Athens University of Economics and Business
Department of Management Science and Technology

Statistics for BA II
Professor: D. Karlis

Project II
Telemarketing Dataset

Marianna Konstantopoulou
A.M: P2822122

M.Sc. Business Analytics
Part Time 2021-2023

Athens, 06/04/2022

Table of Contents

Chapter 1: Introduction - description of the problem, data, aim, background information…….……. iii

Chapter 2: Classification methods………………………...……………………………………..…… iv

Chapter 3: Clustering methods.……...……………………….……………………………….…..…...x

Chapter 4: Conclusion……….……………………………..……………………………...…...xiii

Chapter 1

Introduction
- description of the problem, data, aim, background information

This information is about telemarketing phone calls to sell long-term deposits. During a campaign, the agents call a list of customers to promote the product (outbound) or, if the client phones the contact-center for any other purpose, he is urged to subscribe to the product (inbound). As a consequence, the outcome is a binary variable indicating whether the interaction was successful or failed.

This study takes into account real data gathered from a retail bank from May 2008 to June 2010, in a total of about 40K phone interactions. More than one contact with the same consumer was frequently necessary to determine if the product (bank term deposit) will be subscribed to ('yes') or not ('no').

The first aim of our project was to create a predictive model to classify whether a client will buy or not the new product. Our dataset contains variables about the bank client data such as his age, education, marital and job status, his housing and personal loans, as well as details about the type and frequency of contact (about current or previous campaigns). Important indexes and rates are also included (employment variation rate, employment variation rate, employment variation rate, employment variation rate and the number or employees).

The second goal of this project was to use some of the variables mentioned previously to cluster the clients and to characterize the clusters.

Chapter 2

Classification methods

To start our analysis, we will load our data in R and check the data types. We will remove a column that is not useful for our research: pdays (we can extract the same information from the "previous" variable so there is no need to keep both). We need to change the rest of the variables' data types to the correct type. Age of the client (age), duration of last contact (duration), the number of contacts performed during this campaign (campaign), the number of contacts performed before this campaign (previous) and the indexes/rates employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m) as well as the number of employees (nr.employed) are all numeric. We also have a number of categorical variables: job, marital status, education, default credit, housing and personal loan, contact communication type, last contact month and day of the year, day of the week, outcome of previous marketing campaign and of course our response variable whether the client subscribed a term deposit or not. It is also important to perform a check for potential missing values. In our case there are no missing values in our data set.

In order to start our classification, we will need to split our data set to train, test and validation data set. I followed a split of 60% train, 25% test and 15% validation. The first method I used was the **Logistic Regression**. The logistic regression model is a type of Generalized linear model (GLM), which is an extension of linear regression. The GLMs are identified by the combination of the assumed distribution and the link function *g*. It is worth noting that the link function may be viewed as a bridge between the linear predictor and the mean of the distribution function, as shown below:

$$E[Y|X] = \mathbf{X}\beta = \mu = g^{-1}(\mathbf{X}\beta)$$

The logistic regression assumes the dependent variable follows Bernoulli distribution with logit link *g*, which can be written as follows:

$$g(x) = \log(\frac{x}{1-x})$$

This means that $g^{-1}(x)$ will be:

$$g^{-1}(x) = e^x / (1 + e^x)$$

Since our Y response variable can only take 2 possible values (0: unsuccessful and 1: successful contact). This means that:

$$Y_i \sim B(1, p_i) \text{ independent for } i=1,\ldots.39883,$$

$$p_i = P(Y_i = 1)\text{: successful contact}$$

In order to implement Logistic Regression, we will need the following transformation as link function, which is the logit function:

$$\log \frac{pi}{1-pi} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},\ i = 1, \ldots., p$$

We will conduct LASSO as a variable selection technique. We use cross validation and select the largest value of lambda such that error is within 1 standard error of the minimum Using the lambda we selected, we receive the estimated coefficients under the lambda.1se. Our model has subscribed as a response and as predictors all variables and intercept except for age, loan, previous and euribor 3 month index.

