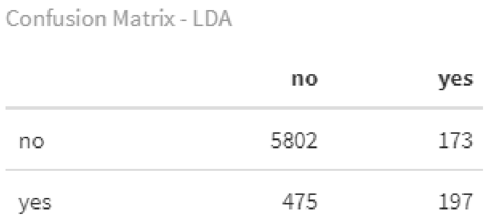
	no	yes
no	35837	59
yes	3894	93

Figure 8: Confusion Matrix of Decision Tree

The accuracy is **90,08%** , almost the same with randomness. We will examine below if the accuracy varies when using the cross-validation technique.

Linear Discriminant Analysis (LDA)

Logistic Regression does not coincide with LDA (due to different objective functions) but usually they are close.



Confusion Matrix - LDA

	no	yes
no	5802	173
yes	475	197

Figure 9: Confusion Matrix LDA

Same way as previously we can see the accuracy, precision and recall for LDA method. We found that the accuracy of the model is **89,87%**. Moreover, the recall in this method is better than the precision.

1.3 Methods comparison

To avoid overfitting of our data we must implement a cross validation. We used 6-folds cross validation. The graph below represents the goodness of the models we implemented to classify the clients. The red line represents the baseline scenario. If we randomly select an individual to call we have 90% possibility to refuse to subscribe to long term deposit. Having that in mind, a good model for prediction is a model with accuracy higher than simple randomness, namely, 0,90.

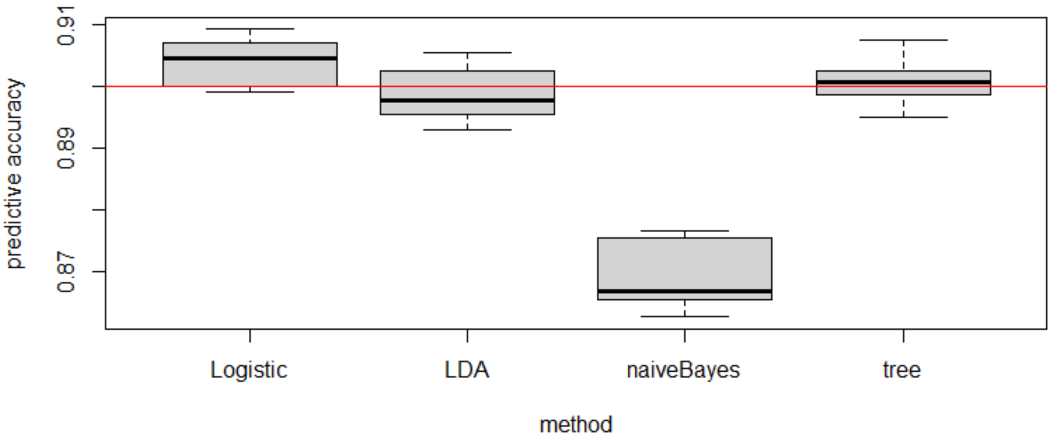


Figure 10: Methods comparison based on Accuracy

To conclude, the logistic Regression it seems to be the simplest and the most accurate method for classify the phone calls. For all experiments the accuracy of the model was quite better than 0,90. The next best method is the decision tree because half of times achieves a better accuracy than the randomness. Moreover, LDA most times predicts lower than 0,90 as we can see from the line at the corresponding boxplot. Finally, as we noted before naïve bayes is a very strict method and it is obvious from the graph as the accuracy is in the range of 0,87.

PART 2: Clustering

In this part of the report, we want to create well separated clusters in order to identify groups of individuals that may have similar characteristics. Finding these groups can lead to identify behavioral characteristics. The goal is to find observations that inside cluster are similar but observations from different clusters are different.

For this part we will use the following attributes: Age, job, marital, education, default, loan, housing, campaign, pdays, previous, poutcome. Below we will examine if all these variables contribute in finding actual clusters or we can eliminate the variables.

2.1 Which method to use?

The selection of the method for clustering is defined from the distance we select to use to find the similarities between the groups.

