How to do Effective and Successful Bank Telemarketing

Arindam Barman[1], Mohamed Elmoudni[1], Shazia Khan[1], Kishore Prasad[1]

[1] City University of New York (CUNY)

Author note

How to do Effective and Successful Bank Telemarketing

Abstract

The objective of this project is to analyze and improve a Portuguese bank's telemarketing campaign efficiency by identifying socio-economic attributes of customers as the driving factor for term deposit product selection. As methodology, we will be using the Cross Industry Data Standard Process for Data Mining (CRISP DM) framework for this project. We will start with the business case, followed by data exploration, data preparation, modeling, evaluation, and recommendation from final model. The dataset has 16 variables related to customer's socio-economic conditions and about 41188 customer records. The response is binary variable, the campaign response. We will create different models - Logistic Regression, Classification Tree, and Random Forest. To evaluate and select from the three models, we used accuracy, (AUC), F1 score etc. With the given dataset, the response is disproportionate to the population with 10% success. This specific correlation incurred some challenges in the model. Hence we had to use the Area under curve (AUC) metrics for our final selection rather than the accuracy number. Based on our model comparison Random Forest has been found as the most efficient model with AUC score of around 92% for the given case scenario. Among predictor variables, we found that the "duration" variable is the most important predictor; with longer duration calls resulting into more productive discussions and success of the campaign. The next important predictor variables are inter-bank transfer rate (euribor3m) and (nr.employed), high transfer rates and number of bank employees respectively lead to successful campaigns.

Keywords

"Logistics Regression Model, Classification Tree, Random Forest, Area under curve (AUC), Predictive modeling, Bank Telemarketing, Direct Marketing, Data Mining"

Introduction

Banks are increasingly concerned about their investment in marketing campaigns. High and fierce bank competitions have reduced the response rate from marketing campaigns to low, sometimes close to single digit. Consequently, banks have invested aggressively in their marketing campaigns to overcome competition and gain edge over their competitors. Adversely, negative impact of mass campaigns also influences bank's brand and value.

Banking companies have started working on addressing this tradeoff. One solution is to be able to identify customers who may have higher chances of response to a marketing campaign. Although the solution is intuitive, it carries multiple challenges such as methods on how to identify those customers and target them for higher responses, the accuracy of predicting responses, and maintaining response success rate above expectations.

Therefore, our objective in this project is to develop a classification solution to enhance the identification of our target customers, customers that are most likely to respond to our bank telemarketing campaign, develop a model to predict customer response with over 90% accuracy.

Literature Review

There have been few papers that have addressed this requirement. A common thread across all papers was the use of GLM based algorithms. In addition, other algorithms used Neural Networks[1], Random Forests[1], KNN[1], CART[2], Naive Bayes[3] and Support Vector Machines (SVM)[3]. Out of these, Neural Networks and Random Forests seemed to stand out to giving better performances[1].

We have not used KNN in our approach as we cannot interpret the effect of different predictors on our dependent variable[1]. We have not used Neural Networks as it does not fit well to data that was not part of the original training dataset[1]. In our approach, we did not use SVM as it requires a lot of processing power and can sometimes be non-responsive[3].

Data Imbalance[1] was another factor that was considered in one of the papers. This was addressed in that approach by using over or Under sampling, or a mix of both, from the training dataset. However, the results from each of these approaches can vary considerably when applied in a real world situation. It will also differ based on the algorithms that will be applied. We have not addressed this in our approach since we believe that the data imbalance will be inherent in real data and the applied model should appropriately apply some bias.

Based on the literature review, we decided to apply GLM, CART and Random Forests for training the predictive models. Duration was one of the variables highlighted in almost all papers. Some of the papers resorted to extensive feature engineering[1] [3]. However, the results in such papers showed that the basic variables like Duration were the ones that had higher predictive power as opposed to other exotic features. Again in our approach, we did not delve deep into feature engineering and stuck to the basic feature engineering. The advantages of extensive feature engineering seemed to be negligible.

Methodology - CRISP-DM

In this project we will be using CRISP DM methodology.

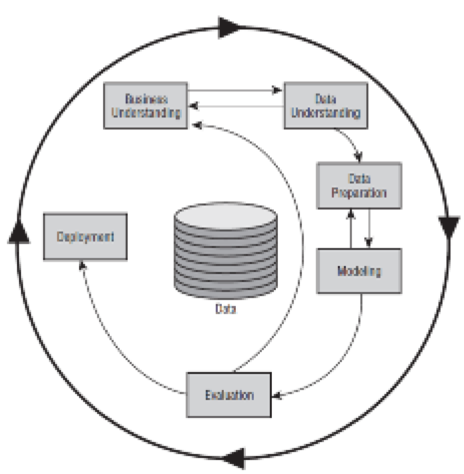


Figure 1. CRISP-DM Methodology

As per wikipedia, “Cross Industry Standard Process for Data Mining, commonly known by its acronym CRISP-DM, was a data mining process model that describes commonly used approaches that data mining experts use to tackle problems. Polls conducted at one and the same website (KDNuggets) in 2002, 2004, 2007 and 2014 show that it was the leading methodology used by industry data miners who decided to respond to the survey. CRISP-DM breaks the process of data mining into six major phases.[9]. The sequence of the phases is not strict and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions and subsequent data mining processes will benefit from the experiences of previous ones."

Below is the summary of the CRISP-DM methodology:

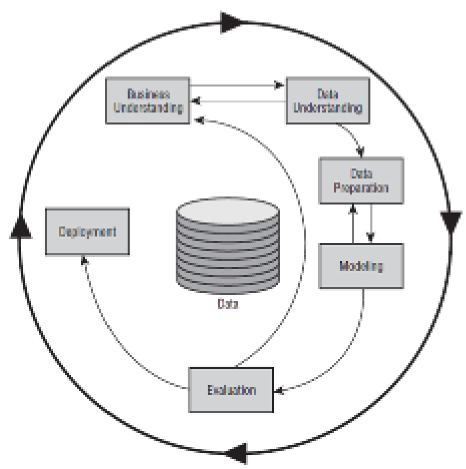


Figure 2. CRISP-DM Methodology

Business Understanding

The data consists of client's personal and financial activity profile. In addition, past campaigns' statistics and results were collected for further analysis. We will be leveraging all the collected data in our analysis in our project. The client data will help us identify potential customers that would respond to our campaign. Financial activities will help create financial profiles of customers. And finally past campaign results will be used for our prediction models.

Data Exploration

In this section, we will using exploratory plots, predictor and response variable association, and counts of response by each variable. During the data exploration, using various charts and tables we will analysis how predictor variables impact the response variable. In addition, we will identify outliers, missing data, as well as any invalid data.

Data Preparation

In this section, we will treat outliers, missing data, as well as "unknowns" values. In addition, as most of our variables are categorical, we will create dummy variables to convert categorical data to numeric. Data treatment will mostly consist of using median values for categorical and mean for numeric values.

Modeling

As our response variable is binary, we will confine our modeling techniques to three modeling methods: Logistics Regression, Classification Tree, and Random Forest Model. Below is brief description of each modeling technique.

Logistics Regression. Logistic Regression is a probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor. Logistics model doesn't suffer a lot from severe class imbalance. Logistic Regression creates log odds of the response as a linear function of predictor variables. Many of the categorical predictors in the data set for this project have sparse and unbalanced distributions. Using logistics model with the given set of data would need adjustment of variables to fine tune the model.

Classification Tree. Classification Tree is used to predict the outcome of a categorical response variable. The purpose of the analyses via tree-building algorithms is to determine a set of logical conditional split that permit accurate classification of cases and accurate prediction. Effectiveness of classification tree model with binary variable is one of the reason for selection for this analysis study. This model though has problem with over fitting. We will also create Random Forest model to overcome that.

Random Forest Model. Random Forests technique grows many classification trees for given set of response and predictor variables. Each tree gives a classification, and all the outputs from different trees are "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Over fitting problem with the classification tree can be overcome by this approach with weighted average of more number of trees. This method is good for prediction but a little bit difficult to interpret. Since we are facing the binary category, Random Forest is a good classification method to try.

