ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

STATISTICS FOR BUSINESS ANALYTICS II

Project II

Name: Ioannis Dekoulakos

A.M.: P2822110

Tutor: Dimitrios Karlis

Academic Year 2021 - 2022

Part time program
2nd Quarter

Πίνακας περιεχομένων

1. INTRODUCTION – DESCRIPTION OF THE PROBLEM

Marketing campaigns basically constitute a technique of outsourcing by organizations with the goal of improving the financial posture of their businesses. Telemarketing is way of straight marketing where a dealer come up to the buyer directly or through telephonic calls and influences them to purchase the products. Nowadays, telephone has been broadly used. It is cost effective and keeps the customers up to date.

In Banking sector, marketing is the backbone to sell its product or service. Banking advertising and marketing is mostly based on an intensive knowledge of objective information about the market and the actual client needs for the bank profitable manner. Banking advertising and marketing is mostly based on an intensive knowledge of objective information about the market and the actual client needs for the bank profitable manner.

Screening of targeted customers for telemarketing that are more likely to subscribe products will reduce the cost of marketing. Using available information and customer metrics, it is possible to build and establish automated protocols for selecting customers in advance. Such a protocol allows one to reduce the time and costs of campaigns and performing fewer and more effective phone calls will diminish client stress and intrusiveness.

The scope of this project is devited into two parts. The first is to build a model which will predict the outcome of the campaign held by the bank, namely whether a client will subscribe to term. The second part is to cluster the clients and to characterize the clusters.

For this purpose, a real dataset collected from one of the retail banks, from May 2008 to June 2010, in a total of near 40K phone contacts. These campaigns consist of information on phone calls where people were asked if they want to subscribe to a bank term deposit. The phone call is considered a success if a term deposit is made.

The dataset encompasses 4 main groups of features:

- Demographic Information of the clients — *age, job, marital, education, default, housing, balance, loan*
- Time Characteristics of the Call — *day, month, duration*
- Characteristics of the Campaign — *contact, campaign, pdays, previous, poutcome*
- Social and Economic context attributes — emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

SUBSCRIBED (target**)** — Indicates whether client has subscribed to a term deposit

2. PART 1 CLASSIFICATION

In this case we will use machine learning to understand pattern and predict classification or label, we use three predictive models to predict using training and testing data.

Model will be used: Logistic regression, Naive Bayes & Decision Tree.

These models are categorized as supervise learning, it is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately.

2.1 DATA CLEANING

Importing the records of telemarketing phones call created an R data frame which has 21 attributes and 39883 observations (columns & rows as well).

Upon importing the dataset, actions regarding changes in data types of variables, exclusion of specific columns without valuable information, as well as merging some levels of categorical variables were considered. All performed transformations will be analyzed below.

Some variables that were given as numeric may have been more useful as categorical variables. For instance, it makes more sense to see how *age* might influence the decision to subscribe across a few broad categories (more details with a histogram can be founded in Appendix 6, Figure 6.1). Additionally, the variable education will be merged into 4 levels and (illiterate, basic, intermediate, advance) as well.

According to the variable *pdays* more than 75% of the data set wasn't contacted previously by another campaign (more details with a histogram can be founded in Appendix 6, Figure 6.2), 999 means no previous contact, so it will be removed because it offers the same information as the variable *poutcome*, it does not offer any extra valuable information (more details with a barplot can be founded in Appendix 6, Figure 6.3).

Table 1. List of the attributes from the dataset after cleaning

variable	*Description*	*Type*
age	Age of the client	Factor
job	Type of job	Factor
marital	marital status	Factor
education	education of the client	Factor
default	has credit or default?	Factor
housing	Indicates whether client has a housing loan	Factor

loan	Indicates whether client has a personal loan	Factor
contact	Type of call contact communication	Factor
season	Season of call	Factor
day_of_week	day of call	Factor
duration	duration of call (sec)	num
campaign	Number of contacts made during current campaign for this client	num
poutcome	outcome of the previous marketing campaign	Factor
emp.var.rate	Employment variation rate	num
cons.price.idx	Consumer price index	num
cons.conf.idx	Consumer confidence index	num
euribor3m	Euribor 3 month interest rate	num
nr.employed	Number of employed citizens in Country	num
previous	number of contacts performed before this campaign	num
SUBSCRIBED	has the client suscribed the term deposit?	Factor

2. 2 DATA DIVISION, MODELLING, AND RESULTS

We use 6-k cross validation, we create 6 folds of observations of almost equal size, and we leave out one-fold each time, use the rest for model fitting and the one left out for prediction. We will use the same folders for each method.