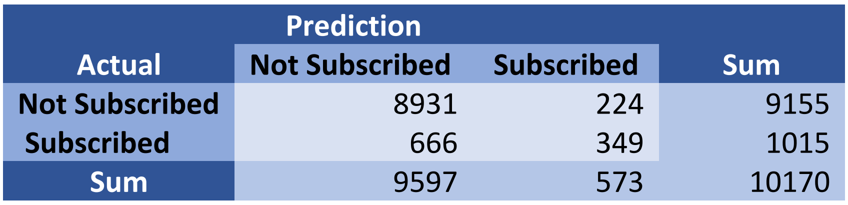
To select our final model, we are using stepwise methods as well. Since we are focusing on prediction, we used the Stepwise procedure according to AIC. Our final model has subscribed as response and the predictors are job, marital, default, contact, month, duration, campaign, poutcome, emp.var.rate, cons.price.idx and nr.employed.

Next step would be to check for multi-collinearity. We used the GVIF test and decided to remove variable emp. var. rate since it caused multicollinearity issues to our model.

This is the final model we selected:

logit (*subscribed*) = 75.3 – 0.371* *jobblue-collar* – 0.08 * *jobentrepreneur* – 0.37 * *jobhousemaid* - 0.04 * *jobmanagement* + 0.28 * *jobretired* + 0.49 * *jobself-employed* - 0.22 * *jobservices* + 0.40 * *jobstudent* – 0.78 * *jobtechnician* + 0.06 * *jobunemployed* + 0.01 * *jobunknown* – 0.13 * *marritalmarried* + 0.008 * *maritalsingle* + 0.17 * *maritalunknown* - 0.29 * *defaultunknown* – 8.4 * *defaultyes* – 0.25 * *contacttelephone* + 0.66 * *monthaug* + 0.39 * *monthdec* + 0.47 * *monthjul* + 0.67 * *monthjun* + 1.27 * *monthmar* – 0.63 * *monthmay* + 0.05 * *monthnov* + 0.45 * *monthoct* – 0.2 * *monthsep* + 0.004 * *duration* – 0.05 * *campaign* + 0.4 * *poutcomenonexistent* + 1.66 * *poutcomesuccess* + 0.007 * *cons.price.idx* – 0.01 * *nr.employed*

We will now use the ROC curve to select the best threshold. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. When using a threshold of p = 0.5, using our test data we receive the following confusion matrix:



	Prediction		
Actual	Not Subscribed	Subscribed	Sum
Not Subscribed	8931	224	9155
Subscribed	666	349	1015
Sum	9597	573	10170

Table 1: Logistic regression confusion matrix for threshold p = 0.5

This indicates that the precision is 0.60 and the recall (sensitivity) is 0.34, while the F1-score is 0.44. We will prefer to compare our models with F1-score since we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives).

One of the best methods to select the best p is to find the coordinates of FP and TP in the ROC curve, which minimize the distance from the point to (0, 1). In other words, find the p, which makes the decision rule to be a perfect classifier. In Figure 1, the red dot represents the optimal point with threshold 0.087.