Evaluation

The objective is to build a model that can predict likelihood of response from a customer. The following evaluation criteria will be used to assess our model performance:

- The Hosmer-Lemeshow test assesses the model calibration and how predicted values tend to match the predicted frequency when split by risk decides. This test will be used for Logistics regression model validation.
- AUC along with model Accuracy will be used for model evaluation. Accuracy is calculated based on certain threshold; whereas AUC is overall performance evaluation of a model as various points.

AUC criteria will be given more weight to assess our model evaluation for its high predictability for our dataset type as it has binary response variable.

Experimentations and Results

Data Exploration

The dataset is available on the UC Irvine Machine Learning Repository website. There are two different data sets available. We chose to use the dataset with additional attributes, "bank-additional", which has 41,188 records and it has 20 attributes and 1 response variable. The data consists of four groups of information.

- Client's personal information
- Client's bank information
- Bank's telemarketing campaign information
- Social and economic information

The main problem with the dataset is that it consists of many missing values which are labeled "Unknown". The missing data consists of 26% of the data. We decided to retain the missing data to help with our regression modeling. The other problem with the data is that only 12% of the data shows the response variable to be "y". We looked at each variable and the unique values contained in each variable and what they represented. We can divide the variables in the following three categories:

- Binary values of "yes" and "no" with null values given as "unknown".
- Categorical values with "unknown" as missing values. The categorical variables require dummy variables to be created for each unique value. We included "unknown" as one of the dummy variable. - numeric values with "999" as indication of null value. We created a variable to indicate if the data was missing or present.

Also following two areas have been explored in the training data set.

- Missing values and Unique Values
- Variables relationship to y (y was given as our response variable)

We also investigated how the initial data aligns with a typical logistic model plot. Recall the Logistic regression is part of a larger class of algorithms known as Generalized Linear Model (GLM). The fundamental equation of generalized linear model is:

g(E(y)) = a+ Bx1+B2x2+ B3x_3+…

where, g() is the link function, E(y) is the expectation of target variable and B0 + B1x1 + B2x2+B3x3 is the linear predictor B0,B1,B2, B3 to be predicted. The role of link function is to "link" the expectation of y to linear predictor. In logistic regression, we are only concerned about the probability of outcome dependent variable success or failure. As described above, g() is the link function. This function is established using two things: Probability of Success as p and Probability of Failure as 1-p. p should meet following criteria: It must always be positive (since p >= 0) It must always be less than equals to 1 (since p <= 1).

Variable	Data.Type	Analysis
age	Numeric	No significant trend with responses variable, better response with age grp<30 & >55
job	Catagorical	12 levels, proportion of responses from admin and blue collar job profiles are higher
marital	Catagorical	4 levels, % response from marital status from single is greater compare to other grp
education	Catagorical	8 levels, responses from education with university degree are higher
default	Binary	3 levels, response is from no default group is dominant and some responses from unknown
housing	Binary	3 levels, no significant difference in association for three different groups
loan	Binary	4 levels, no significant difference in association for three different groups
contact	Catagorical	2 levels, responses from cellular contact is higher
day_of_week	Catagorical	5 levels, response from customer is better on Wed,Thu, Tue
month	Catagorical	10 levels, there is significant variations of responses from Customers
duration	Numeric	closely associated with response variable with threshold for positive response
campaign	Numeric	Number of campaign has impact on positive response of the campaign
pdays	Numeric	This variable does not seem to have strong relationship with response variable
previous	Numeric	previous contacts seems to have influence on the positive response of the campaign
poutcome	Catagorical	have relationship with campaign outcome, earlier success has better response to positive outcome
emp.var.rate	Numeric	lower the variation rates higher the number of positive outcome
cons.price.idx	Numeric	lower consumer price index seems to have higher positive response rate
cons.conf.idx	Numeric	lower confidence index brings more success to the campaign as people tend to spend less that time
euribor3m	Numeric	lower rate has association with more number of positive cases
nr.employed	Numeric	lower the number of employee higher the number of positive responses

Figure 3. Variable Analysis

Data Preparation

The main objective in the transformations is to achieve linear relationships with the dependent variable or, consequently, with its logit. As discussed above, we carried out the following transformations:

- Convert Binary variable to 0 and 1 from yes and no

- Create dummy variables for categorical variables
- Data Summary Analysis
- Correlation of Variables with y

Model Building

In this section experimentation will be carried out with the data by formulating three different types of models with three different approaches. The following are the three different approaches that will be used here:

- Model 1: This model will be created by using logit function of Generalized Logistics Model (GLM).
- Model 2: This model will be created by using Classification tree function.
- Model 3: This model will be created by using classification technique Random Forests model.

There are two data set given with the business case training and test set. Training set will be used to train the model and the test set will be used to evaluate the model performance.

Logistics Regression - Model 1. Logistics regression function GLM has been used to classify the campaign response variable. Basic model generated by using GLM function has been enhanced by making necessary adjustments to non-associated predictor variables shown as “NA” in basic model output. Next the model has been validated by using k=5 fold cross validation press to do necessary adjustment to the model.

A total 10 iterations been performed before final selection of variables were made. AIC value from model 1 and model1_update (enhanced) model were same 13776. Hence removing variables from basic model does not help performance wise but reduced complexity with less degrees of freedom. By using k=5 cross validation, ($delta) error value came out to be low 0.06289177.

Variable Importance	Variables	Odd Ratio
***	duration, campaign, emp.var.rate	1.004, 0.957, 0.182
***	cons.price.idx, job_blue_collar, contact_telephone	8.64, 0.615, 0.541
***	month_may, month_aug, month_nov, month_mar	0.481, 1.80, 0.526, 5.72
**	education_secondary, month_jun, poutcome_failure	0.858, 0.443, 0.448
*	cons.conf.idx, job_housemaid, job_services	1.022, 0.661, 0.707
*	job_admin., job_technician, job_self_employed	0.778, 0.772, 0.680
*	job_entrepreneur, day_of_week_mon, day_of_week_wed	0.672, 0.844, 1.185

Figure 4. Variable Importance

Classification Tree - Model 2. The basic idea of classification tree model is to predict a response variable y for the campaign from predictor variables. The model does its prediction by growing a binary tree. At each node in the tree, a test is applied to one of the inputs. Depending on the outcome of the test, two routes will be created and decision will be made to either traverse to the left or the left of the right of the mode. Eventually a leaf node is reached where a prediction is made about the binary outcome of campaign response. Model 2 has been rated using the Classification function from ROCR package. Basic model has been optimized using prune function.

The following are the most important variables from this model: duration, nr.employed, euribor3m, emp.var.rate, cons.conf.idx, cons.price.idx. Total 6 leaves (decision points) have been formed from this model. Complete Classification tree is given below in the diagram, figure 5.

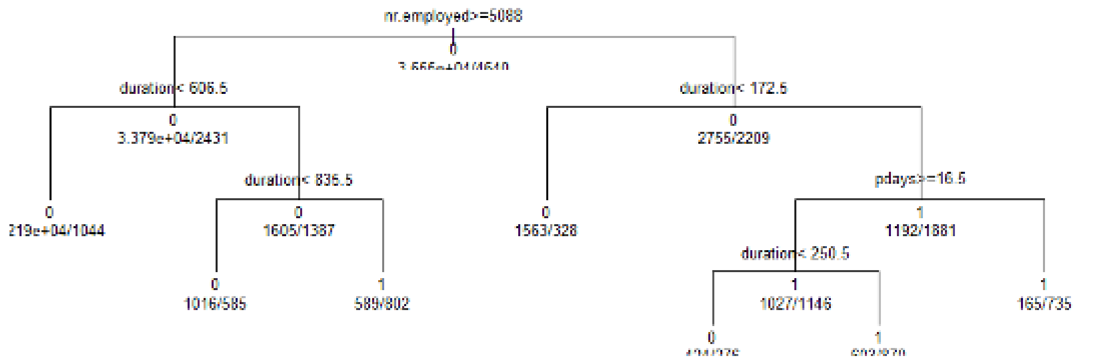


Figure 5. Classification Tree

RandomForest- Model 3. In Random Forests many classification trees are formed to classify campaign response variable y. Each tree creates separate set of classification, each tree is voted for performance for that classification. The forest chooses the classification having the most votes (over all the trees in the forest). One model will be created using this method with tree size 50. Then this model will be evaluated with a model of tree size 100.

