Client Settings as a Determinant of Bank Telemarketing Success Rate on Long-Term Deposits

Project II: Supervised and Unsupervised Data

Anastasios Moraitis

p2822124

Statistics for Business Analytics II, MSc in Business Analytics

School of Business, AUEB, Athens, Greece

April 2022

Abstract

This study aims to determine the factors affecting the ability to predict the decision of a given bank customer towards selecting a long-term deposit product. It concentrates the key factors that best predict the possibility the customer subscribes to the product and, in general, which variables contribute to a successful contract. It also aims to cluster a subset of data (client and other attributes) to evaluate the ability to explain the behavior of different clients on the subscription. To this end, we draw 49 months of real data collected from one of the retail banks, in a total of near forty thousand phone contacts. These data are split between bank client data, data related with the last contact of the current campaign and other attributes, and social and economic context attributes. Overall, we find that a mix of characteristics impacts customer acquisition.

Keywords: Telemarketing sales, Marketing strategy, Long-Term Deposits, Prediction, Classification, Clustering

Client Settings as a Determinant of Bank Telemarketing Success Rate on Long-Term Deposits

Project II: Supervised and Unsupervised Data

To effectively predict potential clients ready to invest in new products businesses include marketing selling campaign, and, more precisely, direct marketing. Telemarketing is highly likely to be included in this set of tactics. Telemarketing has always been a popular way to expand a business, but the use of personalized scripts and the wide availability of phone numbers for consumers has been an easy way for individuals to end calls before they begin. When done correctly, however, telemarketing can be a useful advertising medium for businesses. This project aims to help banks to increase the accuracy of their customer profiling through predicting as well as identifying a group of customers who have a high probability to subscribe to a long-term deposit. The remainder of this paper proceeds as follows. Section II and III presents the materials and methods used to achieve the objective of this research, Section IV presents the experimental results of the classification processes used for prediction, Section V describes the clustering method used and presents the results, and finally Section VI concludes with some indication for future work.

2. Data

Bank telemarketing data has been made available. The dataset contains subscription statuses obtained after each call to the targeted customer, thus making it easy to attach bank client attributes, features related with the last contact of the current campaign and other attributes, with social and economic ones among them. The data have been cleaned and transformed and we will end up using the final data used by Moraitis (2022). In addition, we will run an ad-hoc bivariate analysis on some of the variables to determine whether further elimination of features or categorizations should be made in our data. That will not only help us deduct features, but also provide sufficient insights for the later clustering process.

The additional data treatment is reported early in this report for clarity, and further transformations based on the EDA will be presented in the next section. It is known that transformations on the original data will reduce accuracy of the classification models, but it is something we will not avoid, because we want our models to be interpreted by humans, and not end up with models difficult to explain.

In Tables I and II, the descriptive measures of factorial and numeric variables have been put together to give us insights about each of the feature. These are also attached to give us a rough description of all the variables given and are used throughout are process.

Table 1 - Descriptive statistics of factorial variables

Feature	*Categories - Proportion (Count)*					
job	admin.	blue-collar	entrepreneur	services	student	technician
	25.25% (10067)	23.22% (9260)	3.63% (1449)	9.83% (3920)	1.87% (745)	16.48% (6572)
	housemaid	management	retired	self-employed		
	2.63% (1048)	7.2% (2869)	3.94% (1570)	3.52% (1402)		
marital	divorced	married	single			
	11.22% (4472)	61.12% (24369)	27.66% (11030)			
education	Primary	Secondary	Tertiary			
	32.56% (12981)	24.06% (9592)	43.38% (17298)			
default	no	yes				
	99.99% (39867)	0.01% (4)				
housing	no	yes				
	46.36% (18486)	53.64% (21385)				
loan	no	yes				
	84.4% (33651)	15.6% (6220)				
contact	cellular	telephone				
	62.9% (25080)	37.1% (14791)				
dow	Fri	Mon	Thu	Tue	Wed	
	19.06% (7599)	20.65% (8232)	20.85% (8315)	19.68% (7845)	19.76% (7880)	
poutcome	failure	nonexistent	success			
	9.75% (3889)	87.82% (35015)	2.43% (967)			
subscribed	no	yes				
	90% (35885)	10% (3986)				
pdays (contacted)	In first 10 days	After 10 days	First time			
	2.24% (894)	0.4% (158)	97.36% (38819)			
age	(16,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]

	0.3% (121)	17.28% (6888)	40.04% (15966)	25.26% (10071)	15.33% (6113)	0.97% (388)
previous	Never contacted	Less_than_or_3_times	More than_3_times			
	87.82% (35015)	12.1% (4823)	0.08% (33)			
quarter	Q1	Q2	Q3	Q4		
	21.96% (8756)	31.59% (12596)	45.78% (18252)	0.67% (267)		

Table 2- Descriptive statistics of numeric variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
campaign	39,871	2.589	2.801	1	1	3	56
emp_var_rate	39,871	0.131	1.572	-3.400	-1.800	1.400	1.400
cons_price_idx	39,871	93.553	0.572	92.201	93.075	93.994	94.465
cons_conf_idx	39,871	-40.463	4.611	-50.000	-42.700	-36.400	-26.900
euribor3m	39,871	3.710	1.690	0.634	1.405	4.961	5.045
nr_employed	39,871	5,173.223	64.626	4,992	5,099.1	5,228.1	5,228

