Athens University of Economics and Business

MSc in Business Analytics

Statistics For Business Analytics II

Project II 2021-2022

Bank telemarketing phone calls dataset

Name: *Panagiota Gkourioti*

Student ID: *P2822109*

Professor: *Karlis Dimitrios*

Contents

1. Introduction – description of the problem

Telemarketing is a widely adopted direct advertising approach towards potential customers through telephone communication. Customer targeting consists of identifying prospective buyers of a product or subscribers of a service offered within the context of a marketing campaign. It is beneficial for businesses to constrict the range of potential customers and to explore their characteristics, thus increasing the success rate as well as efficiently reducing the marketing costs.

This marketing technique also enables banks and other financial organizations to focus on those customers who present the largest likelihood of subscribing to their products, offers, and other packages. Most often than not, identifying these group of customers poses a challenge to financial institutions.

The current study focuses on a retail bank that conducted a telemarketing campaign, aiming to persuade customers into subscribing for long-term deposits. Within the campaign, the agents make phone calls to a list of clients to sell the product (outbound) or, if meanwhile the client calls the contact-center for any other reason, he is asked to subscribe the product (inbound). The dataset contains information collected from a retail bank between May 2008 to June 2010 for all the customers who were contacted during a particular year to open term deposit accounts in a total of near 40 thousand phone contacts. They include both numerical and categorical variables, regarding the socio-demographics of the clients, the client-bank relationship, the contact context and the socio-economic macro context.

This project uses Classification and Clustering Machine Learning techniques to make predictions about the outcome of the campaign.

The first part will focus on the supervised learning problem, which uses already labeled data to find specific relationships or structure and produces correct output data. Specifically, we will use classification algorithms to construct several models, which will try to efficiently predict the outcome of the bank telemarketing campaign having as input both numerical and categorical variables, regarding the socio-demographics of the clients, the client-bank relationship, the contact context and the socio-economic macro context.

The second part deals with the unsupervised learning problem, where no labels are provided, therefore a cluster analysis is performed that groups data into clusters, by attempting to find similarities and dissimilarities among them based on their distance.

2. Model building

Part 1: Classification

This section will focus on constructing several models that attempt to predict whether the client will subscribe to a Term Deposit. Data Classification is the use of Machine Learning techniques to organize datasets into related sub-populations, not previously specified in the dataset. This can uncover hidden characteristics within data and identify hidden categories that new data belongs within.

The method used for splitting the dataset will be K- Fold Cross-validation. This is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. We decide to use 6 folds (k=6) for our analysis to achieve comprehensive results of the prediction tested on different parts of the dataset. This method generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

Logistic Regression

The first method used for model building is logistic regression, which is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary), specifically in the current study, whether or not the client subscribed a long-term deposit (success/failure), and the predicted values are probabilities.

The large number of potential useful features for a model demands a stricter choice of relevant attributes. Feature selection is a useful step to discard irrelevant inputs, leading to a simpler data-driven model that are easier to interpret and provides better predictive performances. To achieve that, we implement variable selection algorithms (LASSO - AIC).

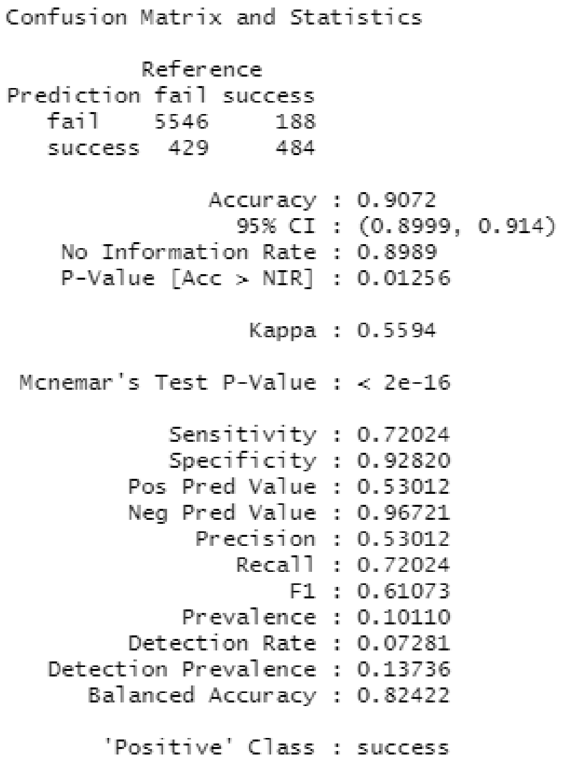
Initially, LASSO method filters the full model and selects the most appropriate covariates. It forces the sum of the absolute value of the coefficients to be less than a fixed value, which sets certain coefficients to zero, thereby removing them from the model. In logistic regression, Lasso regularization works by adding a penalty term to the log likelihood function. The regularization parameter is “lambda” and measures the degree to which the coefficients are penalized. By choosing lambda within one standard error from the minimum, LASSO shrinks more parameters towards zero and the model becomes significantly simpler, while also dealing with the multicollinearity issue.

Afterwards, we implement AIC variable selection. By comparing the AIC improvements from dropping and adding each candidate variable from the current model, we end up with a model having the smallest AIC. The model ends up with ten variables: age, job, education, day_of_week, month, duration, campaign, poutcome, emp.var.rate and nr.employed.