From figure 6, it can be seen that classification error rate to classify negative responses reduces with the increase in number of trees. However, there is no significant change in error rate for positive response. There is only a slight reduction in error rate for negative responses when tree size is increased to 100 from 50. The number of variables that were tried at each split is 7 with negative classification rate of 0.03 and positive classification error rate of 0.51. Below is a chart showing the importance of various variables used in the model.

Results from Models

Logistic Regression Results. The result from Logistics Regression model has a very high accuracy rate of 91.42% when the model was evaluated using the validation data set. The AUC for this model was comparatively lower (0.702), which indicates poor fit of the model. By using Hosmer-Lemeshow goodness-of-fit (GOF) tests, when model was evaluated, p value came to be greater than 0.05. With this test, if the p value is lower than 0.05 model is rejected and if it's high, then the model passes the test. The egression model passed this

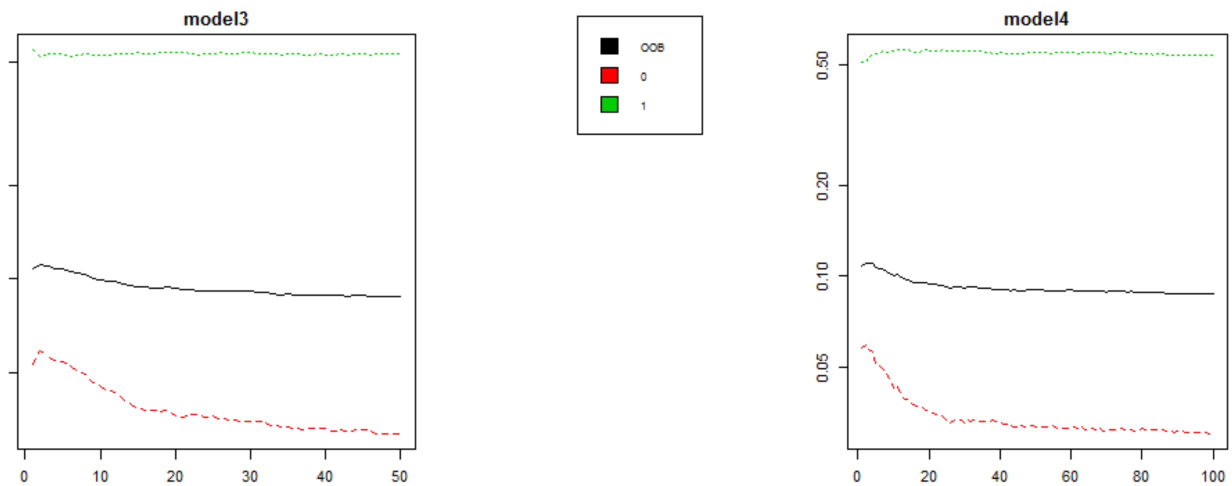


Figure 6. Classification Error Rate Comparison

test.

Classification Tree Results. The results from the Classification Tree Model-this model has also very high accuracy rate of 91.81% which is very good fit. The model also has AUC value of 0.865 which seem to be in line with given high accuracy.

Random Forest Results. The result from Random Forest Model-The model created using Random forest has an accuracy of 98.64% which is extraordinary results and gives a rise to a suspicion. The model is able to separate out the classification based on certain variable. When we looked at the importance of variable "duration" it becomes apparent that this variable is being used in a big way to classify response accurately. It can be seen that this model also shows similar kind of trend in classification of data in earlier stages with very stiff line till true positive rate of 0.4 and then sharp increase in false positive rate.

Discussion and Conclusions

Table 1

Comparison of the Models

	Model	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	GLM	0.9142996	0.0857004	0.4323725	0.6678082	0.9331069	0.3607211	0.7029638
2	CRT	0.9181840	0.0818160	0.5343681	0.6548913	0.9440149	0.4377405	0.8650875
3	RF	0.9881039	0.0118961	0.8980044	0.9926471	0.9876044	0.9002912	0.9485933

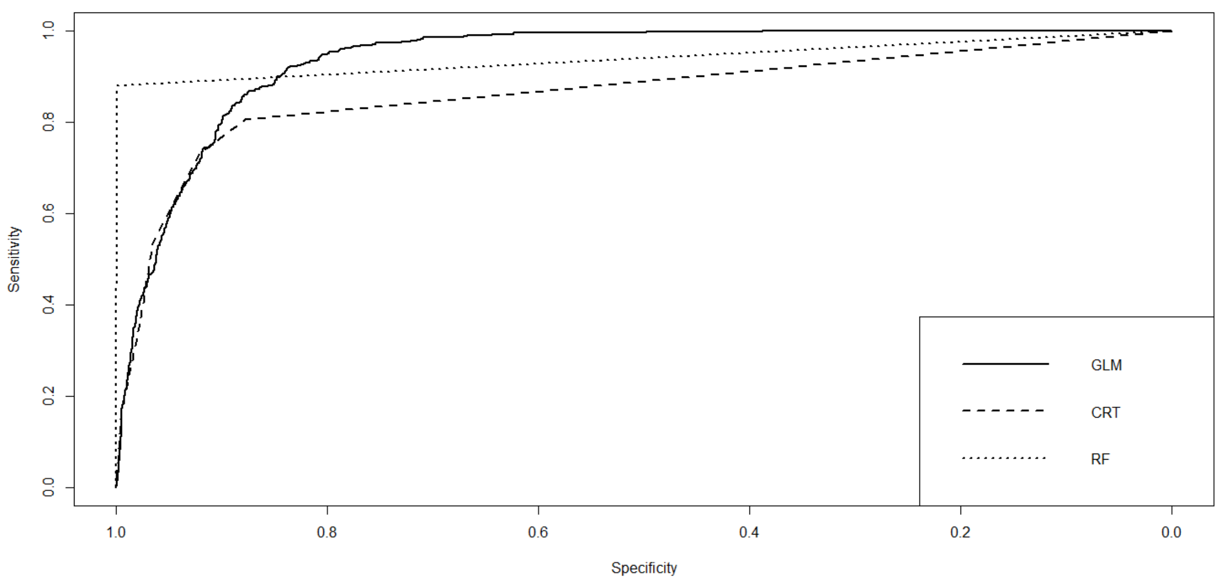


Figure 7. ROC Curves

Final model selection

Based on the Accuracy of the model, model 1 and model 2 are very close around 91% accuracy with probability threshold of 0.5. Model 3 has much higher value of 98%. However, the Accuracy is not always the key criteria for a model as Accuracy is calculated based on a defined threshold. In addition, due to imbalance of data of 10% to 90%, the distribution of response variable forced to choose the model based on other criteria. The model based on

AUC value is model 3 having AUC value of 0.9398 which is a very good score. Model 3 stands out among the three models.

Key predictor variables

For all three models it is found variables "duration" is the most important variables by far. The duration variable has positive impact in campaign outcome. It could be due to the fact that longer the customer stays on phone, a more productive conversation is taking place to get the customer start their term deposit Account. The variable "euribor3m" is also most important variable which denotes inter-bank interest rate in Eurozone. The term deposit interest rates are generally interlinked and tend to go up together. This variable has positive impact on response variable. The predictor "nr.employed" denotes number of employees for the bank. This variable also has positive impact on campaign response. Also, the more number of employees, the more visible the bank is, and in turn more customers it gets through the campaign. Among the negative variables "emp.var.rate" has negative impact on the response. A negative rate of this variable indicates issues with economy and lower economic activities. That in turn could impact the savings rate and people tend to use their savings during such time.