3. Descriptive Analysis

Univariate

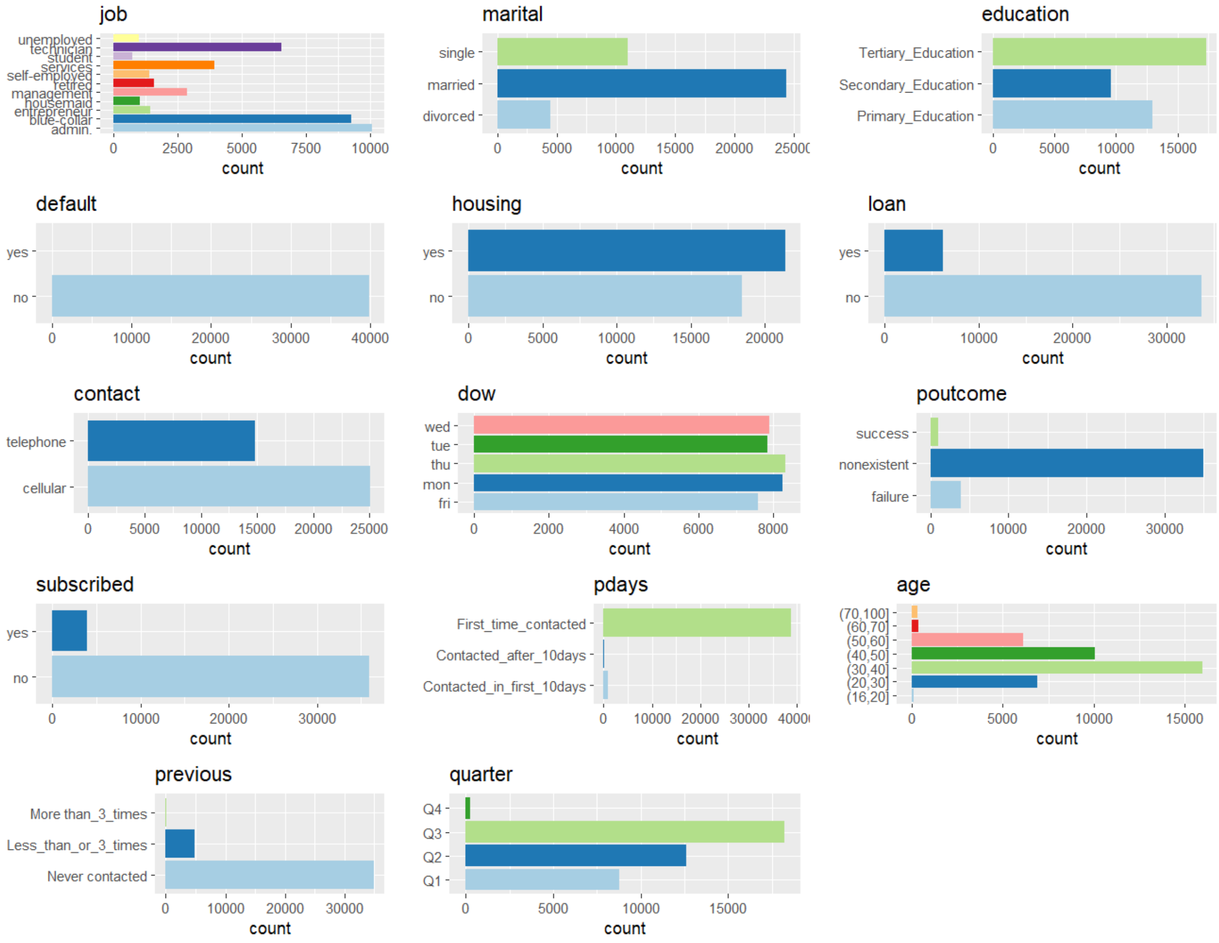
We are looking again at each variable, respectively, as a reminder to our previous analysis. The duration feature is already removed because it is only known after the fact. Let's now go over the numeric features.

[Frequency of features here]

In Figure A, we notice that our numeric variables are related to social end economic context. These have to do with employment variation, consumer price index, Euribor 3-month rate, etc. We notice that there is no sense of normality in these data, even from the figure, but we hereby attach the table with the normality tests.

Concerning the factorial data, as introduced from our previous analysis, or imported as is from the source, we present their distribution throughout the dataset. Our categorical data are presented now. Overall, we notice that we have a lot of features with their values categorized. Response Rate is significantly higher for prospects that have been contacted within 27 days as compared to first time contacts.

Categorical variables distribution



Bivariate

We are now going to compare various features together. And most importantly we are going to look if it explained (related) by another feature in our dataset. But first we are going to look at the correlation between the numeric variables.

[Correlograms - add subscribed]

From Figure XY, there is a relationship between the features *pdays* and *previous*. Although a good practice seems to be to merge these two, it is not advisable because we would lose the ability to look at those two separately in terms of our prediction. A similar correlation exists between the latter (*previous)* and *poutcome*. This was expected. If no previous contact is made, there is no previous outcome. The 3 variables are tightly coupled and may cause issues in predicting the response variable, but we are willing to let the models decide which of them are not useful in our prediction.

We now aim to find height correlations in our, so called, independent features. From the remaining, we noticed a strong correlation exists for our *poutcome* feature compared with *previous* and not *pdays*. For evidence we show that in Figures 2 and 3. This is an alert for us, and we will remove the variable so that we will not attempt a model with aliased coefficients in it.

Figure 1- poutcome versus pdays

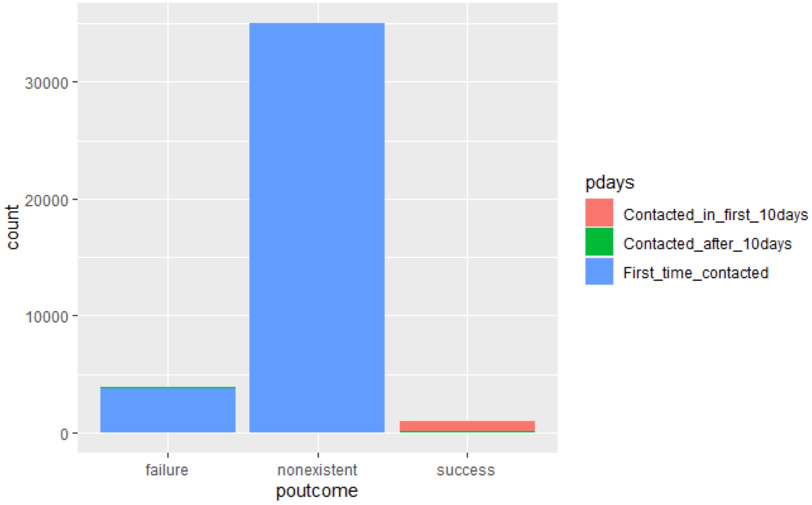
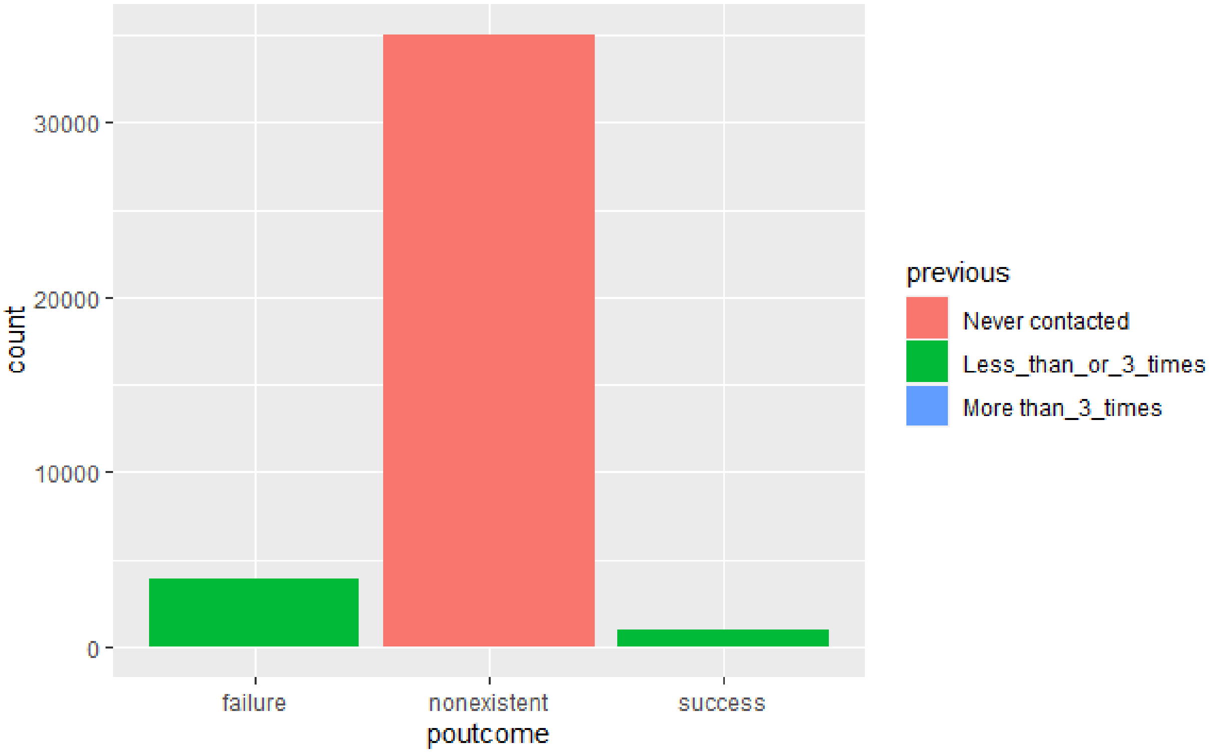