An important parameter that we must keep in mind when we are called upon to choose a distance is the type of data we have. Our dataset consists of mixed data (both continuous and categorical attributes) and therefore, we have to use the appropriate distance.

We used the “Gower” Distance because it combines the Euclidean distance for numeric attributes and the Simple matching distance for the binary ones. With this approach the distances can be compared.

2.2 How many clusters?

Selecting the number of clusters is done based on after – processing of results. We used the silhouette figure to help us identify the best number of clusters. For the best cluster we have high prices of the silhouette width. From the plot below we can assume that the best option is to use 3 clusters.

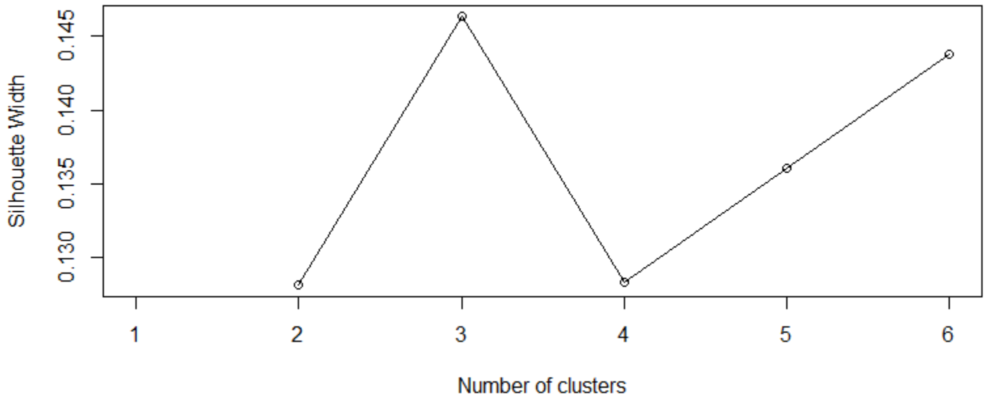


Figure 11: Silhouette values per number of clusters

2.3 Variables used for clustering

Variable selection is also performed to identify the variables that are most influential for determining cluster membership. Variables to be used are important. Their selection can be based on arguments that they shall take different values for different clusters. Too many variables can keep us away from seeing what is really going on

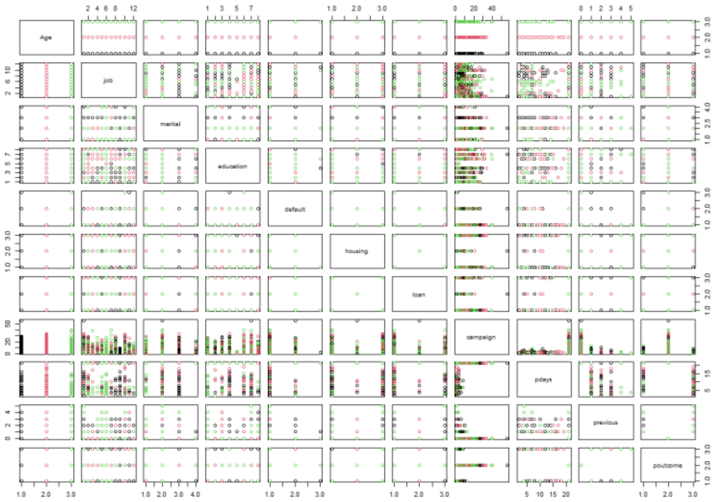


Figure 12: pairwise associations

After some processing we selected to exclude from the dataset the following variables: default, previous and poutcome. These variables do not give us enough information for create clusters and find similarities.