2.2.1 Logistic Regression

Variable selection techniques should be used to select the most important features for this model. The stepwise procedure with AIC method generated a model ('model_aic') using 15 of 20 input variables (dropping the *variables job, housing, loan, previous, nr.employed*) , with residual deviance: 15619 on 39848 degrees of freedom and AIC: 15689.

The AIC tries to select the model that most adequately describes an unknown, high dimensional reality. This means that reality is never in the set of candidate models that are being considered. So, AIC is better choice than other method for predictions (when the most variables are categorical lasso 'kills' them).

Many machine learning algorithms can predict a probability or scoring of class membership, and this must be interpreted before it can be mapped to a crisp class label.

This is achieved by using a threshold where all values equal or greater than the threshold are mapped to one class and all other values are mapped to another class.

Logistic Regression has been analyzed in previous work. The probability

$$p_i = \frac{\exp\,(\beta_\iota)}{1+\exp(\beta_\iota)}$$

$C_{i\,,}$ i= 1, … n., x_i is a vector and β a vector of coefficients.

will be in the unit interval, i.e. [0,1].

For calculation of the threshold, we will use the ROC curve.

For this purpose, we will split the data to train and test set 70-30% as well. The prediction of this action will be used to create the ROC curve.

ROC curves are frequently used to show in a graphical way the connection between sensitivity and specificity for every possible cut-off for a test or a combination of tests

So, we need to plot the true positive rate against the false positive rate for the logistic classifier at a variety of thresholds. The optimal cut off point would be where “true positive rate” is high and the “false positive rate” is low.

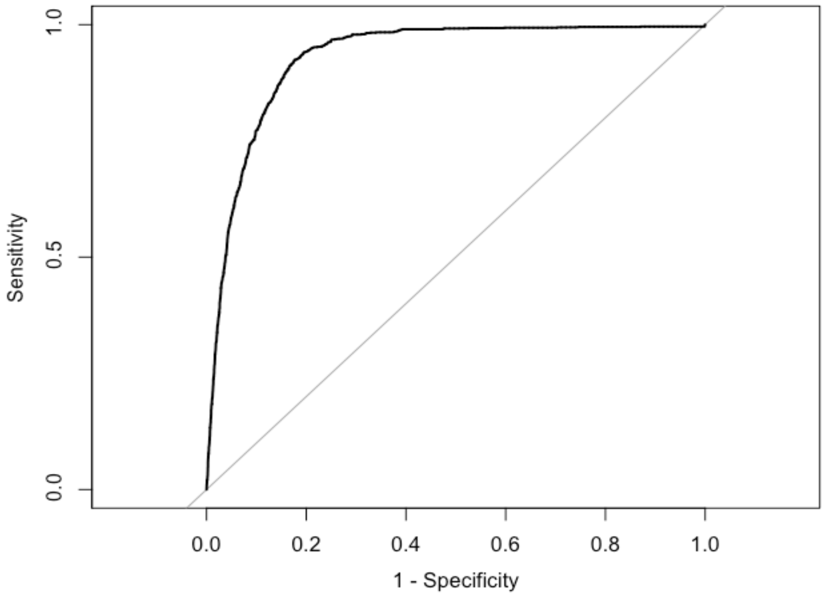


Figure 1 ROC curve

In this case the 0.095 (around) probability cut off points will used to classify observations (as the most optimal trade off between specificity and sensitivity

AUC is the area under the ROC curve. AUC values indicate the success of the model predicting the two classes. If the AUC value is close to 1, it means that the classification model can predict the two classes well. In this case area under the curve is 0.93

LOGISTIC REGRESSION WITH 6-FOLD CROSS VALIDATION

Then, we are ready to build the first model, 6-fold cross validation will used to achieve comprehensive results of the prediction tested on different parts of the dataset.

Evaluation of the logistic regression :