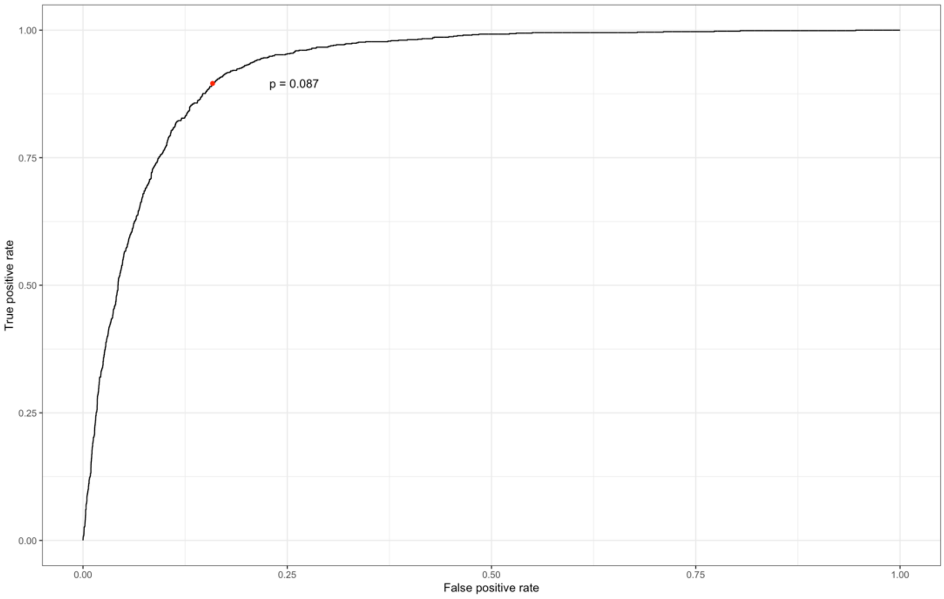


Figure 1: Logistic regression ROC curve and indication of the optimal threshold p

The confusion matrix for the new threshold p = 0.087 is the following:

	Prediction		
Actual	Not Subscribed	Subscribed	Sum
Not Subscribed	7695	1460	9155
Subscribed	106	909	1015
Sum	7801	2369	10170

Table 2: Logistic regression confusion matrix for threshold p = 0.087

This means that the precision is 0.38 and the recall (sensitivity) is 0.89, while the F1-score is now higher 0.53.

Next method we will use will be **Naive Bayes**. Naive Bayes is a probabilistic classifier inspired by the Bayes theorem under a simple assumption which is the attributes are conditionally independent. There is an unobserved class variable C that we want to predict and several features X_1, X_2, …., X_p that depend on the class variable. New observations are assigned to a class according to Bayes Theorem

$$P\,(Ci = k \mid yi) = (\pi_\kappa\, f\,(yi \mid Ci = k))/\sum_{j=1}^{K} f\,(yi \mid Ci = j)$$

where π_κ denotes the prior probability of class k.

The confusion matrix created after using the Naive Bayes method is the following:

	Prediction		
Actual	Not Subscribed	Subscribed	Sum
Not Subscribed	7785	313	8098
Subscribed	1370	702	2072
Sum	9155	1015	10170

Table 3: Naive Bayes confusion matrix

This leads us to precision of 0.69 and recall (sensitivity) of 0.33. The F1-score of this method is 0.45, which makes the Logistic Regression with the 0.087 threshold a better classifier so far.

Lastly, the third method we will use is **Random Forest**. In Random Forest we build several decision trees to create a forest, take one prediction from each tree and then combine them to create a generic prediction. We decided to use 200 number of trees, and this resulted in the following confusion matrix:

	Prediction		
Actual	Not Subscribed	Subscribed	Sum
Not Subscribed	8850	548	9398
Subscribed	305	467	772
Sum	9155	1015	10170

Table 4: Random Forest with 200 trees confusion matrix

This basically means that the precision is 0.46 and the recall (sensitivity) is 0.60, while the F1-score is 0.52.

If M is the number of candidate variables, for each tree we select m <<M so an in every node m variables are selected randomly in order to make decisions. To find the optimal m we run our random forest technique multiple times with different m (starting from 2 and with a maximum of 20) and we will select the one that gives us the best f1-score. Our final decision is that the number of variables randomly sampled as candidates at each split will be 17. When running the random forest with 17 variables we received the following confusion matrix:

	Prediction		
Actual	Not Subscribed	Subscribed	Sum
Not Subscribed	8803	515	9318
Subscribed	351	500	851
Sum	9154	1015	10170

Table 5: Random Forest with 200 trees and 17 variables confusion matrix

Finally, this means that the precision of this method is 0.49, the recall is 0.58 and the F1-score is 0.54.

If we compare all 3 methods using the F1-score as our measure of how good our classification was using the test data, we will choose Random Forest as the best method for this classification problem. We will use the validation data set we created at the beginning to evaluate our method. This is the confusion matrix that was created using the validation data set:

	Prediction		
Actual	Not Subscribed	Subscribed	Sum
Not Subscribed	5173	283	5456
Subscribed	216	310	526
Sum	5389	593	5982

Table 6: Random Forest with 200 trees and 17 variables confusion matrix for validation data set

Lastly, our final precision is 0.52, recall (sensitivity) is 0.58 and final F1-score is 0.55. At this point we can also mention the accuracy of the model which is 91.6%, however since our data set has uneven class distribution, accuracy can become an unreliable measure of model performance. This means that intuitions for classification accuracy based on balanced class distributions will be applied and will be incorrect, causing practitioners to believe that a model has high or even great performance when, in reality, it does not.