Figure 2- poutcome versus previous



Now, let us investigate the correlation between the ordinal/categorical variables and our target variable. That will give us enough insight into which variables are useful in predicting the action of the client. In figure XYI, we notice that there is a strong correlation between XYI and subscribed

4. Classification

In this section we will try to expand on 3 classification algorithms and decide on the one that best predicts our target variable. The algorithms in test are: Logistic Regression, Naïve Bayes, and Decision Tree based Classifications. We are going to expand on the results on this section.

The data used for this part are randomly split and used as such: 70% for training of the various models, and 30% used for predicting/fitting. The given dataset is large enough, so the proportion selected for training is large enough to assess each distinct case. The function *createDataPartition* from *caret* package is used, as defined in Hyndman and Athanasopoulos (2013), to create balanced splits of the data. For our dataset, it will use the target variable for the random sampling so that it will occur within each class and will preserve the overall class distribution of the data.

We remember our investigation about Zero Variance and Non-Zero Variance from Moraitis (2022), in which we decided not to remove any variables, after the results of the diagnostic which gave us only near zero variance for two of our variables. A second step is to investigate Multicollinearity in our data. We will again use the Generalized Variance inflation factor (GVIF) to determine whether we should remove or not variables. We except the same results as in our previous analysis. So, the results of that diagnostic as defined by Fox, J. et al (2018) will be the same and they are referenced only in the Appendix. That process will leave us with a model of all the variables which are presently available, minus *euribor3m.*

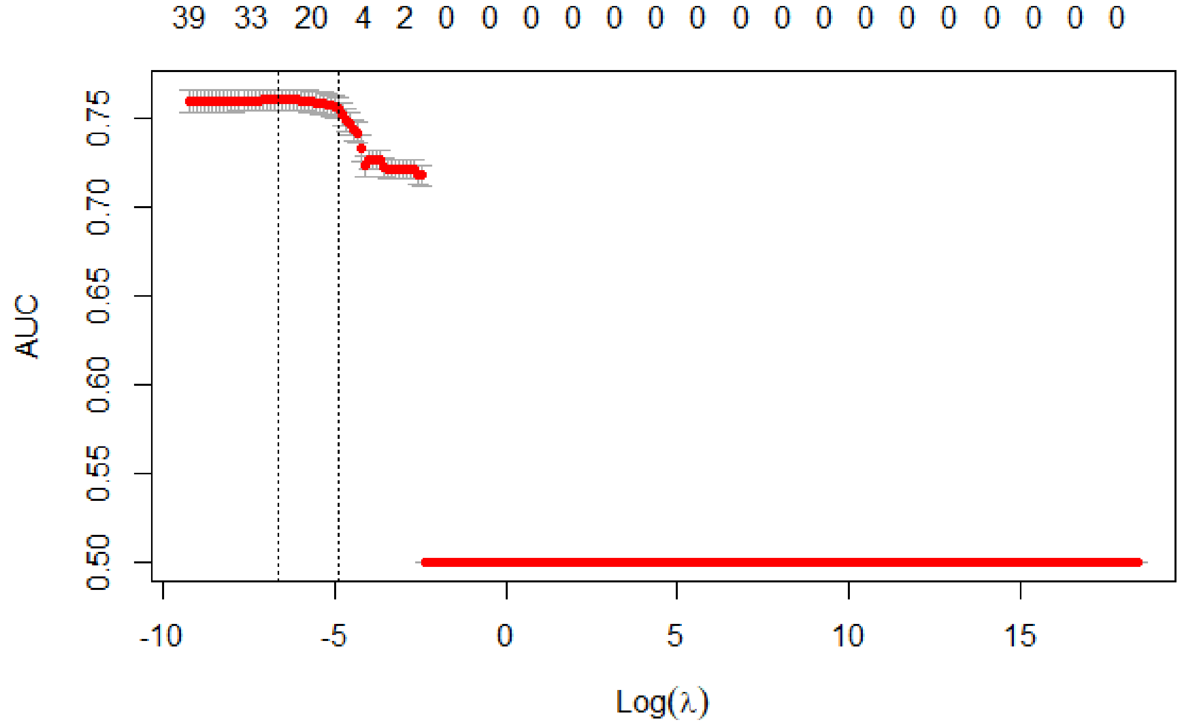
The base logistic regression model is that with the equation:

$$
\begin{aligned}
\text{subscribed} \quad & \sim \mathit{Bernoulli}\left(\text{prob}_{\text{subscribed=yes}} = \hat{P}\right) \\
\log\left[\frac{\hat{P}}{1-\hat{P}}\right] \quad & = 16.88 - 0.22\left(\text{age}_{(20,30]}\right) - 0.36\left(\text{age}_{(30,40]}\right) - 0.47\left(\text{age}_{(40,50]}\right) - \\
& \quad 0.31\left(\text{age}_{(50,60]}\right) - 0.15\left(\text{age}_{(60,70]}\right) + 0.09\left(\text{age}_{(70,100]}\right) - 0.06(\text{campaign}) - \\
& \quad 0.08\left(\text{pdays}_{\text{Contacted}_{\text{after}_{\text{10days}}}}\right) - 1.74\left(\text{pdays}_{\text{First}_{\text{time}_{\text{contacted}}}}\right) \\
& -0.54\left(\text{previous}_{\text{Less\_than\_or\_3\_times}}\right) - 0.26\left(\text{previous}_{\text{More than\_3\_times}}\right) - \\
& \quad 0.21\left(\text{emp}_{\text{var}_{\text{rate}}}\right) + 0.32\left(\text{cons}_{\text{price}_{\text{idx}}}\right) + 0.02\left(\text{cons}_{\text{conf}_{\text{idx}}}\right) - 0.01\left(\text{nr}_{\text{employed}}\right) - \\
& \quad 0.32(\text{job}_{\text{blue-collar}}) - 0.14\left(\text{job}_{\text{entrepreneur}}\right) - 0.1(\text{job}_{\text{housemaid}}) - 0.11\left(\text{job}_{\text{management}}\right) + \\
& \quad 0.11(\text{job}_{\text{retired}}) - 0.04\left(\text{job}_{\text{self-employed}}\right) - 0.16(\text{job}_{\text{services}}) + 0.12(\text{job}_{\text{student}}) - \\
& \quad 0.07(\text{job}_{\text{technician}}) - 0.02\left(\text{job}_{\text{unemployed}}\right) + 0.03(\text{marital}_{\text{married}}) + 0.05\left(\text{marital}_{\text{single}}\right) - \\
& \quad 0.03\left(\text{education}_{\text{Secondary}_{\text{Education}}}\right) + 0.04\left(\text{education}_{\text{Tertiary}_{\text{Education}}}\right) - 8.56\left(\text{default}_{\text{yes}}\right) \\
& -0.06\left(\text{housing}_{\text{yes}}\right) - \\
& \quad 0.07\left(\text{loan}_{\text{yes}}\right) - 0.45\left(\text{contact}_{\text{telephone}}\right) - 0.14(\text{dow}_{\text{mon}}) + 0.1(\text{dow}_{\text{thu}}) + \\
& \quad 0.06(\text{dow}_{\text{tue}}) + 0.19(\text{dow}_{\text{wed}}) + 0.27\left(\text{quarter}_{\text{Q2}}\right) - 0.54\left(\text{quarter}_{\text{Q3}}\right) - \\
& \quad 0.68\left(\text{quarter}_{\text{Q4}}\right)
\end{aligned}
$$

Classification with Logistic Regression

We are now deriving from that model to find a good model for prediction. We will use Lasso algorithm to decide on the features that are important. We are not going to use BIC or AIC, like we did in our previous analysis, because currently we do not need a strict penalty for additional features. In the end, we want a model for prediction, and less variables will reduce the accuracy if they predict the target one.

We also applied LASSO Regression on the sparse matrix of our data and used the λ=exp(x), x is the 1 standard error value of the plot Binomial Deviance-Log(λ), which is reported from cross validation. The mean is somewhat sensitive to the specific run, so our selection is the most regularized model such that error is within one standard error of the minimum. Lasso regression was cross validated with AUC as a metric because we are only interested on the classification of the target variable. We took the features of which the beta coefficients were not equal to 0. We see that the models are nested.



The model we end up is described by the equation below and this will be our *classifier.*

$$
\begin{aligned}
\text{subscribed} \quad & \sim Bernoulli\left(\text{prob}_{\text{subscribed=yes}} = \hat{P}\right) \\
\log\left[\frac{\hat{P}}{1-\hat{P}}\right] \quad & = 57.42 - 0.23\left(\text{age}_{(20,30]}\right) - 0.39\left(\text{age}_{(30,40]}\right) - 0.5\left(\text{age}_{(40,50]}\right) - \\
& \quad 0.34\left(\text{age}_{(50,60]}\right) - 0.19\left(\text{age}_{(60,70]}\right) + 0.07\left(\text{age}_{(70,100]}\right) - 0.06(\text{campaign}) - \\
& \quad 0.1\left(\text{pdays}_{\text{Contacted\_after\_10days}}\right) - 1.75\left(\text{pdays}_{\text{First\_time\_contacted}}\right) - \\
& \quad 0.54\left(\text{previous}_{\text{Less\_than\_or\_3\_times}}\right) - 0.21\left(\text{previous}_{\text{More than\_3\_times}}\right) - \\
& \quad 0.05(\text{emp\_var\_rate}) + 0.02(\text{cons\_conf\_idx}) - 0.01(\text{nr\_employed}) - 0.33\left(\text{job}_{\text{blue}-\text{collar}}\right) - \\
& \quad 0.13\left(\text{job}_{\text{entrepreneur}}\right) - 0.1\left(\text{job}_{\text{housemaid}}\right) - 0.09\left(\text{job}_{\text{management}}\right) + 0.11\left(\text{job}_{\text{retired}}\right) - \\
& \quad 0.03\left(\text{job}_{\text{self}-\text{employed}}\right) - 0.19\left(\text{job}_{\text{services}}\right) + 0.11\left(\text{job}_{\text{student}}\right) - 0.07\left(\text{job}_{\text{technician}}\right) - \\
& \quad 0.02\left(\text{job}_{\text{unemployed}}\right) - 0.36\left(\text{contact}_{\text{telephone}}\right) + 0.37\left(\text{quarter}_{\text{Q2}}\right) - 0.53\left(\text{quarter}_{\text{Q3}}\right) - \\
& \quad 0.69\left(\text{quarter}_{\text{Q4}}\right)
\end{aligned}
$$

Using *caret* package again, we setup the setting of the train function as such: Our training process will classify probabilities, use cross validation with k=5 folds and will compute measures specific to two-class problems, such as the area under the ROC curve (AUC), the sensitivity and specificity. Since the ROC curve is based on the predicted class probabilities (which are not computed automatically), another option is required. The *classProbs* set to TRUE option is used to include these calculations. This setup is used for all the algorithms in test.