Furthermore, multicollinearity between the covariates should be checked using the Vif test. This shows that emp.var.rate and nr.employed are correlated, as previously observed in the correlation table of quantitative variables. Therefore, we decide to remove the emp.var.rate variable to fix the multicollinearity issue and fit again the general linear model.

The next step, after concluding to the final model, is to train our model into the data, using 6-fold cross validation and then evaluate its ability to predict on the unseen data.

We use the confusion matrix in order to evaluate the predicting ability of the model. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. For the logistic regression, it is observed on table 2.1. that the model accuracy on the test dataset is 90.72 % and the balanced accuracy (that takes into account the sensitivity) is 82.42%, which are very high. Moreover, the accuracy of the model is estimated between 89.99% and 91.4% in a confidence interval of 95%.



```
Confusion Matrix and Statistics

          Reference
Prediction fail success
   fail    5546     188
   success  429     484
                                          
               Accuracy : 0.9072          
                 95% CI : (0.8999, 0.914) 
    No Information Rate : 0.8989          
    P-Value [Acc > NIR] : 0.01256         
                                          
                  Kappa : 0.5594          
                                          
 Mcnemar's Test P-Value : < 2e-16         
                                          
            Sensitivity : 0.72024         
            Specificity : 0.92820         
         Pos Pred Value : 0.53012         
         Neg Pred Value : 0.96721         
              Precision : 0.53012         
                 Recall : 0.72024         
                     F1 : 0.61073         
             Prevalence : 0.10110         
         Detection Rate : 0.07281         
   Detection Prevalence : 0.13736         
      Balanced Accuracy : 0.82422         
                                          
       'Positive' Class : success         
```

Table 2.1.: Confusion matrix for logistic regression

Lastly, we plot the ROC curve of the model, which represents the true positive rate against the false positive rate at different classification thresholds. We would like the ROC curve to be stretched as far as possibly to the upper left corner. The higher to the left is the curve, the lower is the percentage of false classified observations. Figure 2.1. displays the six ROC curves of the logistic regression and for each of the 6 folds of the dataset.

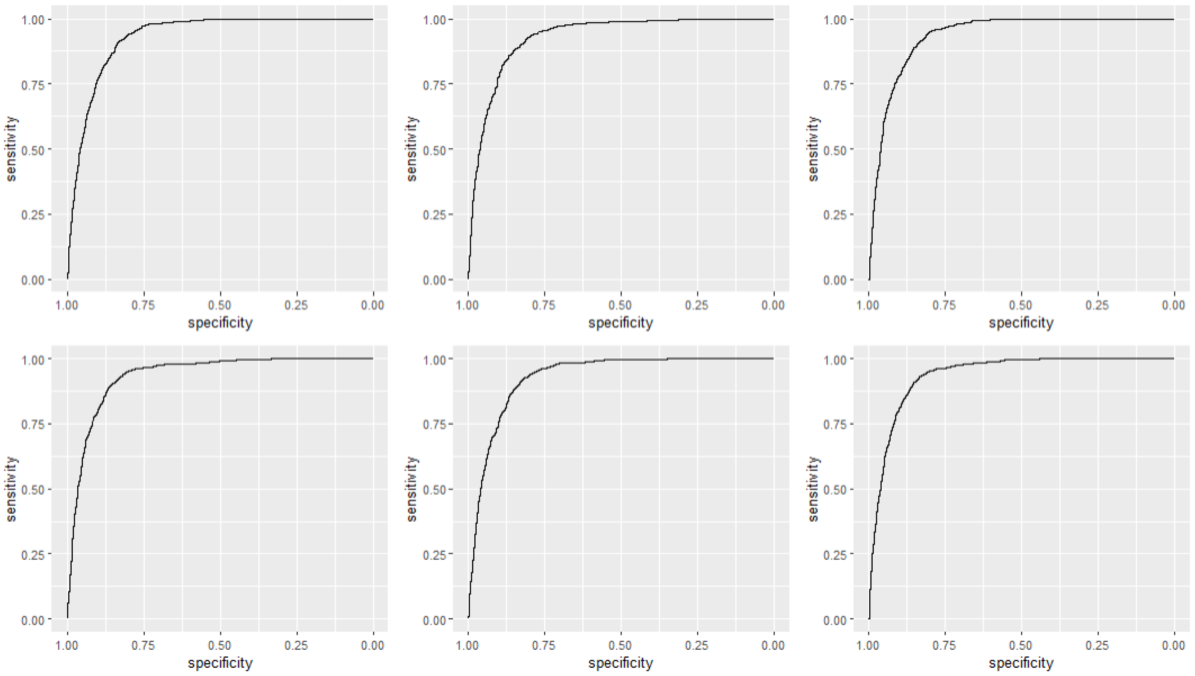


Figure 2.1.: ROC curves of logistic regression for each of the 6 folds

Decision Tree

The second method used for predicting the outcome of the campaign is decision tree, which is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision tree we constructed is presented in Figure 2.2, along with the thresholds and percentages of observations on each node.