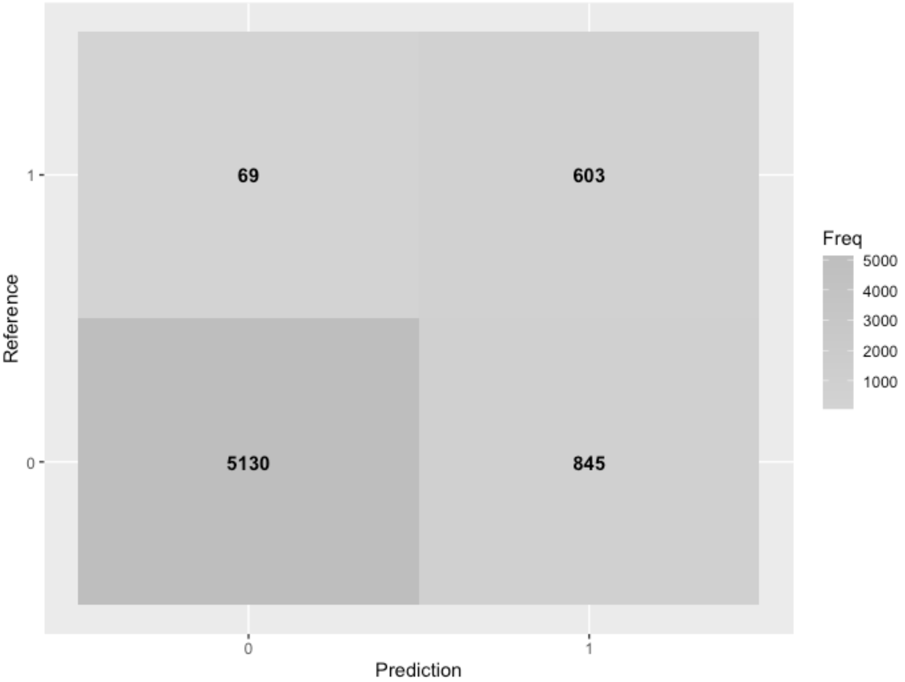


Figure 2 Confusion matrix of Logistic regression

We managed to predict correctly 5130 ‘no’ and 603 ‘yes’ with 0.85% accuracy. At a first glimpse, we can say that the model predicts worse than luck since the data is 90-10% (no-yes as well), but we need to mention that the bank cares to correctly predict the clients which make the subscription. Also, the result is related to threshold which we set, with a higher value threshold, overall, we will take a better accuracy but lower Sensitivity.

Sensitivity is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate. So, in this model the ability to correctly predict the positive class from the total actual-positive class is 0.89

Specificity is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate. So, in this case the ability to correctly predict the negative class from the total actual-negative class is 0.84

The average of Sensitivity & Specificity called Balanced Accuracy = 0.87. It is especially useful when the classes are imbalanced, i.e., one of the two classes appears a lot more often than the other.

2.2.2 Decision Tree

We will use the same folders with the logistic regression for cross validation.

With the rpart function we build the decision tree, in which we use the estimated value that bring the best results and finally we plot the tree.

A decision tree will choose a variable that has the highest information gain and it generate a new node for this until either no carriable is remaining or the cases associated with this leaf node all have the same target value. Splitting will happen at a condition where it maximizes the homogeneity within resulting groups. Outliers will have little influence on the splitting process.

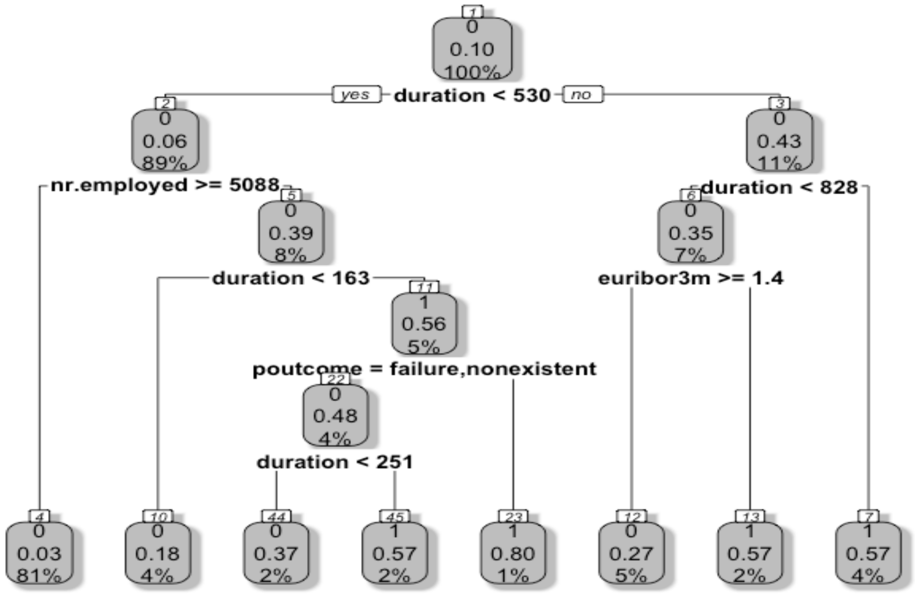


Figure 3 Plot of Decision Tree

Some examples

the response is 0.03 when

- duration < 530
- nr.employed >= 5088

the response is 0.27 when

- duration >= 530 & duration < 828
- euribor3m >= 1.4

The 0.095 probability cut off point was used to classify observations (as the most optimal tradeoff between specificity and sensitivity).