Chapter 3

Clustering methods

Clustering is the process of grouping a population or set of data points into groups so that data points in the same group are more similar to data points in other groups than data points in other groups. Our main aim is to segregate groups with similar traits and assign them into clusters. Our dataset now consists of the following attributes: age of the client (age), the number of contacts performed during this campaign (campaign), the number of contacts performed before this campaign (previous) as well as the categorical variables: job, marital status, education, default credit, housing and personal loan, outcome of previous marketing campaign. We will get rid of the variable pdays as we can extract the same information from the "previous" variable so there is no need to keep both.

Firstly, we will create a distance matrix using Gower distance, since we have categorical variables too. Gower distance measures for each variable its distance in a range between (0,1) and takes the average. Since our data set has 39,883 observations it's nearly impossible for our machine to process and calculate the distances for this large number of observations so we will take a random sample of 10,000 observations, so everything runs smoothly. After using the sample data set to create the distance matrix we will perform Hierarchical clustering. For linkage method we will choose Ward's linkage. Ward's minimum variance criterion reduces the overall within-cluster variation to a minimum. At each phase, the clusters with the shortest between-cluster distance are merged.

After the hierarchical clustering is performed, we can depict the results with Cluster Dendrogram:

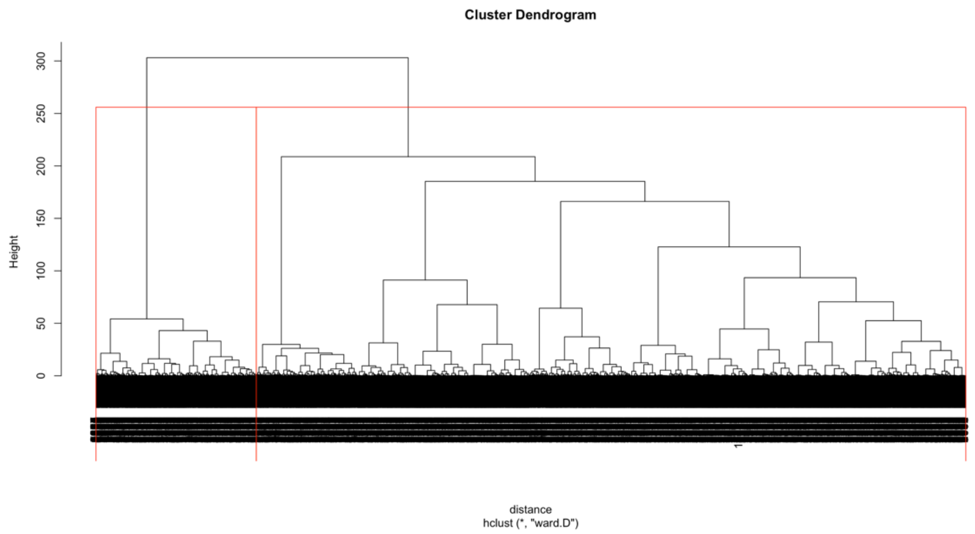


Figure 2: Cluster Dendrogram (rectangles around the branches of a dendrogram highlighting the 2 clusters) with Ward linkage

As it is visible from the above dendrogram we selected 2 clusters. Even though the specific dendrogram is not the best way to distinguish the number of clusters (it could easily be 6 or even 7) we will choose the 2 because we can observe some differences regarding the variables between the clusters, will be explained shortly, and we can easily compare our results with the subscribe variable.