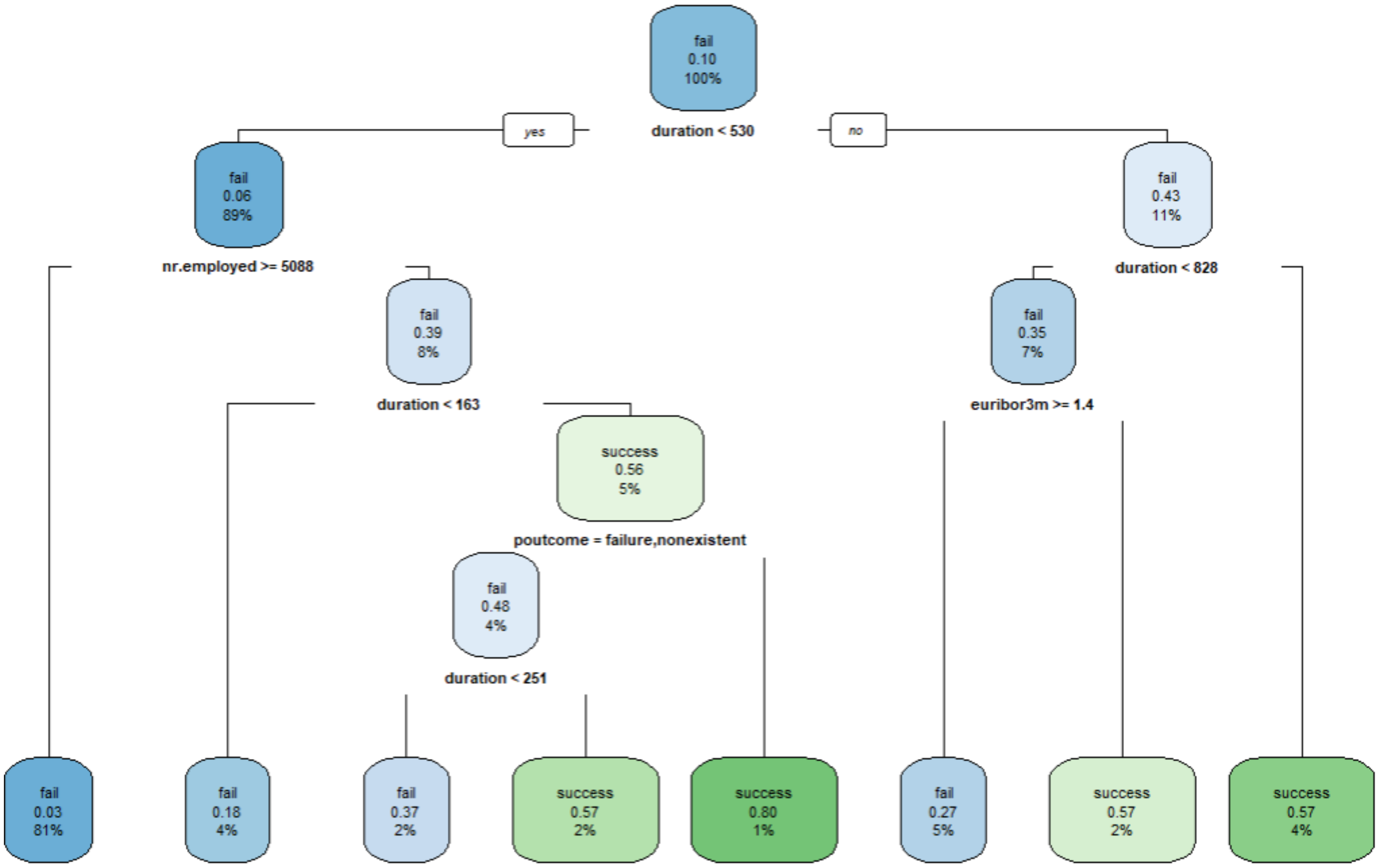
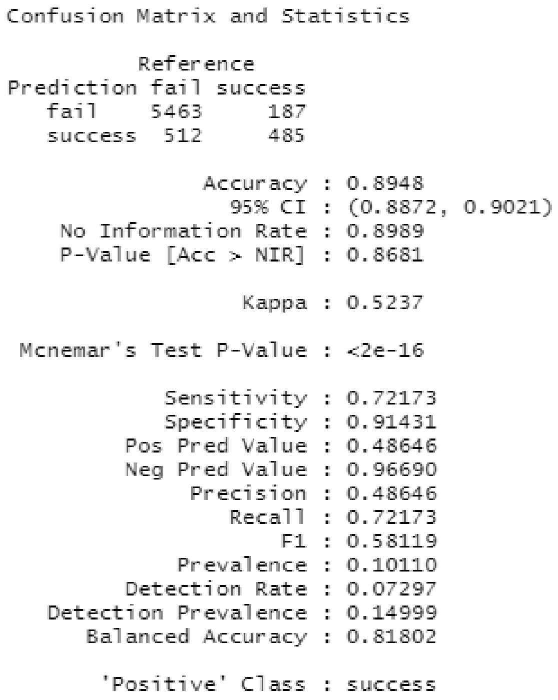