After we trained the model & we used to predict using our data test.

Evaluation of Decision Tree:

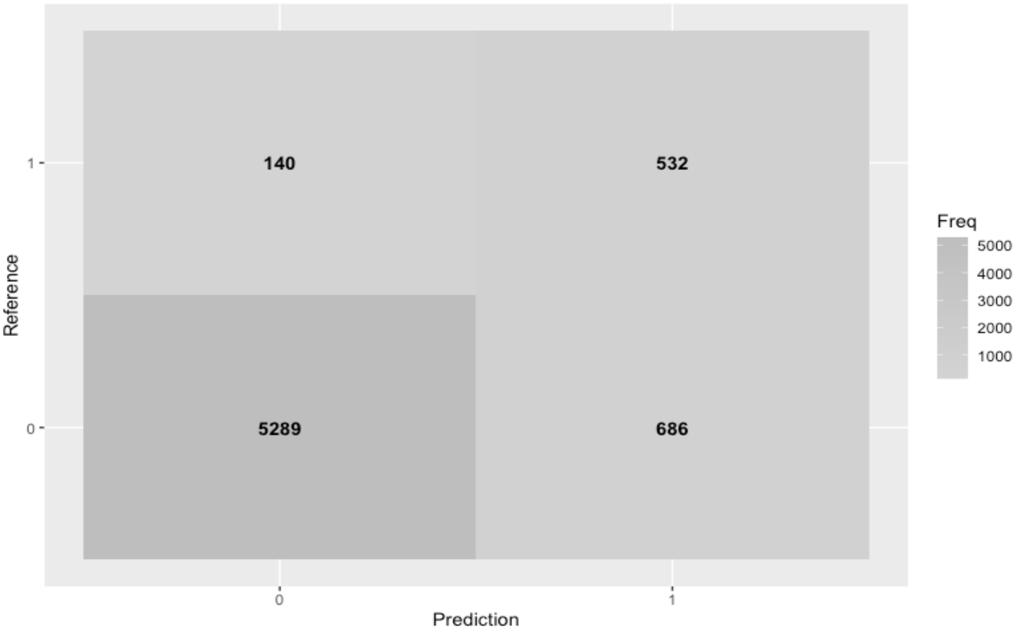


Figure 4. Confusion matrix of Decision Tree

The model managed to predict correctly 5289 ‘no’ and 532 ‘yes’ with 0.87 accuracy. As we mention the scope is to correctly predict the positive clients, the Sensitivity is 0.79 & Specificity 0.87. So, the balance accuracy is 0.83

2.2.3 Naive Bayes

It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge.

We will use the same folders with the previous methods for cross validation. First, we remove the variable *euribor3m* since it is strongly correlated (almost perfectly colinear) with *emp.var.rate* and *nr.employed* and it would create "noisy".

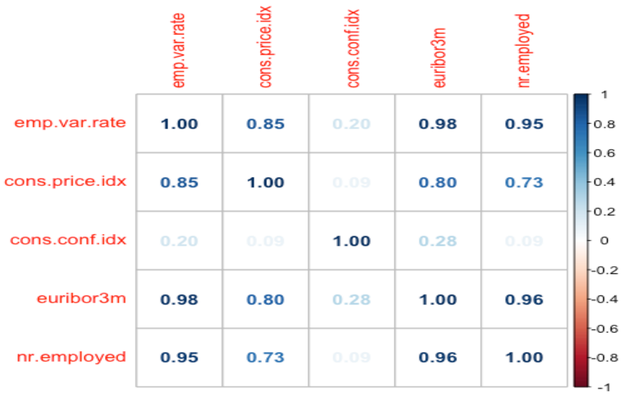


Figure 5 Correlation matrix of social values

Evaluation of the Naive Bayes results:

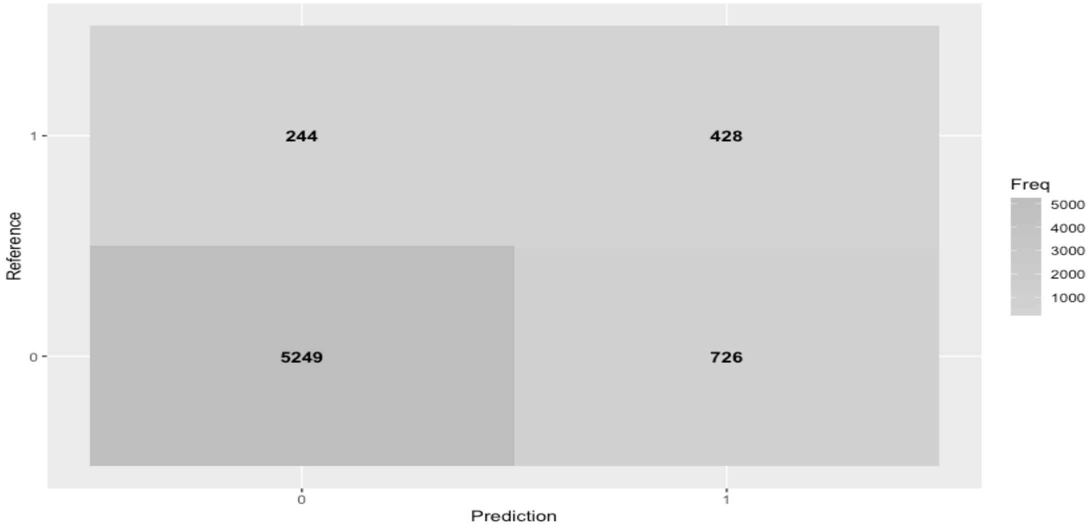


Figure 6. Confusion matrix of Naive Bayes

The model managed to predict correctly 5249 'no' and 428 'yes' with 0.84 accuracy. As we mention the scope is to correctly predict the positive clients, the Sensitivity is 0.64 & Specificity 0.87. So, the balance accuracy is 0.75.

Example: suppose we are interested in only for *emp.var.rate ,cons.price.idx* and the *emp.var.rate* is 0.8 & *cons.price.idx* is 93

A-priori probabilities:

```
Y        0          1

0.90025876     0.09974124
```

emp.var.rate

```
Y           [,1]              [,2]

 0        0.2708666 1.483581

 1        -1.1990950 1.746129
```

cons.price.idx

```
Y           [,1]                [,2]

  0       93.58932 0.5564111

  1       93.20363 0.6047088
```

To classify it is enough to calculate:

P (emp=0.8 | 0) * P(cons=93|0) * P (0)

$f(0.8, \mu=0.27, \sigma=1.4) * f(93, \mu=93.5, \sigma=0.55) *0.90$

P (emp=0.8 | 1) * P(cons=93|1) * P (1)

$f(0.8, \mu=1.1, \sigma=1.74) * f(120, \mu=93.2, \sigma=0.6) *0.099$

*For categorical, the probability calculated directly from the data

2.3 COMPARISON OF THE MODELS AND CONCLUSIONS

Overall, it seems that Decision tree has the higher score regarding the accuracy and Naïve Bayes has the lower.

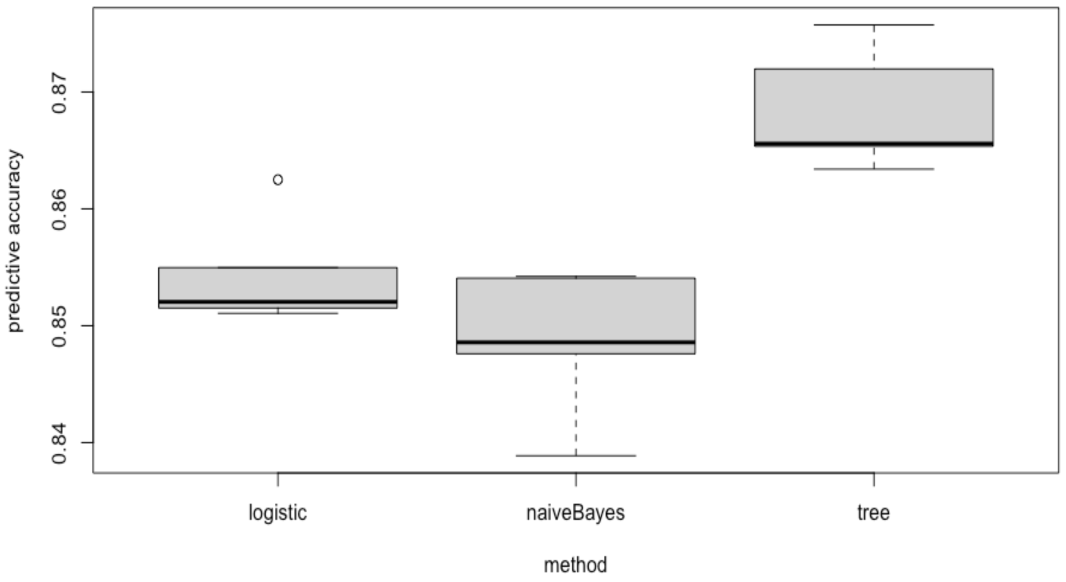


Figure 7. Accuracy of each model

But as we mentioned, the scope for the bank is to correctly predict the clients which subscribed. So, out of the 3 different models presented in this report, **logistic regression** did the best job in determining true positive values and has the best balanced accuracy.