Shortcomings

The Imbalance of response variable only 10% of population, it was the main shortcomings that we have in the model creation. This issue has been addressed partially by using AUC Area Under Curve as the criteria for model selection.

Final Recommendation

In conclusion, it can be suggested to the bank management that the focus should be given in hiring more qualified people, performing more quality and persistent phone calls. In addition, management should try to launch their campaigns during stable macroeconomic environment to maximize their return on investment (ROI)

References

Appendix

- Supplemental tables and/or figures.
- R statistical programming code.

Data Analysis details

Table 2

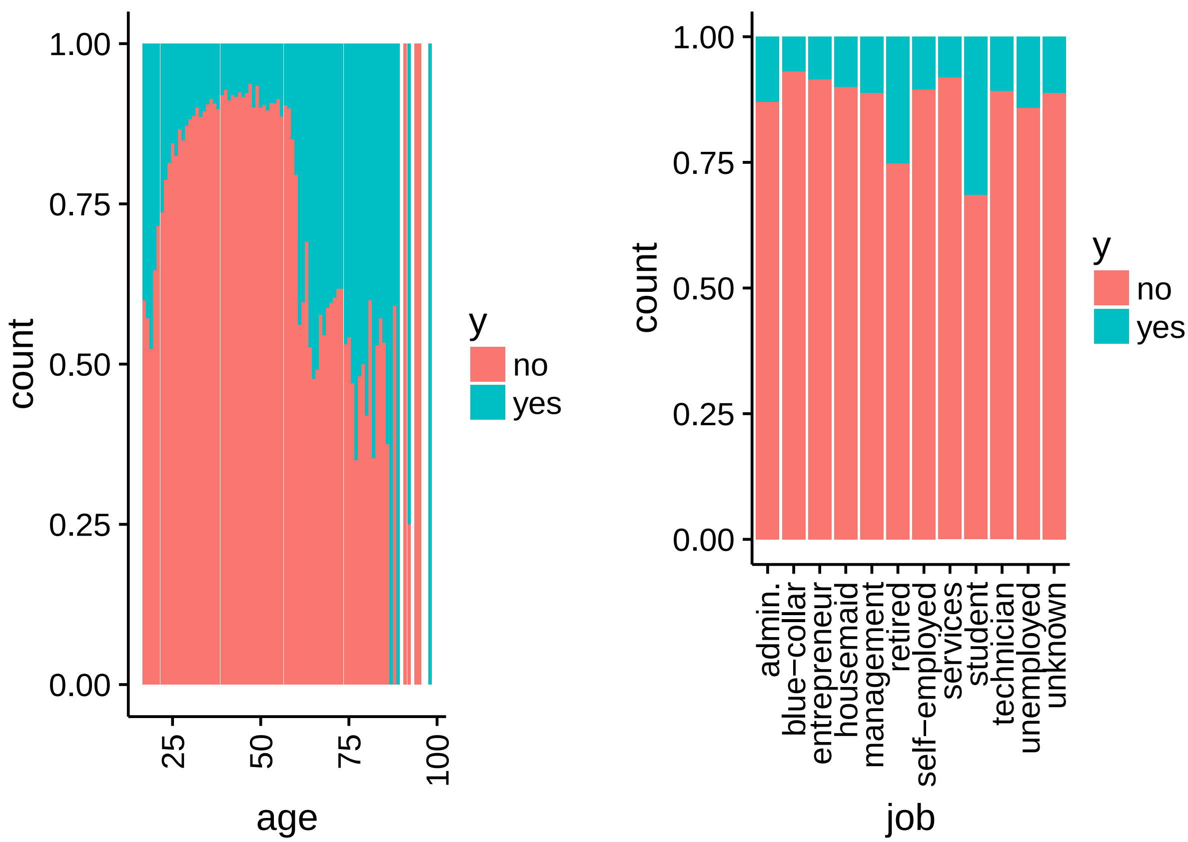
Variable Description

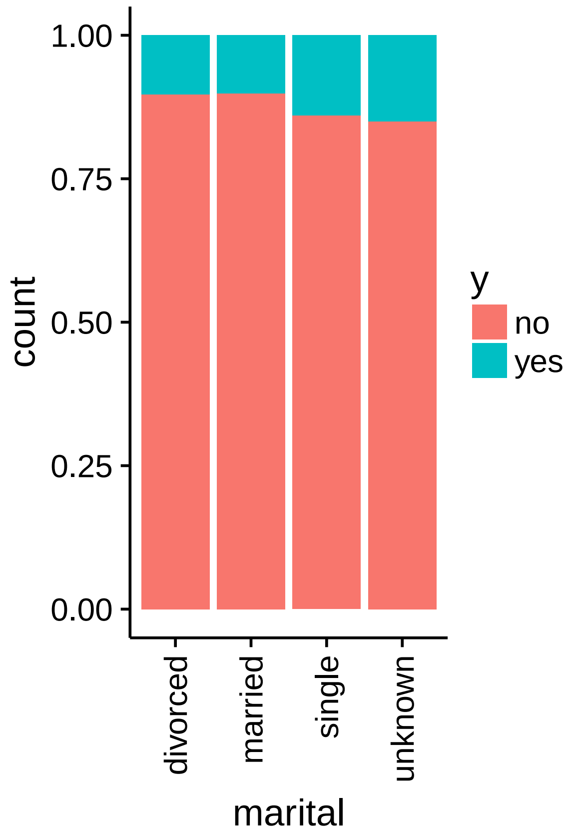
Variable	Data.Type	Type	Description
age	Numeric	Predictor	Client's age
job	Catagorical	Predictor	Client's job
marital	Catagorical	Predictor	Client's marital status
education	Catagorical	Predictor	Client's education level
default	Binary	Predictor	Credit in default?
balance	Numeric	Predictor	Client's average yearly balance, in euros
housing	Binary	Predictor	Client has housing loan?
loan	Binary	Predictor	Client has personal loan?
contact	Catagorical	Predictor	Client's contact communication type
day	Catagorical	Predictor	Client last contact day of the month
month	Catagorical	Predictor	Client last contact month of year
duration	Numeric	Predictor	Client last contact duration, in seconds
campaign	Numeric	Predictor	Client number of contacts performed during this campaign
pdays	Numeric	Predictor	Client days that passed after first contact
previous	Numeric	Predictor	Number of contacts performed before this campaign
poutcome	Catagorical	Predictor	Outcome of the previous marketing campaign
emp.var.rate	Numeric	Predictor	Quarterly employment variation rate

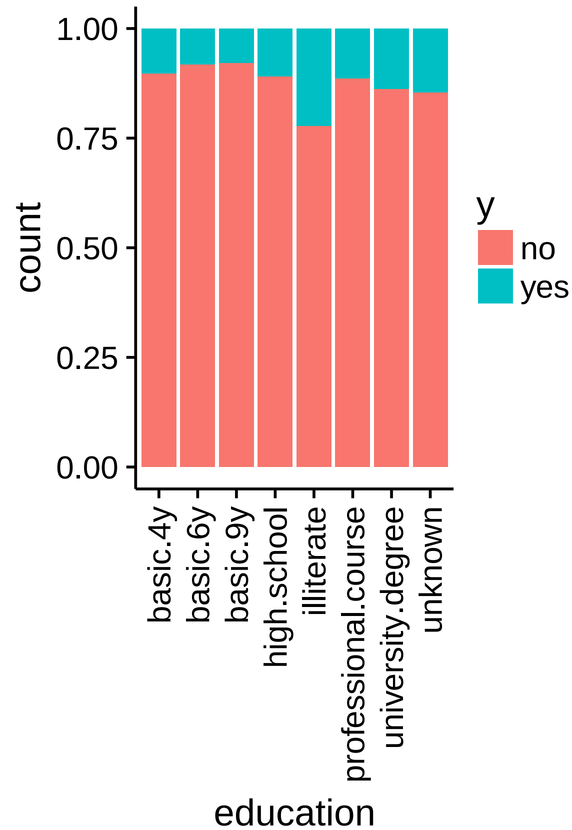
Variable	Data.Type	Type	Description
cons.price.idx	Numeric	Predictor	Monthly consumer price index
cons.conf.idx	Numeric	Predictor	Monthly consumer confidence index
euribor3m	Numeric	Predictor	Daily euribor 3 month rate
nr.employed	Numeric	Predictor	Quarterly number of employees
y	Binary	Response	Has the client subscribed a term deposit?