We will also quickly check the Silhouette value, which is a method of interpretation and validation of clusters of data:

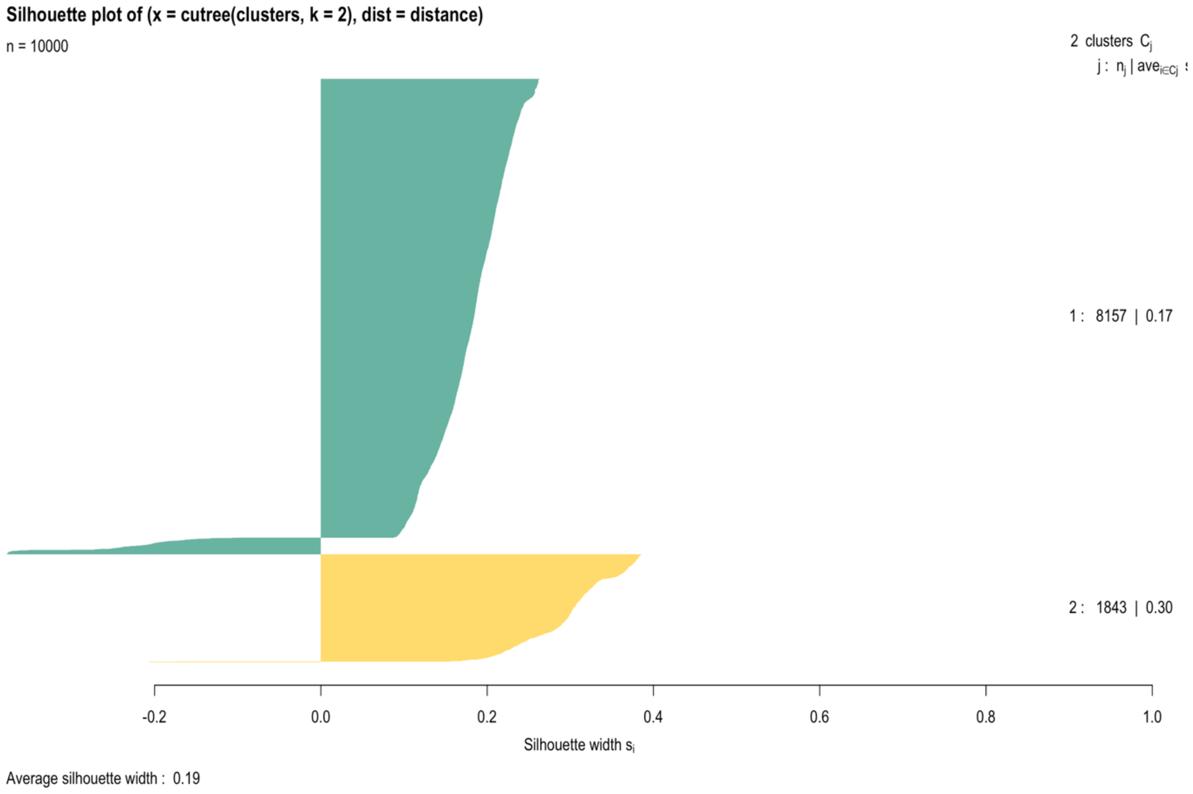


Figure 3: Silhouette plot for 2 clusters (hierarchical clustering)

We can see that there are 8,157 observations in the first cluster and 1,843 observations in the second cluster. The average silhouette width is a measure of how appropriately the data has been clustered and in our case it's 0.19. Knowing that the closer the silhouette width is to one then the more appropriately the datum is clustered, we can see that our clustering is not the best (there are observations with negative silhouette values, which means that these observations would be more appropriate if they were clustered in their neighboring cluster) but it's not as bad either as we can spot differences between our two clusters.

While doing a summary of observations in each cluster we can spot two significant differences: different types of education and jobs in each cluster.