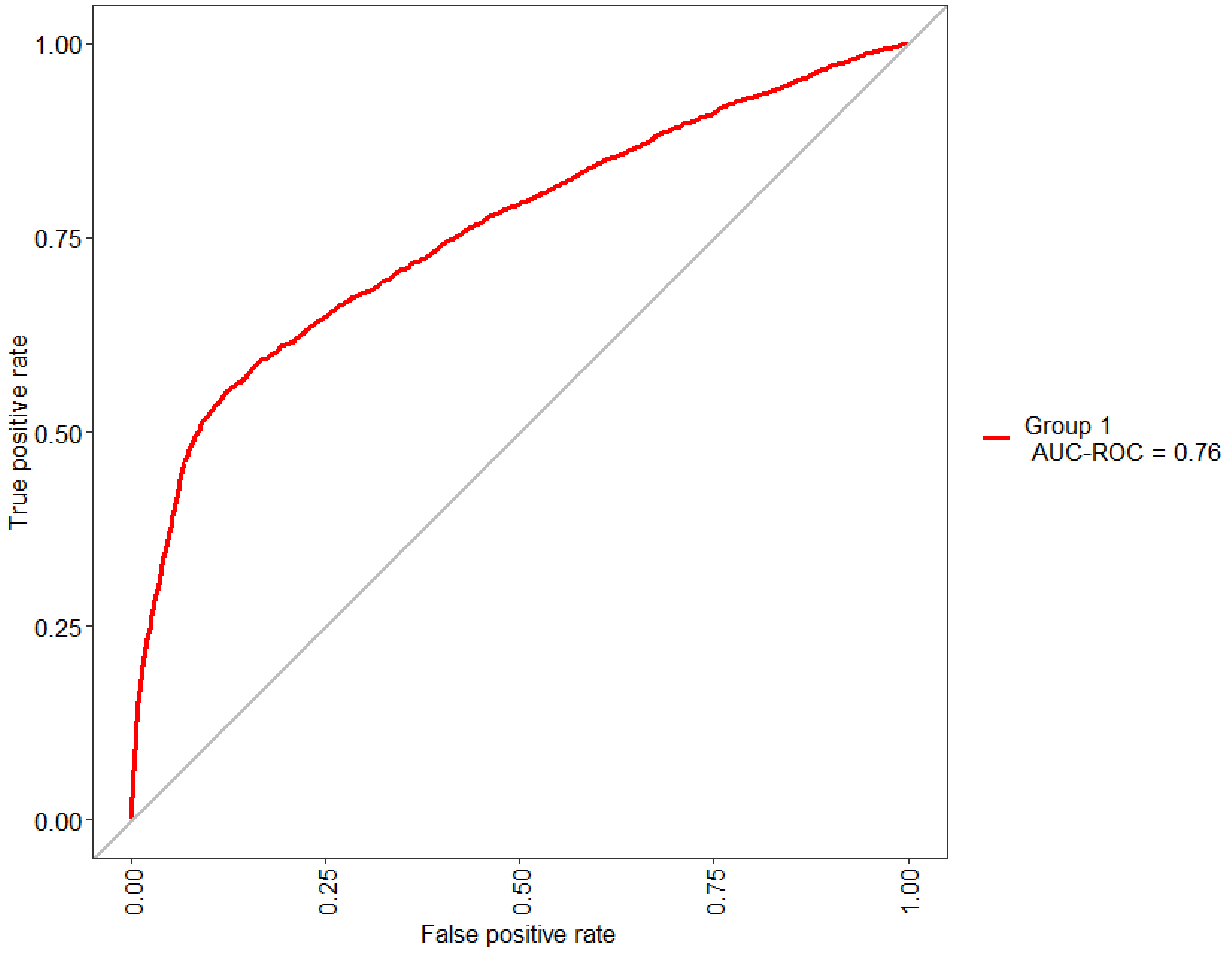
Let us now train the Logistic Regression model using cross-validation as defined above. The train dataset will also be used for validation. We will report an aggregated table at the end of this section with the metrics of all the methods. We set up a grid of tuning parameters for this classification routine, that fits each model and calculates a resampling-based performance measure. (Kuhn, 2008)

Confusion Matrix and Statistics

		Reference	
		no	yes
Prediction	no	10654	1017
	yes	111	178

Metric	Value
Sensitivity	0.14
Specificity	0.98
Pos Pred Value	0.61
Neg Pred Value	0.91

ROC curve



Classification with Naïve Bayes classifier

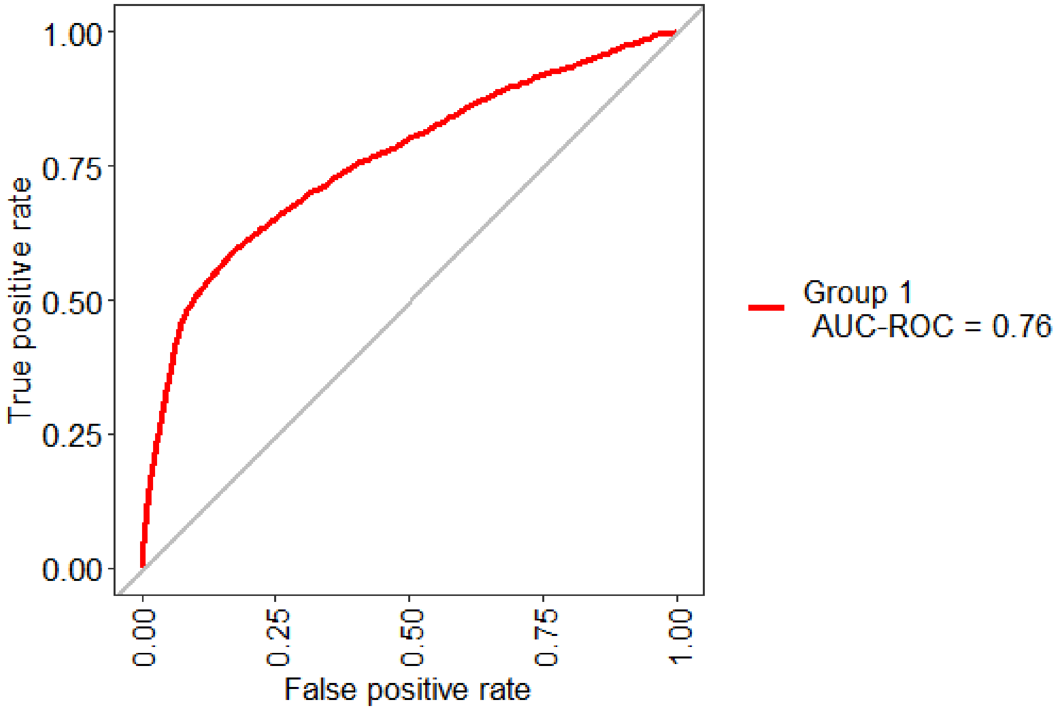
We now run the same process using the Naïve Bayes classifier with the formula given from the full model. We observe that the results in the confusion matrix are unacceptable. The classifier will always label the user as subscribed.