2.4 Evaluate the cluster

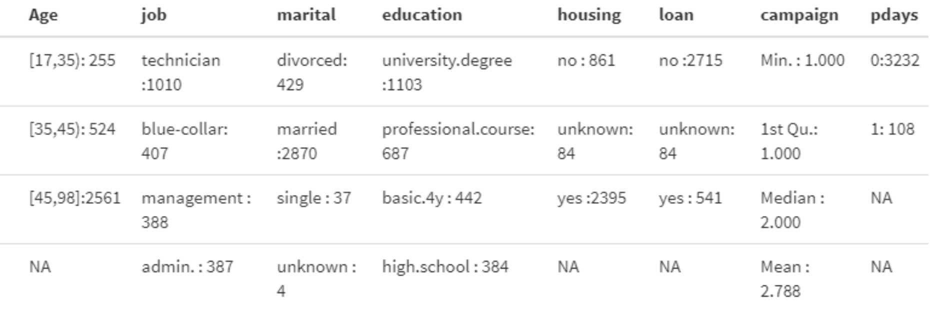
A measure to evaluate clusters is the Adjusted Rand Index between actual labels and clusters. We found out that the ARI for the cluster we constructed is:

2.5 Interpretation of the cluster

We may investigate the characteristics of each cluster with respect to the variables or other variables not used for clustering. Below we can see a summary of the 3 clusters we created.

We will try to describe each cluster.

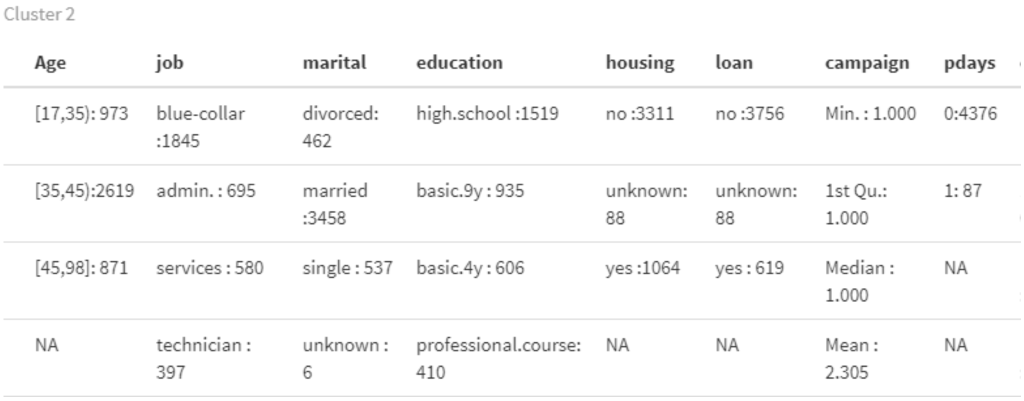
The first cluster refers to married clients of the age 45+, mainly technicians with a university degree who have a housing loan but not a personal loan and have never been called before.



Age	job	marital	education	housing	loan	campaign	pdays
[17,35): 255	technician :1010	divorced: 429	university.degree :1103	no : 861	no :2715	Min. : 1.000	0:3232
[35,45): 524	blue-collar: 407	married :2870	professional.course: 687	unknown: 84	unknown: 84	1st Qu.: 1.000	1: 108
[45,98]:2561	management : 388	single : 37	basic.4y : 442	yes :2395	yes : 541	Median : 2.000	NA
NA	admin. : 387	unknown : 4	high.school : 384	NA	NA	Mean : 2.788	NA

Figure 13: Cluster 1

The second cluster refers to married clients of age of [35,45) with a high school degree and with a housing and personal loan who have never been called before for the campaign.