- Sensitivity 89%
- Specificity 84%
- Balance accuracy 87%

Then follows Decision Tree method with:

- Sensitivity 79%
- Specificity 87%
- Balance accuracy 83%

Final the Naïve Bayes method with:

- Sensitivity 64%
- Specificity 87%
- Balance accuracy 75%

3. Part 2 Clustering

Clustering is a machine learning technique used to find structures within data. Cluster analysis is also known as unsupervised learning, no “truth” is available to verify results.

The goal is to find groups of observations which looks similar into the cluster but difference with the observations of other clusters.

We use the variables: *age, job, marital, education, default, housing, loan, campaign, pays, previous, poutcome* to cluster the clients and to characterize the clusters.

We remove the variable pdays since more than 75% of the data set wasn’t contacted previously by another campaign so, it offers the same information as the variable *poutcome*, it does not offer any extra valuable information.

To group observations together, we need to specify some notion of distance between observations.

Due to the technical limitations, we will take a sample 10000rows (only for the analysis of part 2)

A famous choice for clustering is Euclidean distance. Euclidean distance is only valid for continuous variables, so it is not applicable here. We must use a distance metric that can handle mixed data types. In this case, we will use Gower distance.

Gower's distance can be used to measure how different two records are. The key idea is to measure for each variable its distance in a range between (0, 1) and take the average.

For quantitative variables use the Manhattan distance and for the categorical variables converted them into k binary columns and then and then compute some measure for binary variables. (reference). In R we calculated it using the daisy function.

After we have created a distance matrix, we use Partitioning Around Medoids to split the data set of n objects into k clusters. The PAM algorithm searches for representative objects in a data set group and then assigns each object to the closest group to create clusters. It tries to minimize the sum of dissimilarities between the objects in a cluster and the centre of this cluster.

Specifically, it chooses some random elements to become a medoid, then using the distance matrix, which we have created, assign every element to its closest medoid and for each cluster identify the observation which would have the lower average distance if it became as medoid, so make this observation the new medoid until anyone medoid changes.

Now it’s time to select the number of clusters, for this purpose we use silhouette value

The average value of how similar an observation is to its own cluster compared its closest neighboring cluster over all data of the entire dataset is a measure of how appropriately the data has been clustered.

Thus, silhouette plots and averages may be used to determine the natural number of clusters within a dataset.

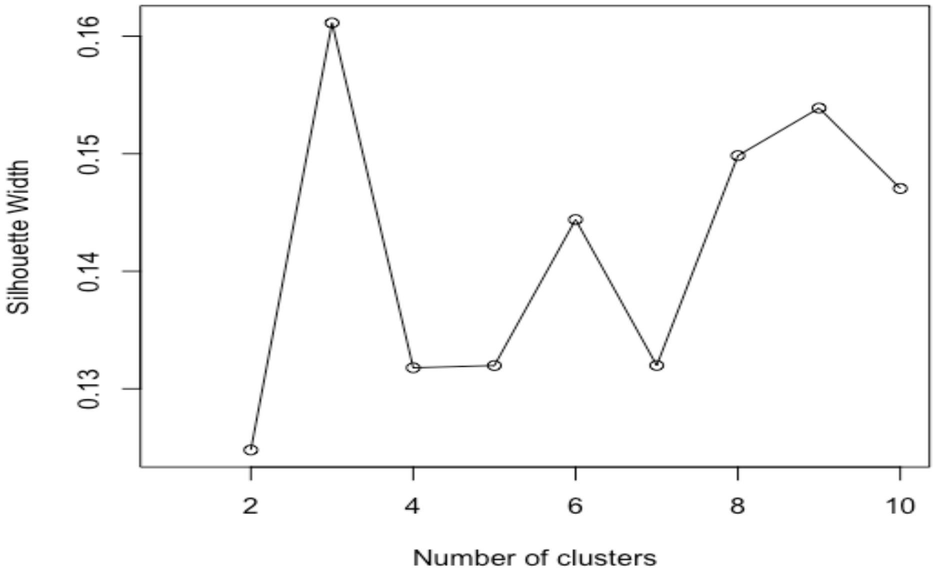


Figure 8. Silhouette plots

We can see that 3 clusters yields the highest value since for this number the Silhouette take the higher value.