Figure 2.2.: Decision tree

Moreover, the results of the confusion matrix are presented in Table 2.2. For the decision tree, it is observed that the model accuracy on the test dataset is 89.48 % and the balanced accuracy (that takes into account the sensitivity) is 81.8%, which are very high. Moreover, the accuracy of the model is estimated between 88.72% and 90.21% in a confidence interval of 95%.



```
Confusion Matrix and Statistics

          Reference
Prediction fail success
   fail    5463     187
   success  512     485

               Accuracy : 0.8948
                 95% CI : (0.8872, 0.9021)
    No Information Rate : 0.8989
    P-Value [Acc > NIR] : 0.8681

                  Kappa : 0.5237

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.72173
            Specificity : 0.91431
         Pos Pred Value : 0.48646
         Neg Pred Value : 0.96690
              Precision : 0.48646
                 Recall : 0.72173
                     F1 : 0.58119
             Prevalence : 0.10110
         Detection Rate : 0.07297
   Detection Prevalence : 0.14999
      Balanced Accuracy : 0.81802

       'Positive' Class : success
```

Table 2.2.: Confusion matrix for decision tree

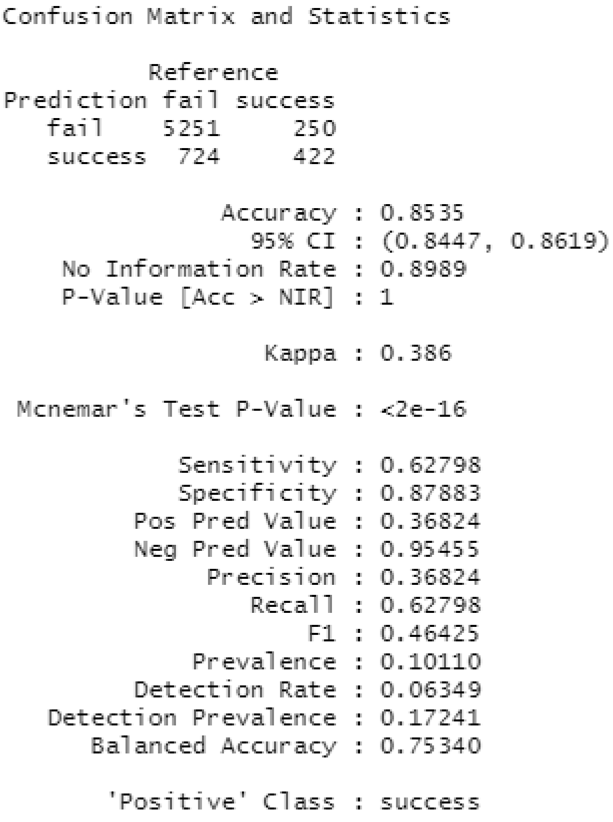
Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. The thought behind naive Bayes classification is to try to classify the data by maximizing P(O|Ci)P(Ci) using Bayes theorem of posterior probability (where O is the Object or tuple in a dataset and "i" is an index of the class).

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

Since Naive Bayes assumes independency between features, we remove the variable euribor3m from the model because it has very high correlation with the other numerical variables, as found in the correlation matrix (See Appendix: Table 2.3.).

The results of the confusion matrix are presented in Table 2.3. For the Naive Bayes, it is observed that the model accuracy on the test dataset is 85.35 % and the balanced accuracy (that takes into account the sensitivity) is 75.34%, which are very high. Moreover, the accuracy of the model is estimated between 84.47% and 86.19% in a confidence interval of 95%.



```
Confusion Matrix and Statistics

          Reference
Prediction fail success
   fail    5251     250
   success  724     422

               Accuracy : 0.8535
                 95% CI : (0.8447, 0.8619)
    No Information Rate : 0.8989
    P-Value [Acc > NIR] : 1

                  Kappa : 0.386

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.62798
            Specificity : 0.87883
         Pos Pred Value : 0.36824
         Neg Pred Value : 0.95455
              Precision : 0.36824
                 Recall : 0.62798
                     F1 : 0.46425
             Prevalence : 0.10110
         Detection Rate : 0.06349
   Detection Prevalence : 0.17241
      Balanced Accuracy : 0.75340

       'Positive' Class : success
```

Table 2.4.: Confusion matrix for Naive Bayes

Comparison of methods

Finally, we produce metrics for each classification method in order to compare their predictive abilities and create boxplots representing their accuracy. Accuracy and Adjusted Rand Index (ARI) are used as metrics and we would like them to be as high as possible. From the Figure 2.3. and the tables below, it can be concluded that Logistic Regression and Decision tree methods perform similarly while the decision Naive Bayes method has poorer results than the previous two. Out of all, logistic regression produces the best results.