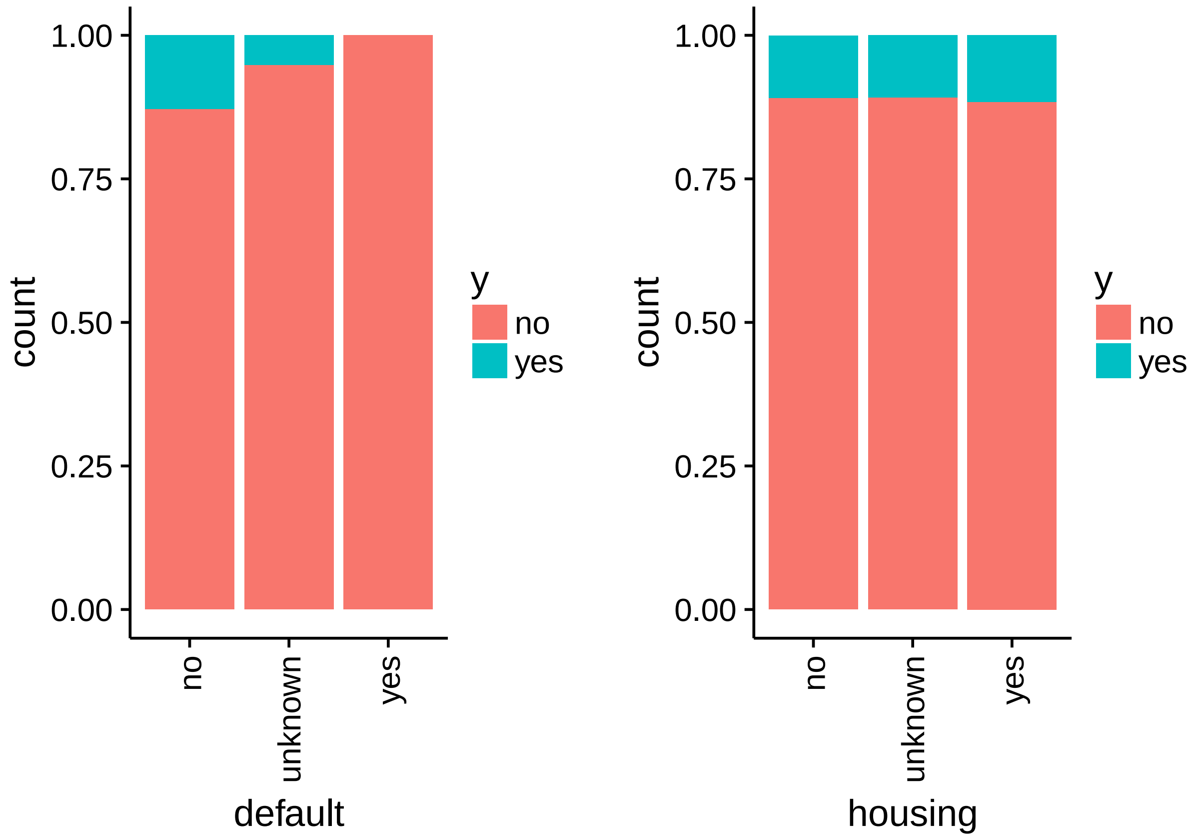
Variable Description.

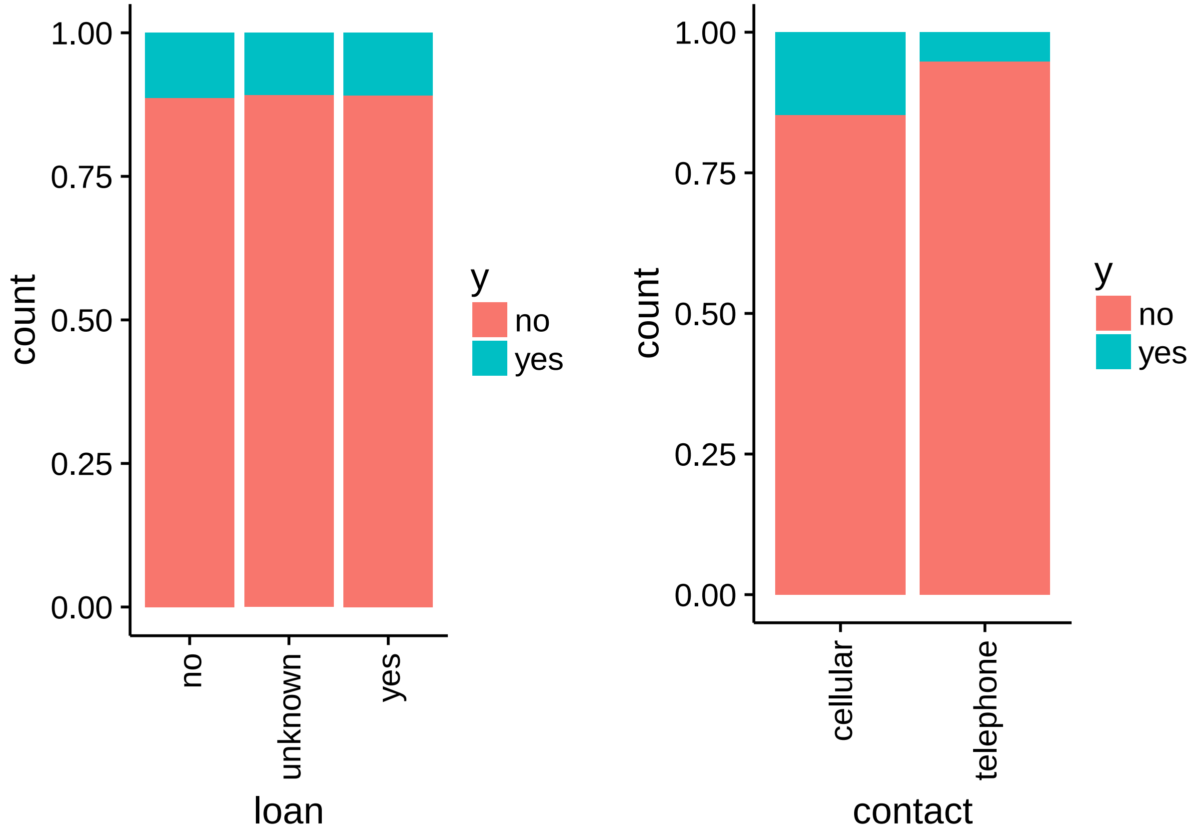
Predictor and Response variable Association.