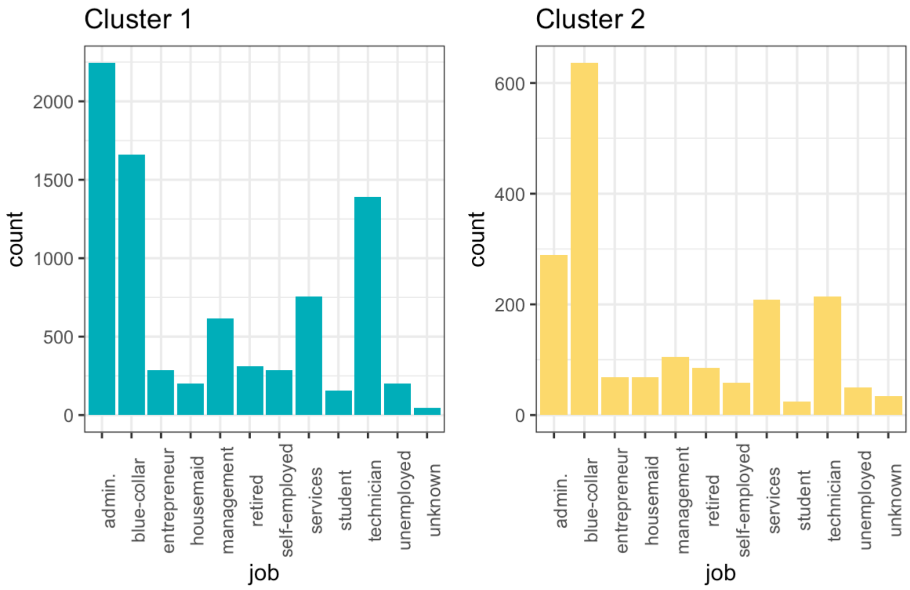


Figure 4: Bar plots for job variable for the 2 clusters

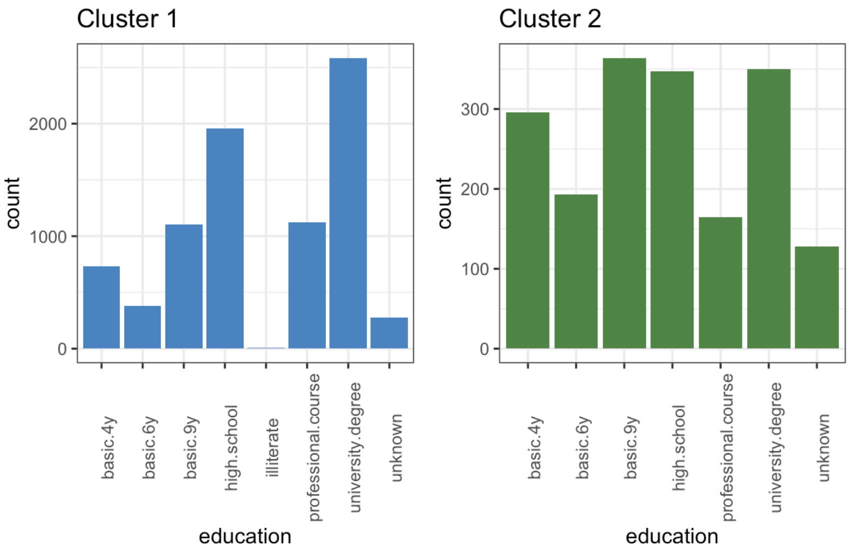


Figure 5: Bar plots for education variable for the 2 clusters

- The most frequent occupation in cluster 1 is administration while on cluster 2 is blue-collar, which basically is workers who engage in hard manual labor
- Most clients on cluster 1 have a university degree while on cluster 2 most clients have completed the 9-year universal basic education program

Lastly, we can compare our clustering with the ground-truth partition using the Adjusted Rand Index. We obtained a negative ARI (-0.045), indicating that the agreement is lower than what would be anticipated from a random outcome. This means that our final clustering doesn’t relate to the subscribe variable.

Chapter 4

Conclusion

In conclusion, finishing the first aim of our project that was to create a predictive model to classify whether a client will buy or not the new product, we used three different classification methods (Logistic Regression, Naive Bayes and Random Forest). Since our data set was pretty unbalanced, we used the F1-score to compare the performance of our methods and we finally selected Random Forest. Using our validation data set this method gave us 91.6% accuracy (which is not the best metric in our case) and F1-score 55.7%.

Regarding the clustering, which was our second aim, we were able to cluster the clients in 2 clusters and characterize the clusters. We used the Gower distance to calculate the distance matrix (since our data set included categorical variables too) and then we performed Hierarchical clustering with Ward Linkage. Our average silhouette width was 0.19 and the Adjusted Rand Index compared to the ground truth was below zero which indicates that our clusters don’t agree with the subscribe variable. We were also able to spot differences between the two clusters regarding the occupation and the education of the clients.