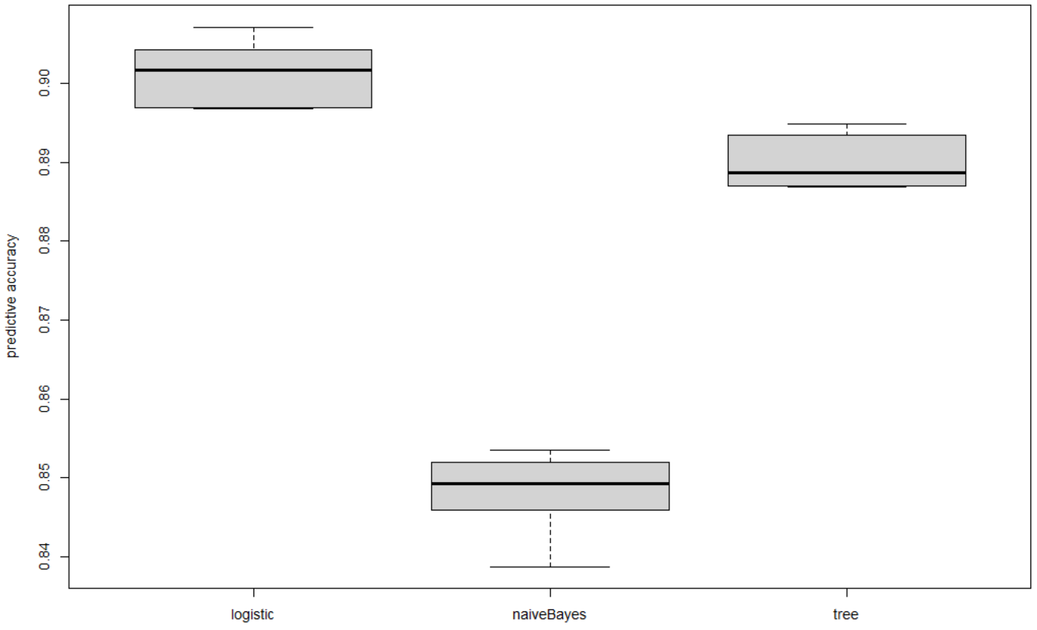
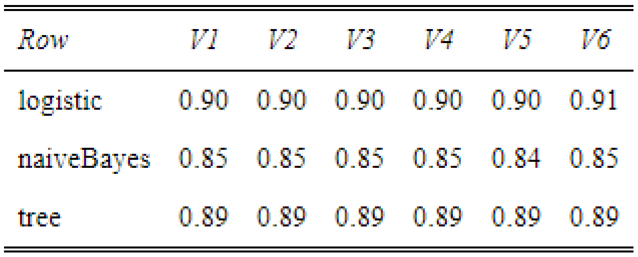
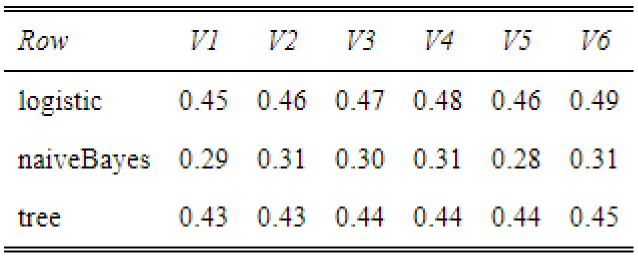


Figure 2.3.: Boxplots of accuracy for each classification method



Row	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*
logistic	0.90	0.90	0.90	0.90	0.90	0.91
naiveBayes	0.85	0.85	0.85	0.85	0.84	0.85
tree	0.89	0.89	0.89	0.89	0.89	0.89

Table 2.5.: Accuracy for each classification method



Row	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*
logistic	0.45	0.46	0.47	0.48	0.46	0.49
naiveBayes	0.29	0.31	0.30	0.31	0.28	0.31
tree	0.43	0.43	0.44	0.44	0.44	0.45

Table 2.6.: Adjusted Rand Index (ARI) for each classification method

Part 2: Clustering

Clustering allows us to better understand how a sample might be comprised of distinct subgroups given a set of variables. This section will focus on the unsupervised learning problem and implement the hierarchical clustering method using Gower distance, partitioning around medoids, and silhouette width.

In order for an algorithm to group observations together, it is initially needed to explore the dissimilarities between observations. The distance metric that we will use for that case is Gower distance, as it can handle mixed data types. The concept of Gower distance is that for each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1. Then, a linear combination using user-specified weights is calculated to create the final distance matrix. Since it requires keeping an NxN distance matrix in memory, we select a subsample of 10000 observations from our original dataset of almost 40000 observations and we proceed to data cleaning.

After calculating Gower's distance, we select an algorithm for clustering. Partitioning around medoids (PAM) is more robust to noise and outliers compared to k-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. The PAM algorithm searches for k representative objects in a data set (k medoids) and then assigns each object to the closest medoid in order to create clusters. Its aim is to minimize the sum of dissimilarities between the objects in a cluster and the center of the same cluster (medoid).

The next step is to select the number of clusters. We use silhouette width, an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster. The metric can range from -1 to 1, where higher values are better. After calculating silhouette width for clusters ranging from 2 to 10 for the PAM algorithm, it can be observed that 2 clusters yield the highest value, close to 1.