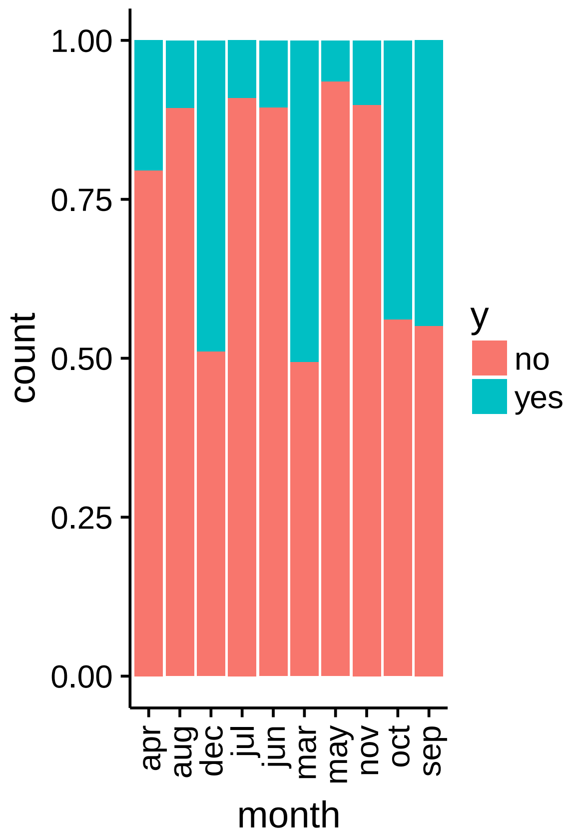


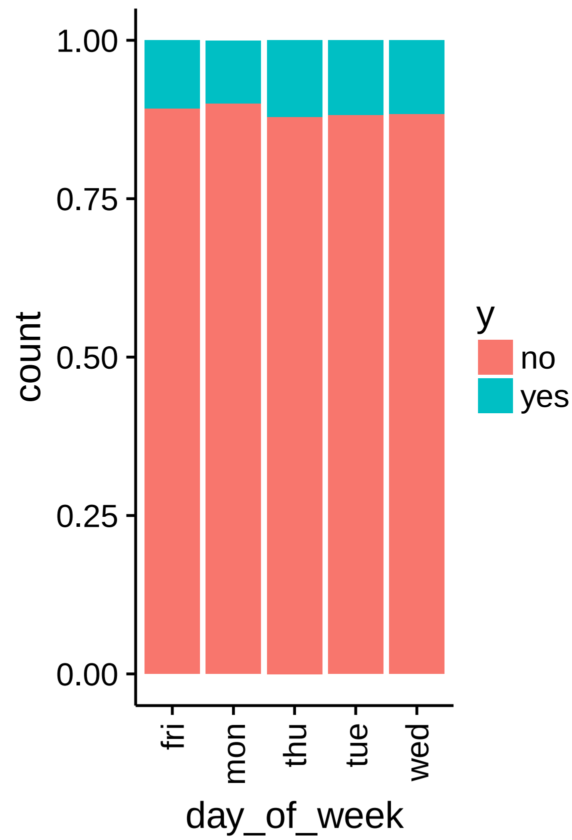

count
1.00
0.75
0.50
0.25
0.00
divorced
married
single
unknown
marital
y
no
yes

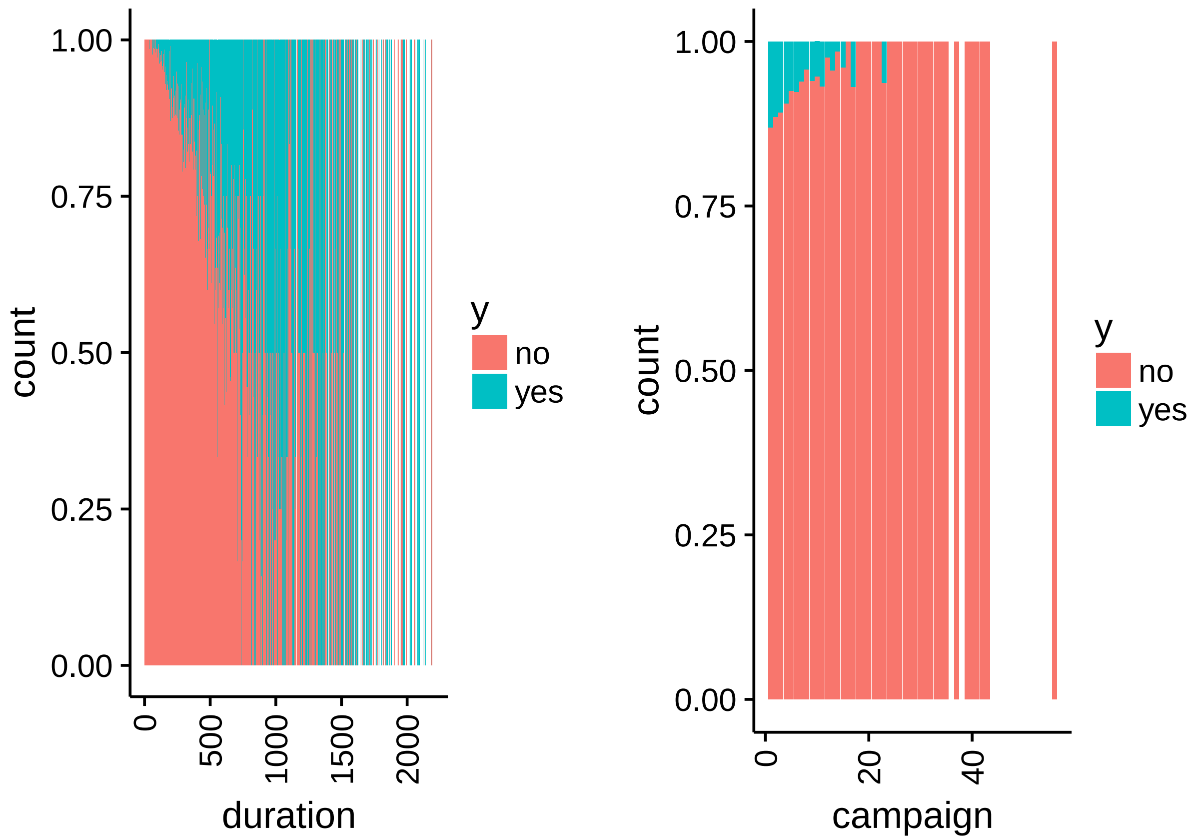

count
1.00
0.75
0.50
0.25
0.00
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education
y
no
yes

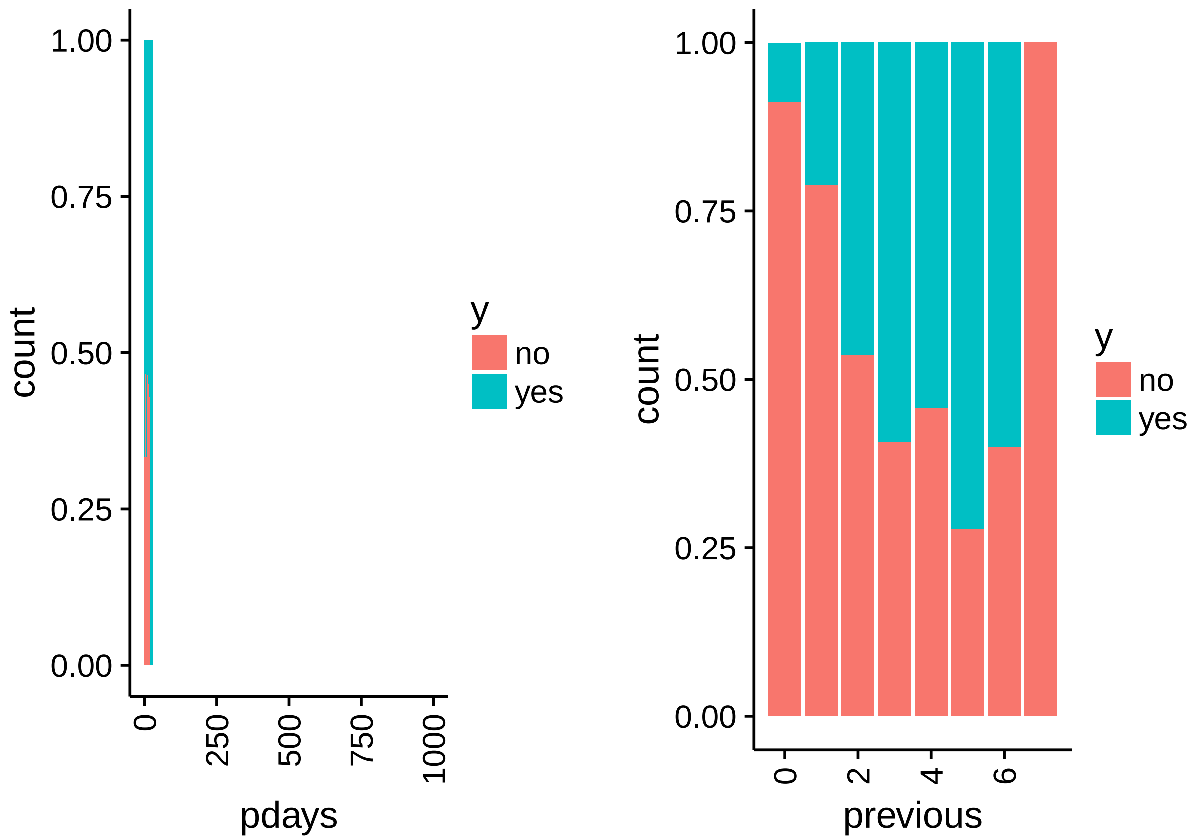

count
1.00
0.75
0.50
0.25
0.00
no
unknown
yes
default
y
no
yes
count
1.00
0.75
0.50
0.25
0.00
no
unknown
yes
housing
y
no
yes