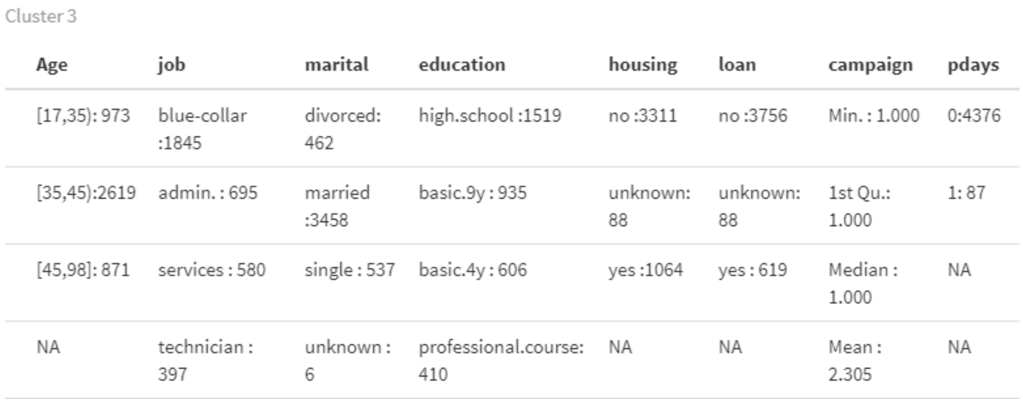


Cluster 2

Age	job	marital	education	housing	loan	campaign	pdays
[17,35): 973	blue-collar :1845	divorced: 462	high.school :1519	no :3311	no :3756	Min. : 1.000	0:4376
[35,45):2619	admin. : 695	married :3458	basic.9y : 935	unknown: 88	unknown: 88	1st Qu.: 1.000	1: 87
[45,98]: 871	services : 580	single : 537	basic.4y : 606	yes :1064	yes : 619	Median : 1.000	NA
NA	technician : 397	unknown : 6	professional.course: 410	NA	NA	Mean : 2.305	NA

Figure 14: Cluster 2

The third cluster refers to married clients of the age [35,45) who are blue collars who finished high school and with both housing and personal loan. Also, these clients have never been called before.



Cluster 3

Age	job	marital	education	housing	loan	campaign	pdays
[17,35): 973	blue-collar :1845	divorced: 462	high.school :1519	no :3311	no :3756	Min. : 1.000	0:4376
[35,45):2619	admin. : 695	married :3458	basic.9y : 935	unknown: 88	unknown: 88	1st Qu.: 1.000	1: 87
[45,98]: 871	services : 580	single : 537	basic.4y : 606	yes :1064	yes : 619	Median : 1.000	NA
NA	technician : 397	unknown : 6	professional.course: 410	NA	NA	Mean : 2.305	NA

Figure 15: Cluster 3

Considering the description of the above clusters we can assume that the variables marital and pdays finally do not contribute to clustering as there are the same for all cluster. This is also the reason that we observe a lot of oversubscriptions in the graph below.

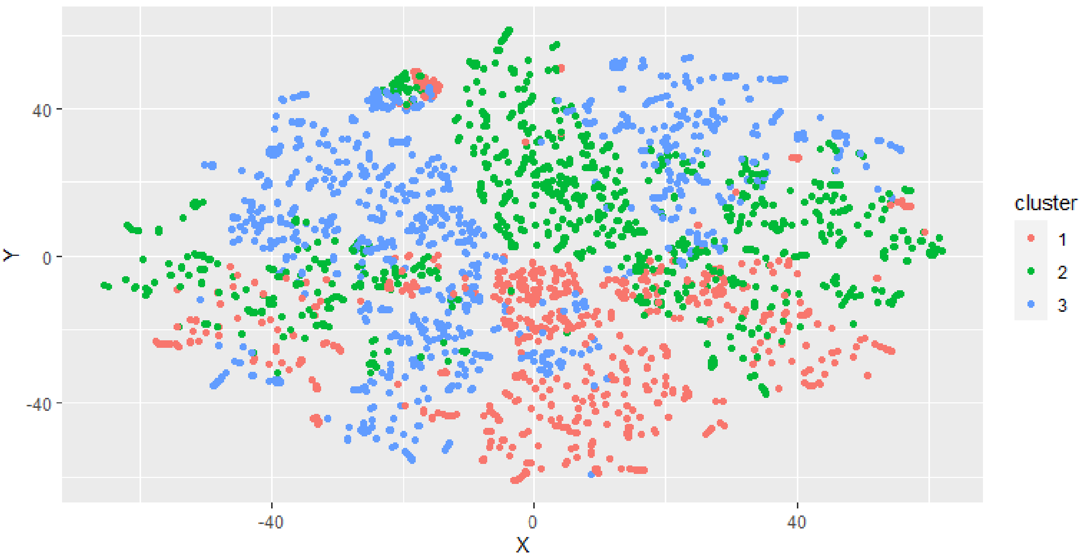


Figure 16: Graphical representation of the clusters