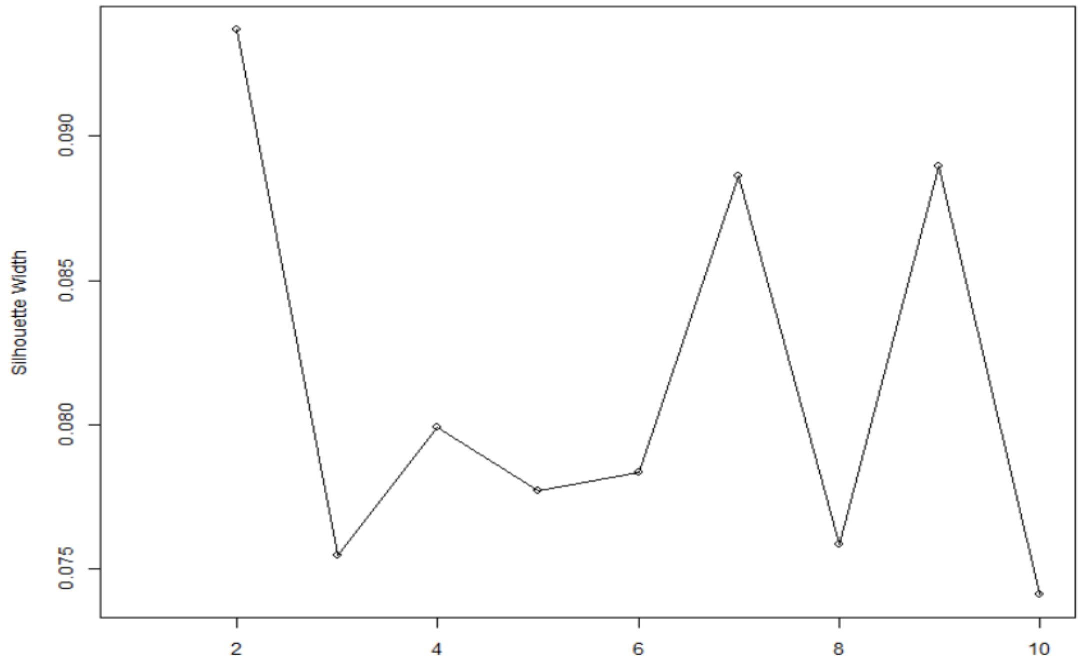


Figure 2.4.: Silhouette width for number of clusters

Finally, we will visualize the results using t-distributed stochastic neighborhood embedding, or t-SNE. This method is a dimension reduction technique that tries to preserve local structure to make clusters visible in a 2D or 3D visualization. Figure 2.4. displays the two clusters that PAM was able to detect. The two clusters are not perfectly separated, but especially the first cluster seems to stand out more. The summary statistics of the two clusters are provided in Appendix, Tables 2.7, 2.8.