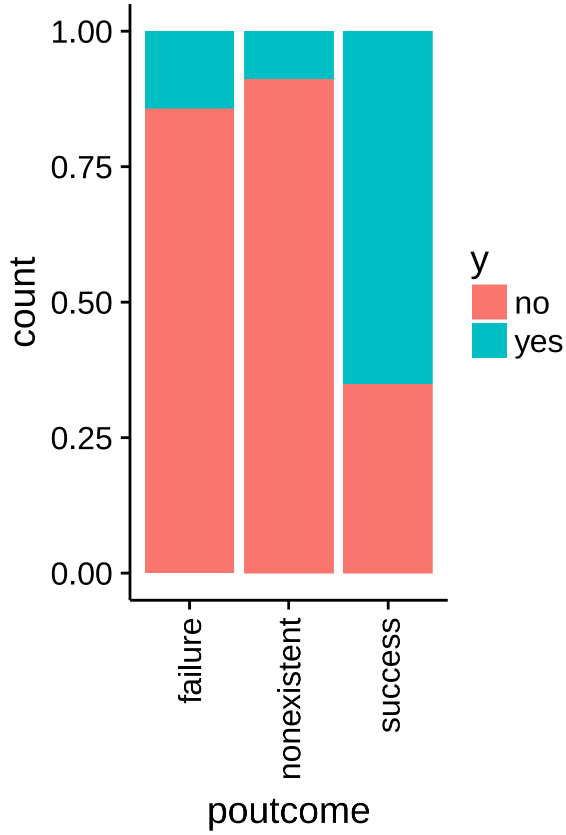

count
1.00
0.75
0.50
0.25
0.00
no
unknown
yes
loan
y
no
yes
count
1.00
0.75
0.50
0.25
0.00
cellular
telephone
contact
y
no
yes

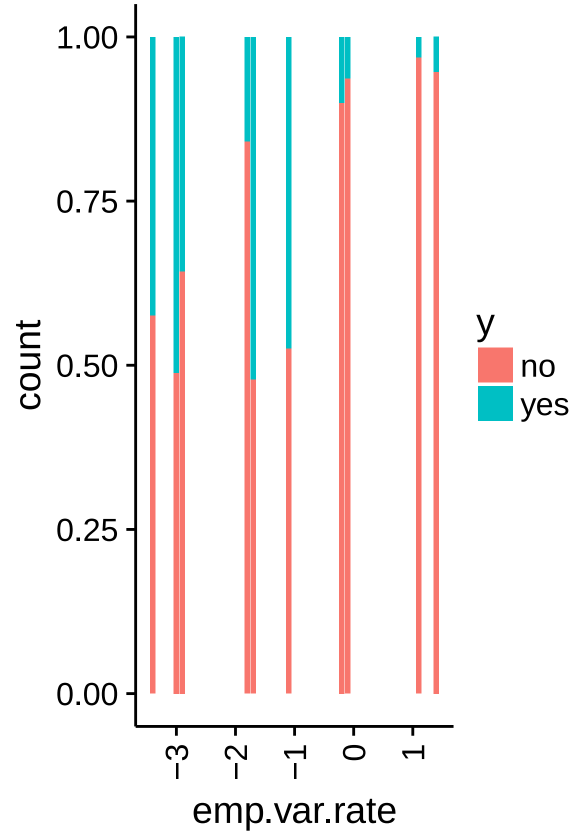

1.00
0.75
0.50
0.25
0.00
count
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month
y
no
yes