After running the PAM algorithm with 3 clusters we can take some summary of each cluster.

- The majority of 1st cluster is 25-50 years old, has admin job, is single with university degree level of education, has no default (default is a term for when a customer fails to pay any kind of debt, he/she has, even if there are some missing payments.), has housing loan but not any other loan, has low number of contacts for this campaign and non-exist in the previous campaign and has very low amount of previous contact.

- The majority of 2nd cluster is 25-50 years old, has blue-collar job, is married with high school level of education, has no default, has not housing or any other loan, has low number of contacts for this campaign and non-exist in the previous campaign and has very low amount of previous contact.

- The majority of 3rd cluster is 25-50 years old, has blue-collar job, is married with high school level of education, has no default, has housing loan but not any other loan, has low number of contacts for this campaign and non-exist in the previous campaign and has very low amount of previous contact.

According to the variables campaign, previous, it seems that there is not any significantly difference between the means of clusters, so may not be useful, something which would be confirmed with an Anova.

Also, in all clusters we observe that in the variable age the observations are gathered almost in the same way, but we have the most observations in the sample in this level. Same the variables default, loan, poutcome maybe does not offer any value information for the clustering. (Reminder, we have taken only 10k like a sample and these levels appears the most times.)

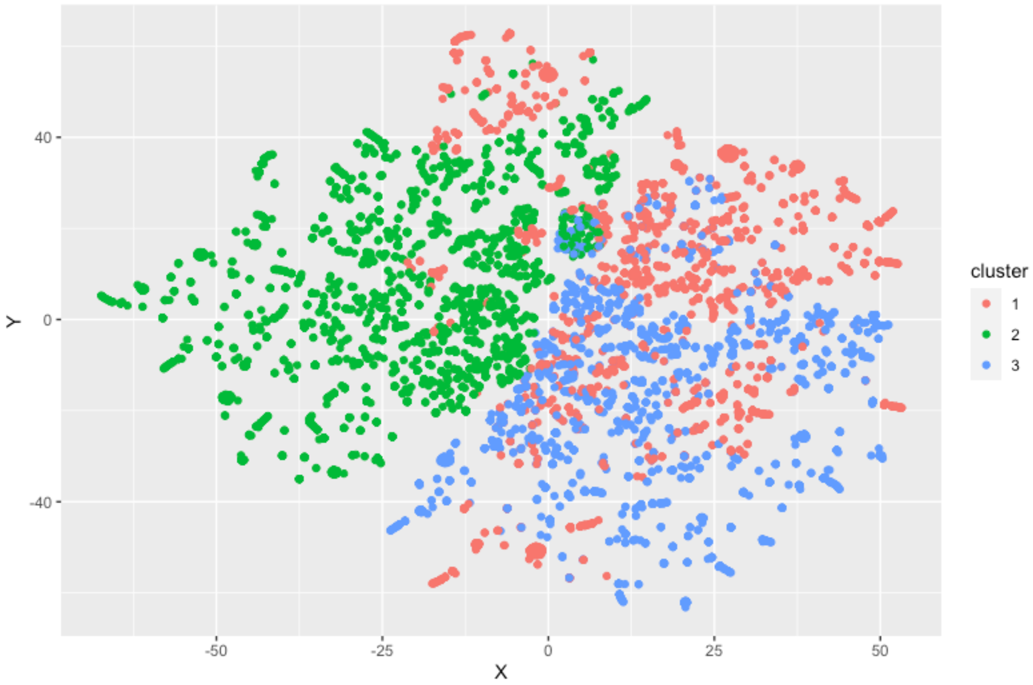
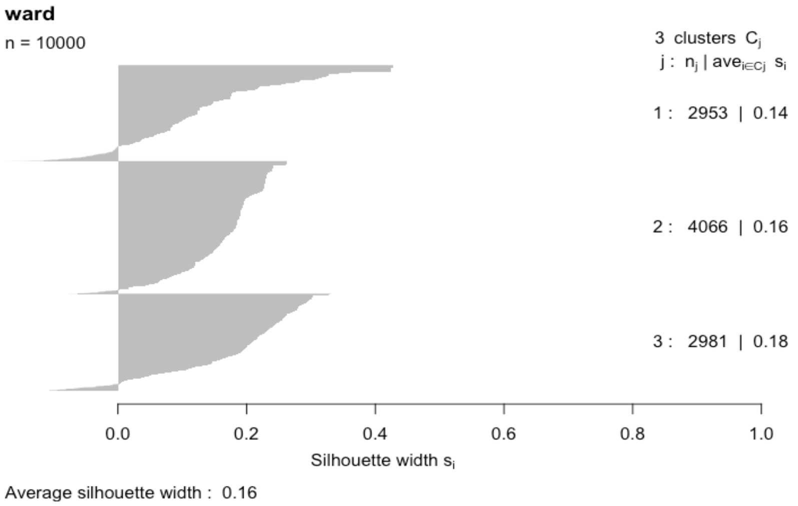