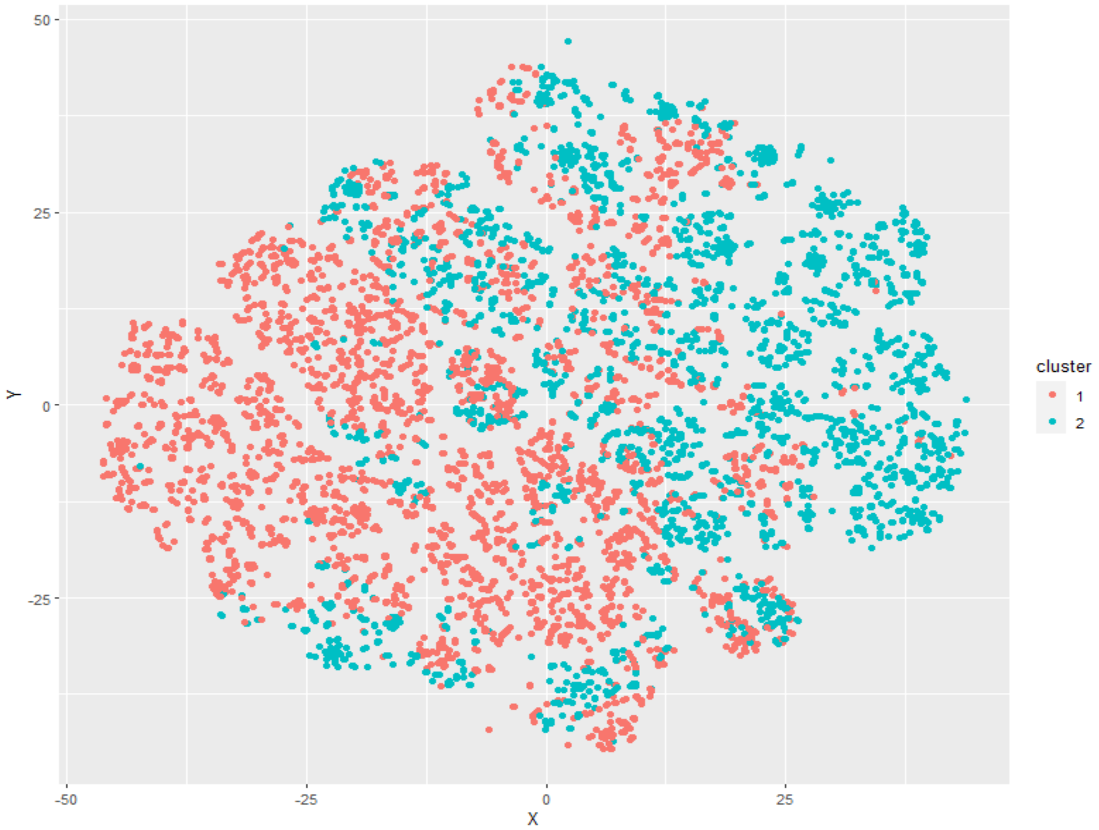


Figure 2.5.: T-SNE visualization of clusters

3. Conclusion

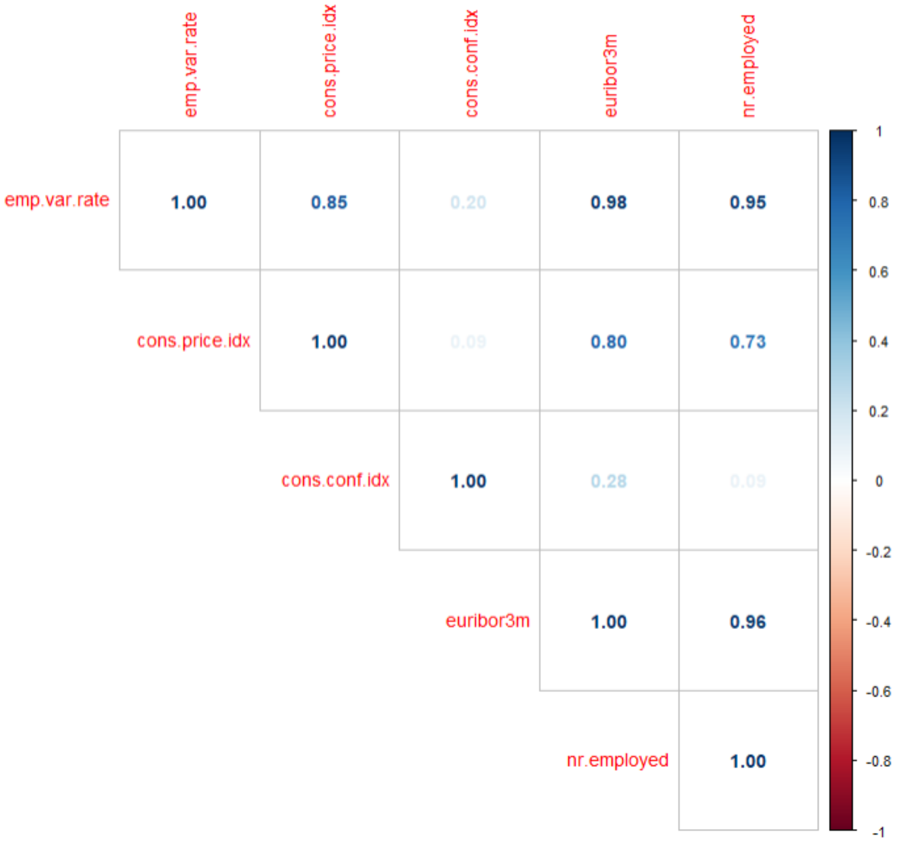
To sum up, regarding classification, it appears that out of the 3 different models presented in this report, logistic regression did the best job in determining true positive values and had the highest accuracy. The results of the decision tree were also very satisfactory, while Naive Bayes did worse with predicting the positive class and had the lowest accuracy. There are also other algorithms that could be used and have perhaps better predictions, for example Support Vector Machines and Random Forest.

Regarding clustering, the results were not equally satisfying, but the algorithm was able to detect two clusters. Gower's distance and partitioning around medoids (PAM) also had some disadvantages, as they required much memory and time for running.

4. References

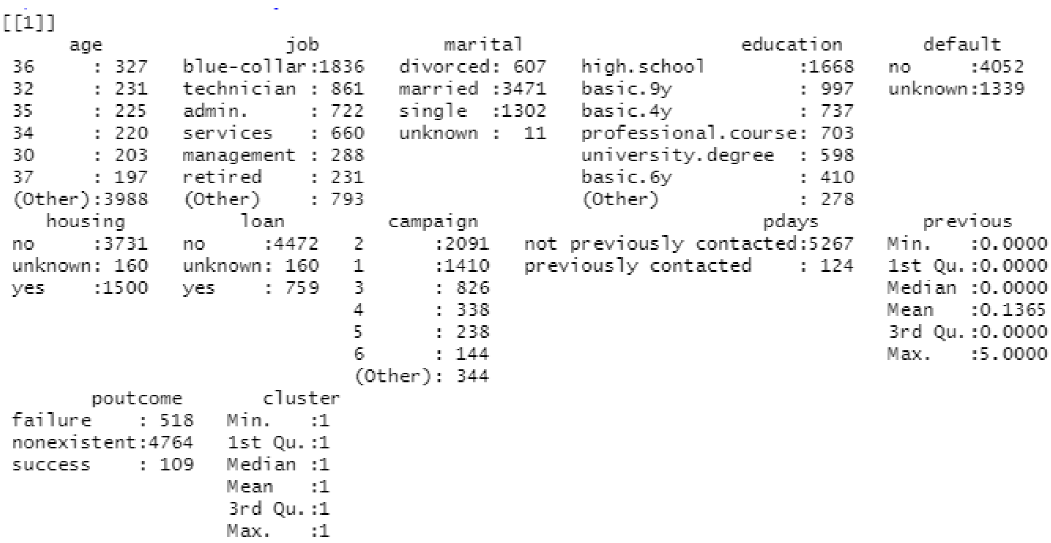
- Statistics for Business Analytics II – course slides.
- Jiang, Y. (2018). “Using logistic regression model to predict the success of bank telemarketing.” International Journal on Data Science and Technology.
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems

5. Appendix



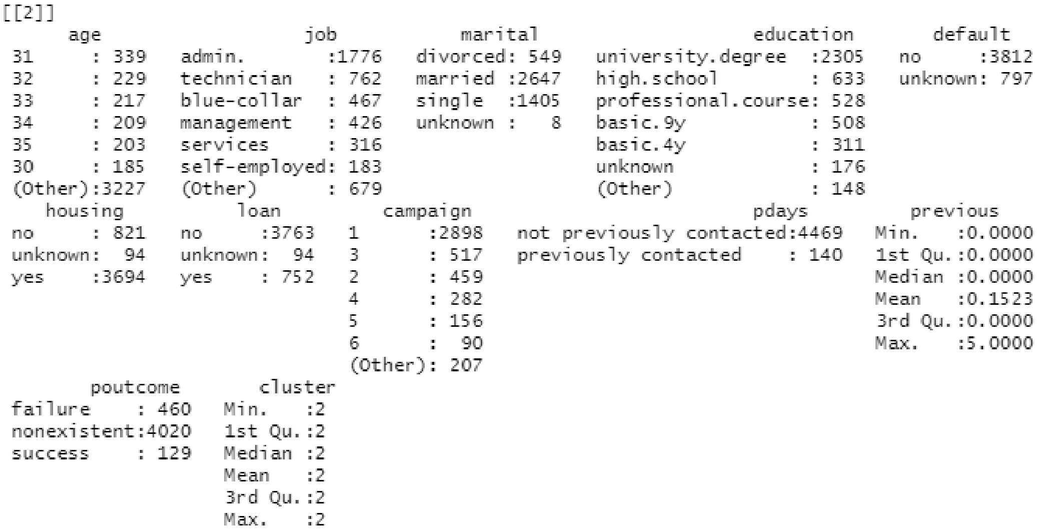
	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
emp.var.rate	1.00	0.85	0.20	0.98	0.95
cons.price.idx		1.00	0.09	0.80	0.73
cons.conf.idx			1.00	0.28	0.09
euribor3m				1.00	0.96
nr.employed					1.00

Table 2.3.: Correlation Matrix



```
[[1]]
      age                  job              marital                     education        default     
 36      : 327   blue-collar:1836   divorced: 607   high.school        :1668   no     :4052  
 32      : 231   technician : 861   married :3471   basic.9y           : 997   unknown:1339  
 35      : 225   admin.     : 722   single  :1302   basic.4y           : 737                 
 34      : 220   services   : 660   unknown :  11   professional.course: 703                 
 30      : 203   management : 288                   university.degree  : 598                 
 37      : 197   retired    : 231                   basic.6y           : 410                 
 (Other):3988   (Other)    : 793                   (Other)            : 278                 
    housing          loan          campaign                         pdays          previous     
 no     :3731   no     :4472   2      :2091   not previously contacted:5267   Min.   :0.0000  
 unknown: 160   unknown: 160   1      :1410   previously contacted    : 124   1st Qu.:0.0000  
 yes    :1500   yes    : 759   3      : 826                                   Median :0.0000  
                               4      : 338                                   Mean   :0.1365  
                               5      : 238                                   3rd Qu.:0.0000  
                               6      : 144                                   Max.   :5.0000  
                               (Other): 344                                                   
        poutcome        cluster 
 failure    : 518   Min.   :1  
 nonexistent:4764   1st Qu.:1  
 success    : 109   Median :1  
                    Mean   :1  
                    3rd Qu.:1  
                    Max.   :1  
```

Table 2.7.: Summary statistics of 1st cluster



```
[[2]]
      age                  job            marital                      education        default
 31     : 339   admin.       :1776   divorced: 549   university.degree  :2305   no     :3812
 32     : 229   technician   : 762   married :2647   high.school        : 633   unknown: 797
 33     : 217   blue-collar  : 467   single  :1405   professional.course: 528
 34     : 209   management   : 426   unknown :   8   basic.9y           : 508
 35     : 203   services     : 316                   basic.4y           : 311
 30     : 185   self-employed: 183                   unknown            : 176
 (Other):3227   (Other)      : 679                   (Other)            : 148
    housing          loan          campaign                           pdays          previous
 no     : 821   no     :3763   1      :2898   not previously contacted:4469   Min.   :0.0000
 unknown:  94   unknown:  94   3      : 517   previously contacted    : 140   1st Qu.:0.0000
 yes    :3694   yes    : 752   2      : 459                                   Median :0.0000
                               4      : 282                                   Mean   :0.1523
                               5      : 156                                   3rd Qu.:0.0000
                               6      :  90                                   Max.   :5.0000
                               (Other): 207
        poutcome        cluster
 failure    : 460   Min.   :2
 nonexistent:4020   1st Qu.:2
 success    : 129   Median :2
                    Mean   :2
                    3rd Qu.:2
                    Max.   :2
```

Table 2.8.: Summary statistics of 2nd cluster