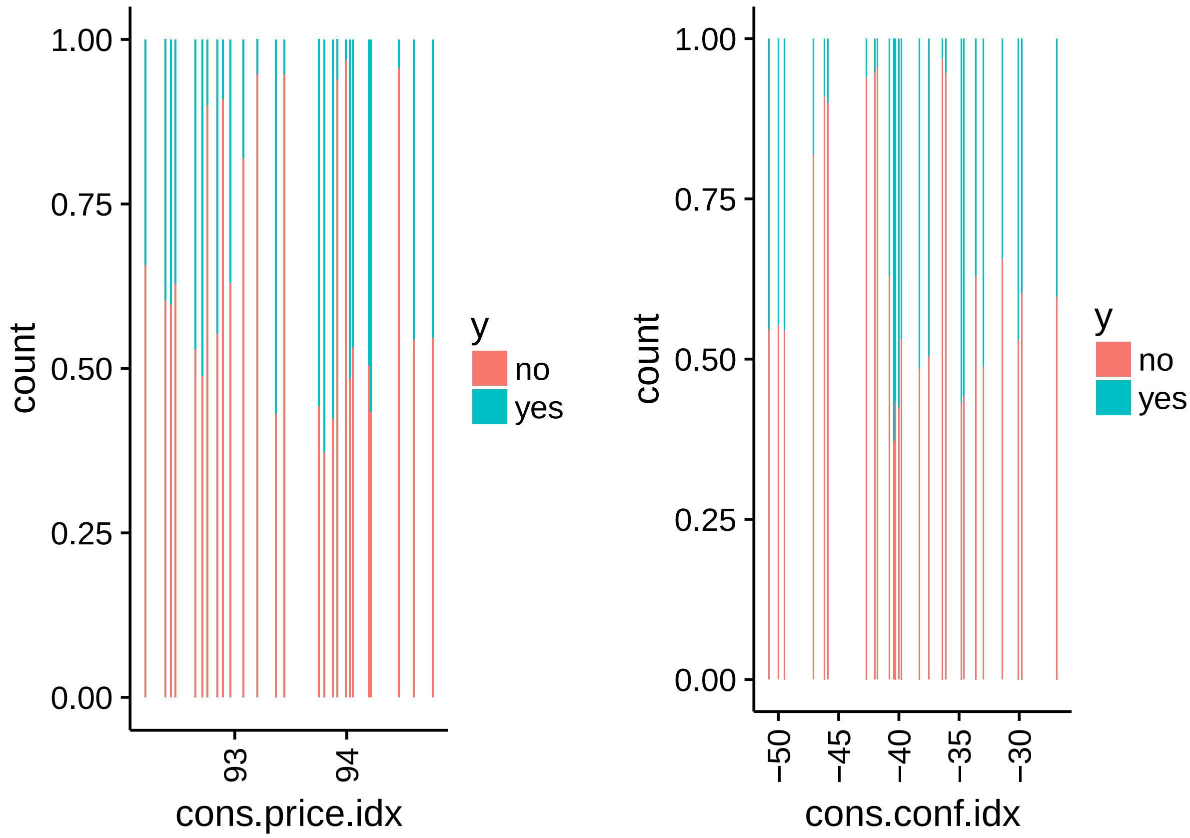

1.00
0.75
0.50
0.25
0.00
count
fri
mon
thu
tue
wed
day_of_week
y
no
yes

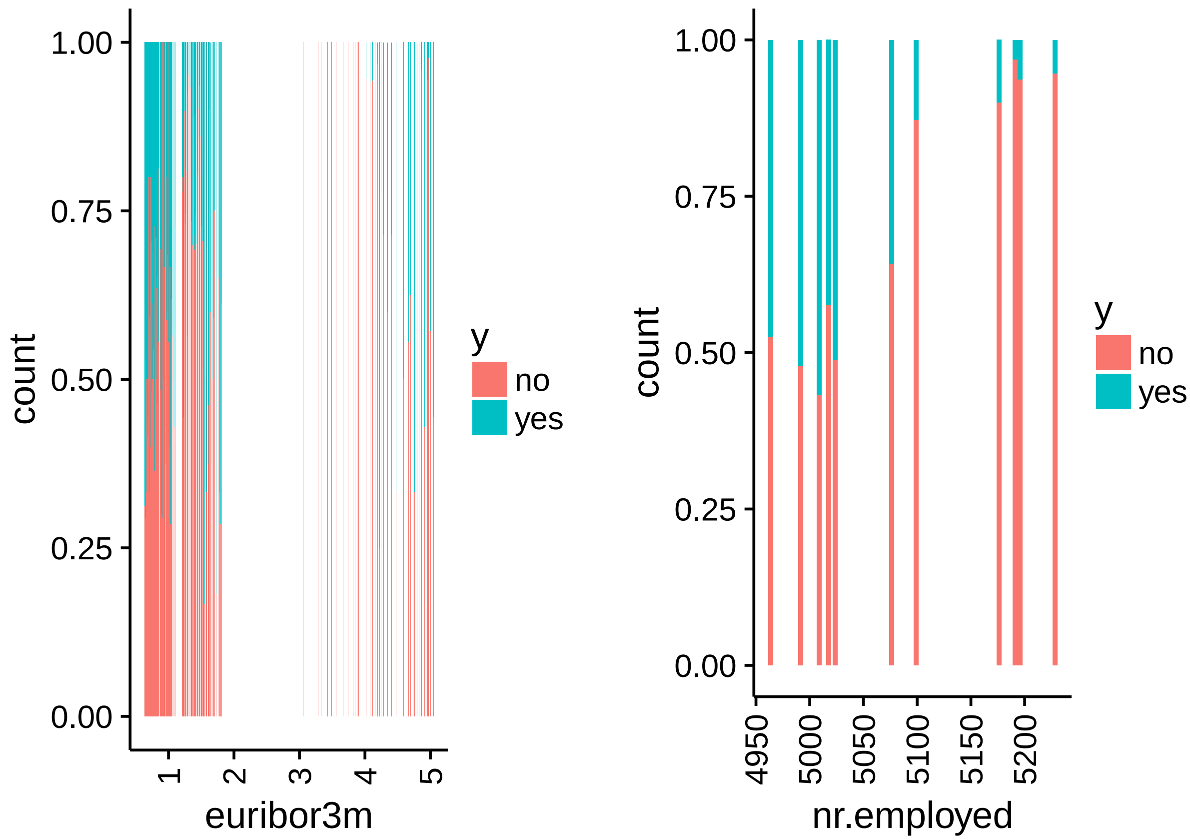

count
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
duration
y
no
yes
count
1.00
0.75
0.50
0.25
0.00
0
20
40
campaign
y
no
yes


1.00
0.75
0.50
0.25
0.00
count
0
250
500
750
1000
pdays
y
no
yes
1.00
0.75
0.50
0.25
0.00
count
0
2
4
6
previous
y
no
yes


1.00
0.75
0.50
0.25
0.00
count
failure
nonexistent
success
poutcome
y
no
yes


1.00
0.75
0.50
0.25
0.00
count
−3
−2
−1
0
1
emp.var.rate
y
no
yes


count
1.00
0.75
0.50
0.25
0.00
93
94
cons.price.idx
y
no
yes
count
1.00
0.75
0.50
0.25
0.00
−50
−45
−40
−35
−30
cons.conf.idx
y
no
yes



Unique Value & Missing value

We see that there are no missing values in our dataset as shown in table 2 and graph format. The unique values are given in the table

Table 3

Missing Values

	Missing Values
age	0
job	0
marital	0
education	0
default	0
housing	0

	Missing Values
loan	0
contact	0
month	0
day_of_week	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	0
emp.var.rate	0
cons.price.idx	0
cons.conf.idx	0
euribor3m	0
nr.employed	0
y	0

Table 4

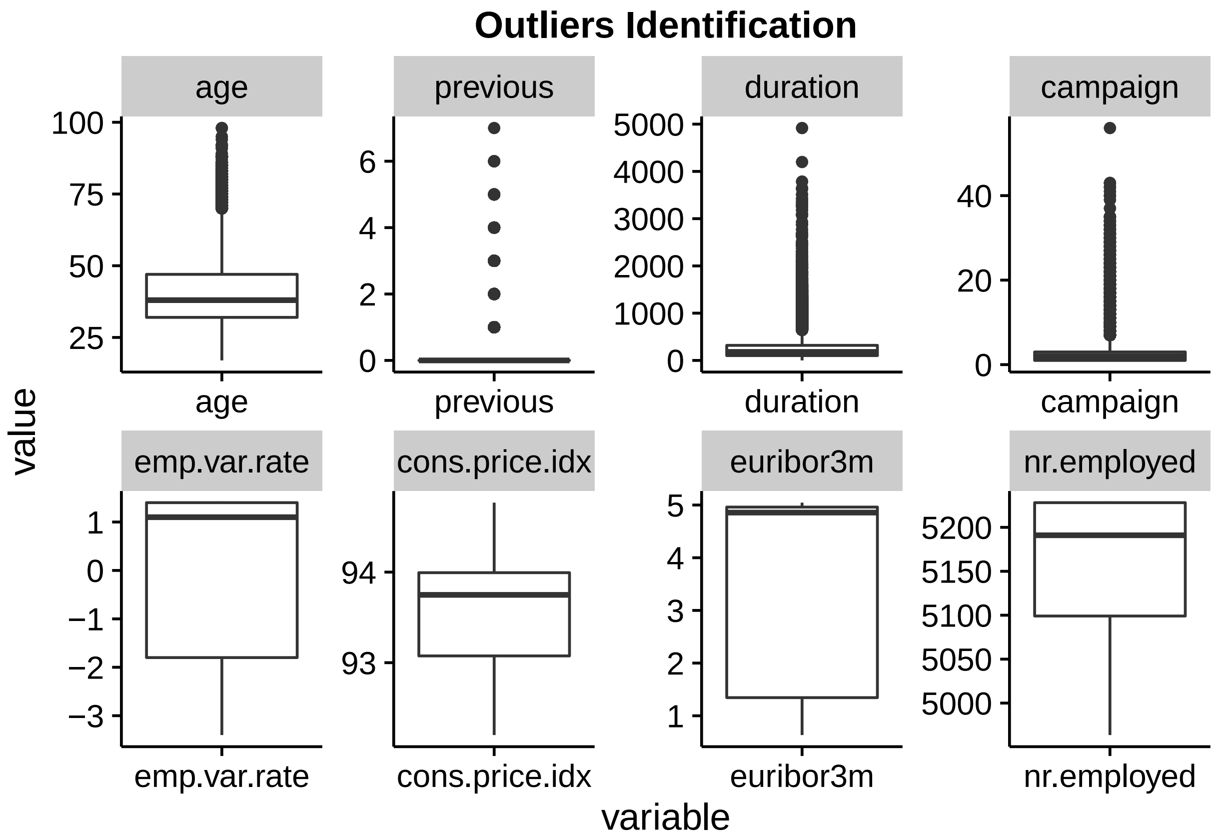
Unique Values

	Unique Values
age	78
job	12
marital	4
education	8
default	3
housing	3

	Unique Values
loan	3
contact	2
month	10
day_of_week	5
duration	1544
campaign	42
pdays	27
previous	8
poutcome	3
emp.var.rate	10
cons.price.idx	26
cons.conf.idx	26
euribor3m	316
nr.employed	11
y	2

Data Summary post conversion.

Outliers Analysis.



Analysis of link functions for given variables.