Confusion Matrix and Statistics

		Reference	
		no	yes
Prediction	no	10765	1195
	yes	0	0

Metric	Value
Sensitivity	0
Specificity	1
Pos Pred Value	Nan
Neg Pred Value	0.9

ROC curve



Classification with Decision Tree

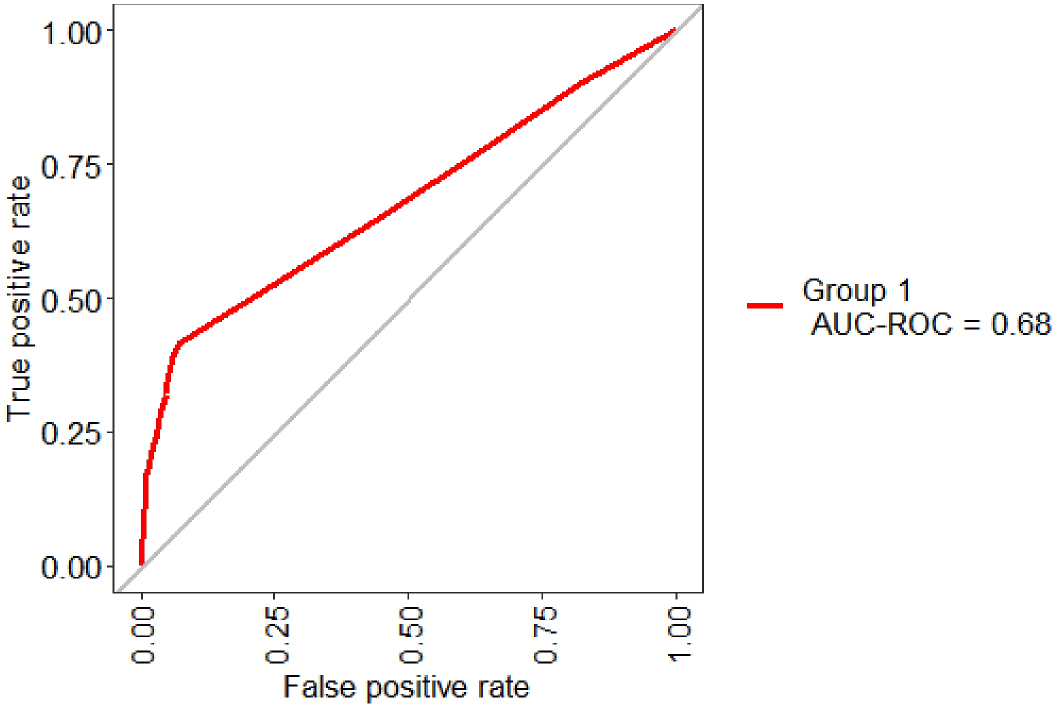
The third algorithm in test is a decision tree that, again, uses the formula of the full model. We let the algorithm decide the best features and how those should be split.

In this method, we notice that the features which decide on the subscription are roughly five, with the most important being nr_employed*, pdays* and *cons_conf_idx*. In general, the number of employees decides on the 91% of the cases. As seen in the tree Visualization, it splits exactly the dataset in 90/10 percent, the same as our target variable splits it.

Classification Matrix and Statistics

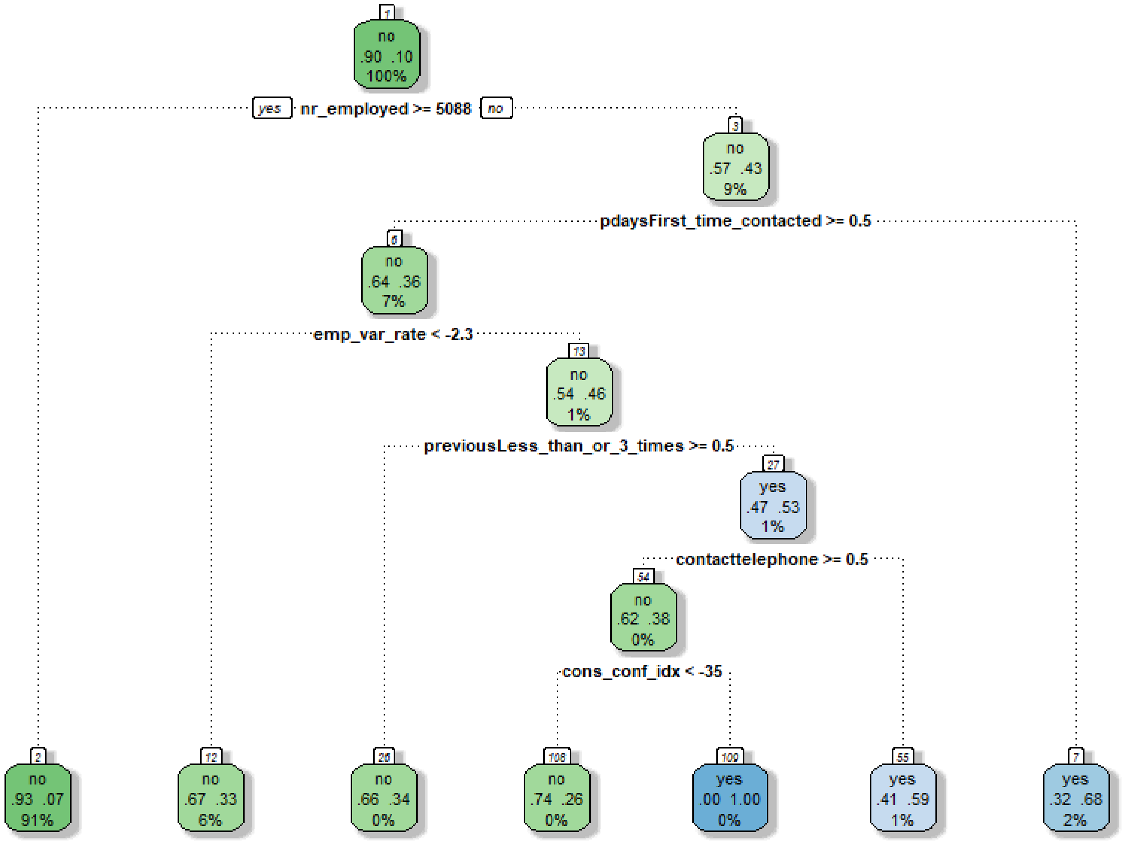
		Reference	
		no	yes
Prediction	no	10660	1010
	yes	105	185

ROC curve



Metric	Value
Sensitivity	0.15
Specificity	0.99
Pos Pred Value	0.63
Neg Pred Value	0.91

Selected Tree Visualization



Variable Importance – TOP 5 features

Variable	Overall
nr_employed	648.828
pdaysFirst_time_contacted	463.088
cons_conf_idx	439.107
emp_var_rate	406.523
cons_price_idx	376.237
previousLess_than_or_3_times	23.96

Summary Results

In conclusion, we have the summary results of the methods, and we are going to reject the Naïve Bayes model, because it will always classify to the negative. It would be better to randomly select a category, being known the subscription rate, rather than predicting with that model. Now, that the other two models are left we will select the algorithm with the best ROC. That will be the Logistic Regression model with 0.75 ROC versus the Decision Tree model which has 0.67. Accuracy on the cross-validation framework for those models is almost equal. The True Positive Rate (Sensitivity) is roughly bigger for the Logistic Regression model, than the Decision Tree model's one. True Negative Rate (Specificity), which is a metric we are interested in this framework, i.e., we want to be sure of the client is not going to subscribe, is also smaller for the Logistic Regression model.

	Logistic Regression	Naïve Bayes	Decision Tree
Accuracy	0.905	0.90	0.906
ROC	0.75	0.76	0.67
Specificity	0.164	0	0.168
Sensitivity	0.99	1	0.989

5. Clustering

Clustering allows us to better understand how our dataset might be comprised of distinct subgroups given a set of variables. In this section we are going to analyze a subset of the base variables. These are the client data and other socioeconomic factors. We will also try to characterize our clusters in terms of the *subscribed* variable. For this task, due to technical limitations, we will take a sample of our initial data (roughly 43% of the initial dataset). Since we have mixed data types, we need to choose an appropriate distance calculation method and clustering algorithm. In addition, we need to select the number of clusters and validate our choice.