Figure 9. Cluster's plot

Final the cluster 1 seems to stand out from the rest while the other two are more difficult to distinguish.



The Average silhouette width value is 0.16 near to 0. It is means that the datum is on the border of two natural clusters. Expected, since we have overlapping clusters

- The first cluster has 2953 obs.
- The second has 4066 obs.
- The third has 2981 obs.

The first cluster has the lower silhouette value maybe some observations could be made in a neighbouring group.

reference*

https://stats.stackexchange.com/questions/55798/what-is-the-optimal-distance-function-for-individuals-when-attributes-are-nomina/55802#55802

Manhattan distance*

https://en.wikipedia.org/wiki/Taxicab_geometry

6. APPENDIX

Only for the first part of data cleaning.

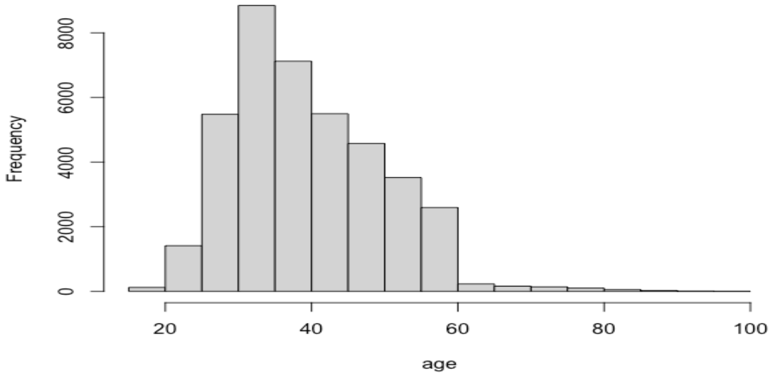


figure 6.1 Histogram for Variable -Age

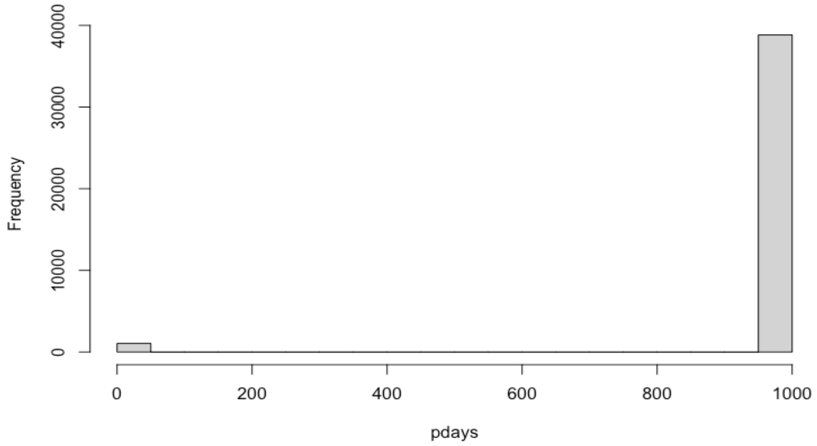


figure 6.2 Histogram for Variable -Pdays

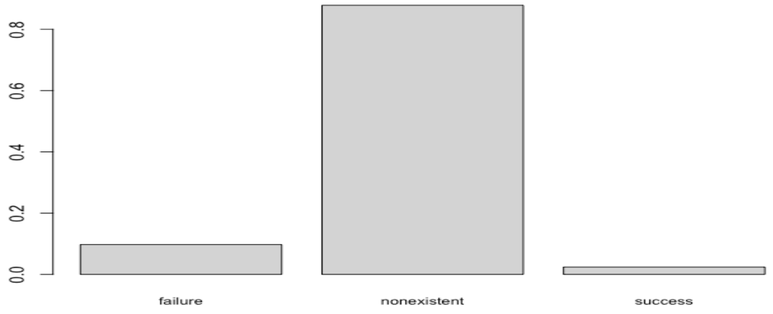


figure 6.3 Barplot for Variable -Poutcome