First, we need to define a way to measure dissimilarity between the different observations. We know that Euclidean is only valid for continuous data. For that purpose, we will be using Gower's (1971) distance.

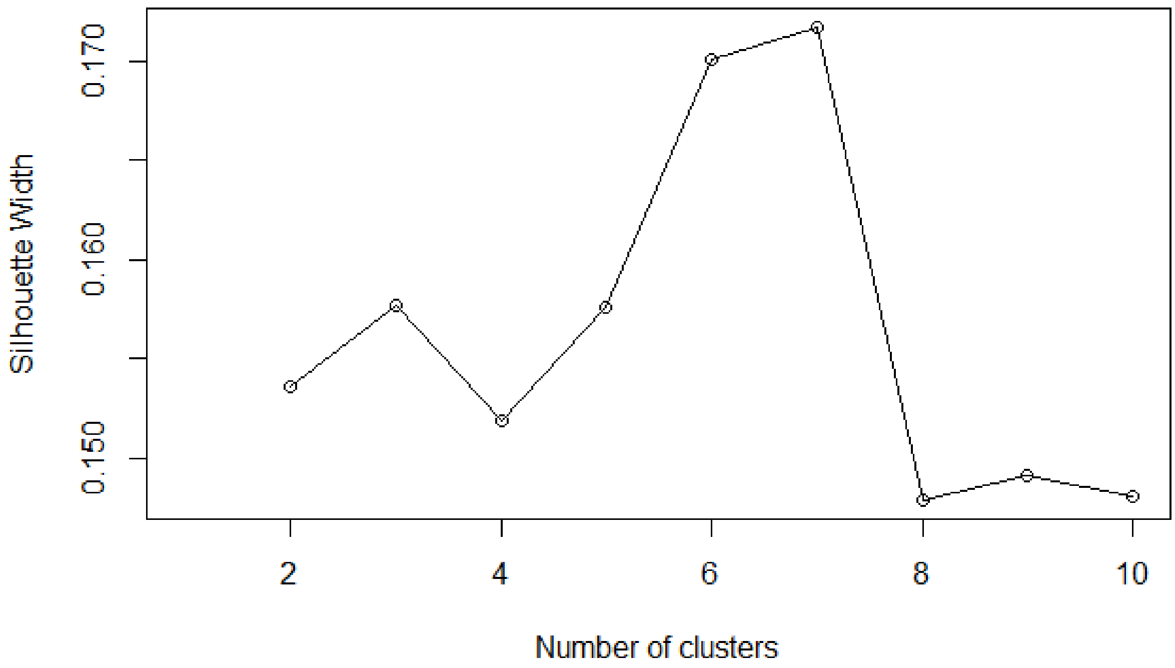
In short, Gower's distance (or similarity) first computes distances between pairs of variables over two data sets and then combines those distances to a single value per record-pair.

So, we now calculate a dissimilarity matrix which contains all the pairwise distances (dissimilarities) between the observations in the dataset.

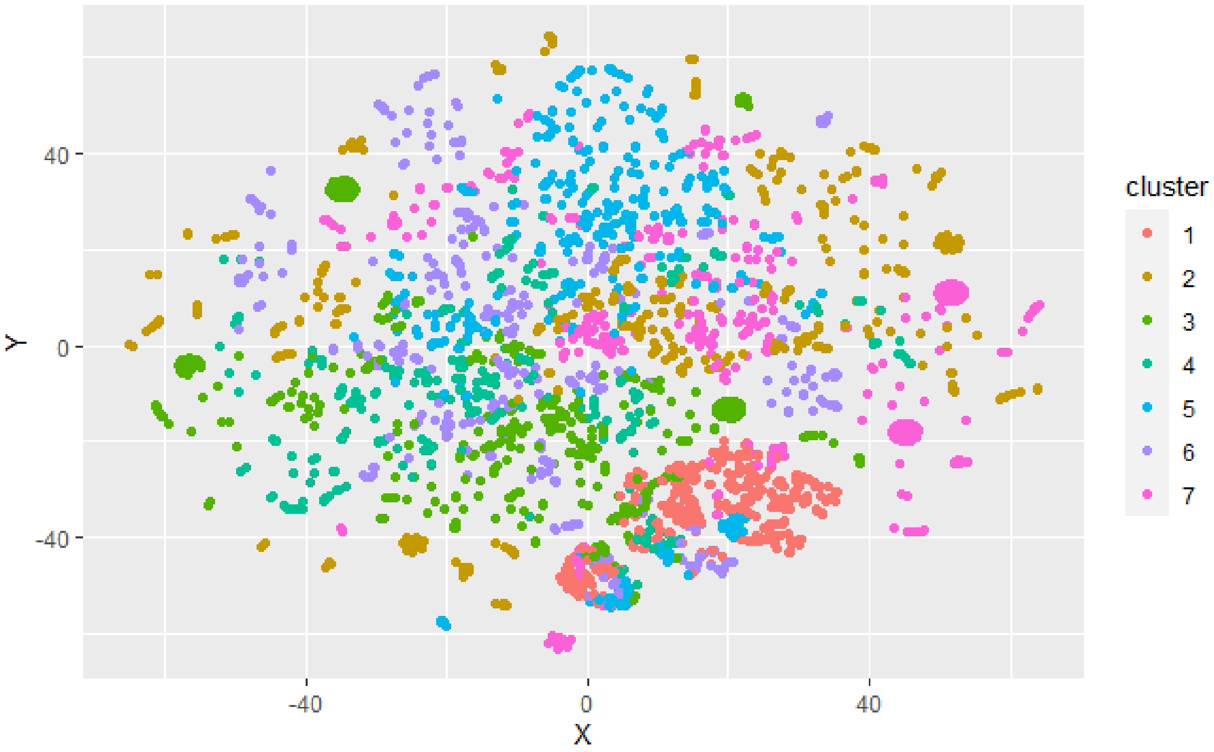
The next task is to choose a clustering algorithm. We will be using partitioning around medoids (PAM) as defined in Kaufman et al (1990). This is an iterative clustering algorithm that first chooses k random entities to become medoids, assigns every entity to its closest medoid. Then, for each cluster identifies the observation that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, it makes the observation the new medoid. If at least one medoid has changed, it returns to the assignment step, otherwise the algorithm ends. This is like k-means, but instead of cluster centers defined by the Euclidean distance, the centers are restricted to be the observations themselves.

Now that we have the distance and the algorithm, we need to select the number of clusters, like we would do in k-means. This task will be using an iterative process that will be calculating the silhouette width, n internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster, for clusters ranging from 2 to 10 with the PAM algorithm. We notice from the figure, that the best clusters number is that with mean width equal to 0.171. This would result in 7 clusters.

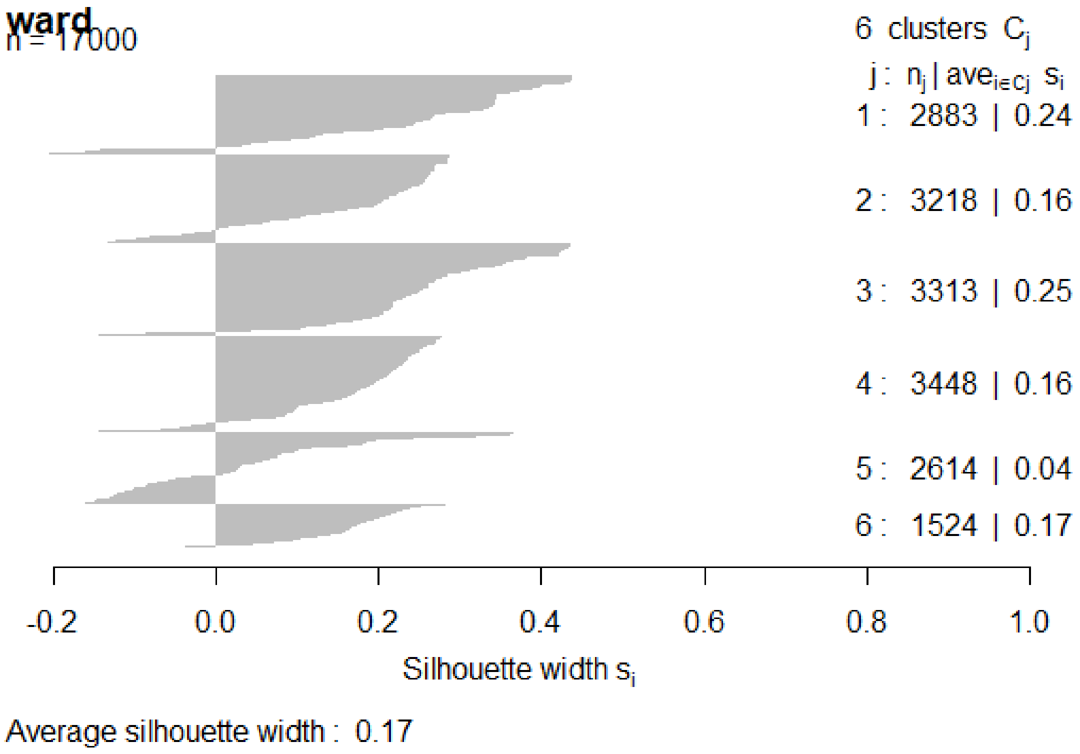
Silhouette widths for various runs of PAM



Now that we have selected the number of clusters, we are going to run the algorithm again for that number of clusters and summarize our result. We see that the average silhouette width is 1,72 and in the following graph we are showing the variables in a lower dimensional space with t-distributed stochastic neighborhood (t-SNE). This is a dimension reduction technique that attempts preservation of local structure to make the cluster visible in the 2 dimensions like this: We see the overlapping between the clusters.



We also provide silhouette graph with average measures of each clustering which show us the worst/best cluster of our result.

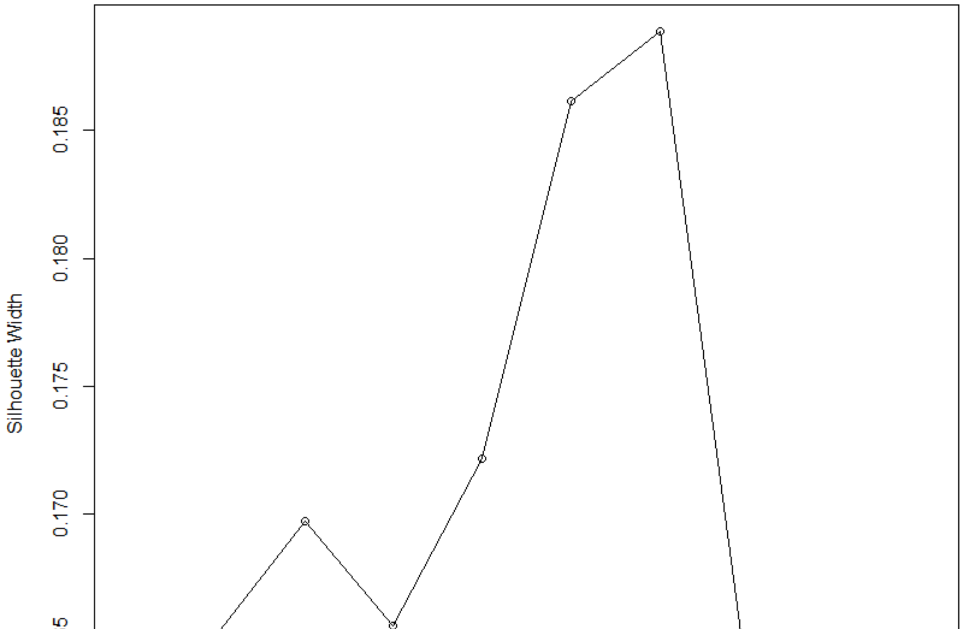


We will now have a look at the so-called medoids, the obserbations serving as centers in our analysis.

#	age	job	marital	education	default	housing	loan	campaign	previous	pdays	poutcome
16698	(30,40]	admin.	married	Tertiary_Education	no	yes	no	2	Less_than_or_3_times	First_time_contacted	failure
16734	(30,40]	admin.	married	Tertiary_Education	no	yes	no	2	Never contacted	First_time_contacted	nonexistent
16943	(40,50]	blue-collar	married	Primary_Education	no	no	no	1	Never contacted	First_time_contacted	nonexistent
16763	(30,40]	services	married	Secondary_Education	no	no	no	3	Never contacted	First_time_contacted	nonexistent
159	(20,30]	admin.	single	Secondary_Education	no	yes	no	2	Never contacted	First_time_contacted	nonexistent

#	age	job	marital	education	default	housing	loan	campaign	previous	pdays	poutcome
1753	(30,40]	technician	single	Tertiary_Education	no	no	no	2	Never contacted	First_time_contacted	nonexistent
16965	(40,50]	blue-collar	married	Primary_Education	no	yes	no	1	Never contacted	First_time_contacted	nonexistent

From the table above we see that 3 variables are constant between the centers.



We are removing these 3 variables and we re-run the process with fewer variables. This is only to find out that more variables should be removed previous and poutcome affect the result and doesn’t allow for fewer cluster to be resulted. Although, silhouette has been improved slightly as show in the figure above.

References

Athanasopoulos, H. a. (2013). *Forecasting: principles and practice.* Retrieved from https://otexts.com/fpp2/

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Kaufman, L. and Rousseeuw, P.J. (1990). Partitioning Around Medoids (Program PAM). https://doi.org/10.1002/9780470316801.ch2

Karlis, Course lectures