How to do Effective and Sucessful Bank Telemarketing

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

[1] City University of New York (CUNY)

Author note

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 we have been around 41188 customer records. The response is binary variable, the campaign response. We will create different models - Logistics Regression, Classification tree, and RandomForest. Several criteria have been used for evaluation those three models. We will use accuracy, (AUC), F1 score etc. as key indicators for our models selection. Based on our model comparison RandomForest 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. With a 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.

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

How to do Effective and Sucessful Bank Telemarketing

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 brand and value.

Therefore, banking companies started working on addressing the 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 identifications of our target customers, customers that are most likely to respond to our bank telemarketing complain, 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.

Literature Review References: - 1) Who Will Subscribe A Term Deposit? Jiong Chen (jc4133), Yucen Han (yh2645), Zhao Hu (zh2210), Yicheng Lu (yl3071), Mengni Sun (ms4783) - 2) Predictive Modeling to Improve Success Rate of Bank Direct Marketing Campaign - Vaidehi R - 3) A Data Mining Approach for Bank Telemarketing Using the rminer Package and R Tool - Sergio Moro, Paulo Cortez , Raul M. S. Laureano

Methodology CRISP-DM:

In this project we will be using 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.

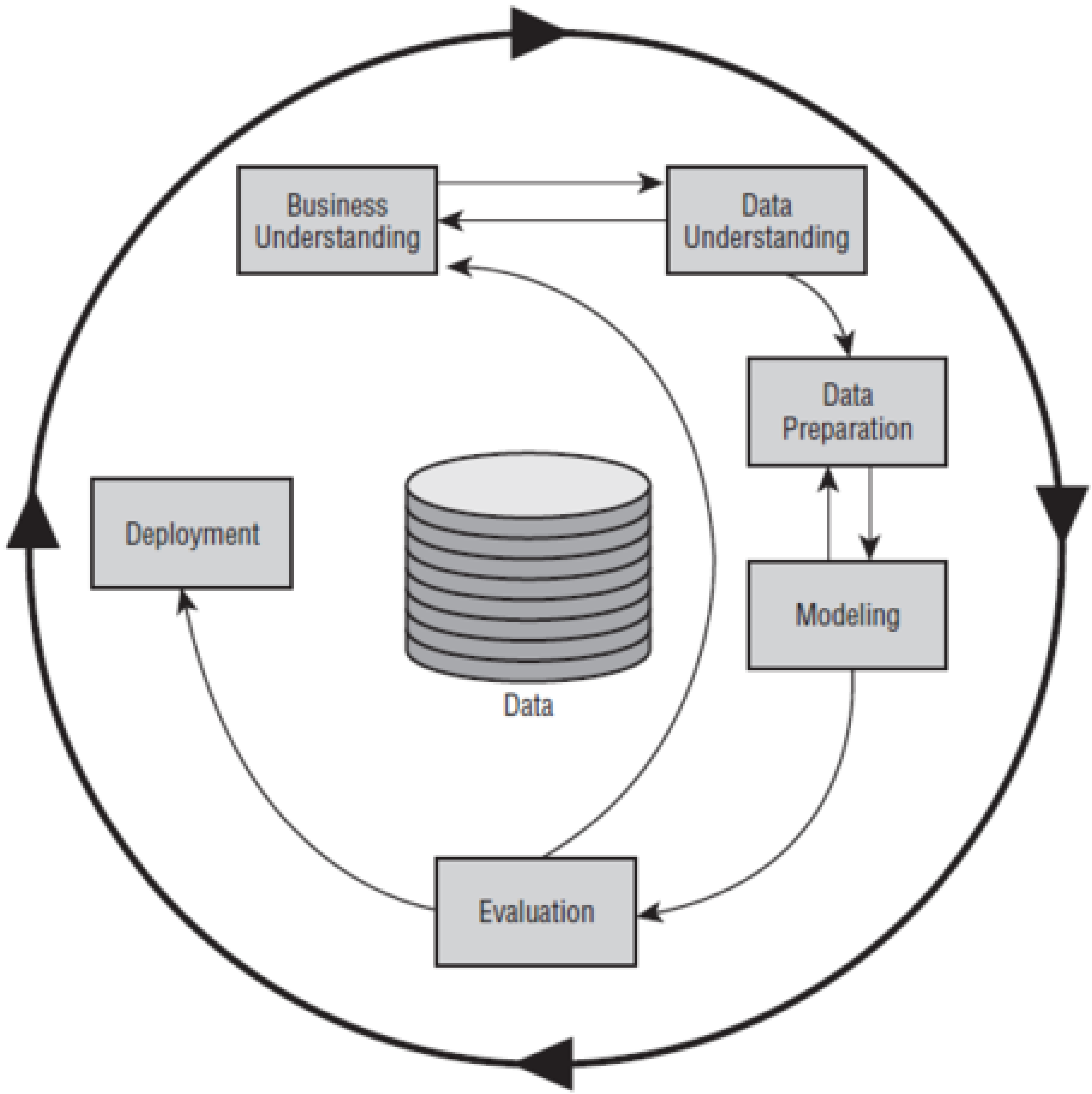


Figure 1. CRISP-DM

The only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. A review and critique of data mining process models in 2009 called the CRISP-DM the"de facto standard for developing data mining and knowledge discovery projects."[6] Other reviews of CRISP-DM and data mining process models include Kurgan and Musilek's 2006 review, and Azevedo and Santos' 2008 comparison of CRISP-DM and SEMMA. Efforts to update the methodology started in 2006, but have As of 30 June 2015 not led to a new version, and the"Special Interest Group" (SIG) responsible along with the website has long disappeared."

Business Understanding

The data is available on website for UC Irvine Machine Learning Repository. There are two different data sets available. The "bank" data has 45,211 records with 16 attributes and 1 response variable. The "bank-additional" data has 41,188 records with additional attributes added to "bank" data, it has 20 attributes and 1 response variable. We chose to use the data with additional attributes.

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

Data Exploration

We used exploratory graphs, Predictor and Response variable Association, count of y by each variable to explore the data. We were able to see using these graphs how the variables were making an impact on y.

Data Preparation

We prepared the data by first changing the response variable from yes and no to 1 and 0. The other variables that were basically binary but had and "unknown" value, we treated it just like a categorical variable and created 3 dummy variables. The categorical variable had dummy variables created to accomodate for all values in the variable including the "unknown" which are the missing variables in the data. we did not omit the missing records but incorporated in our data to see if there was value in it.

Modeling:

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 RandomForest model to overcome that.

RandomForest Model. Random Forests 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

There are number of ways to evaluate the regression and classification models based on the purpose like prediction, classification, variable selection etc. In the given business scenario objective is to classification of the response variable by building a model that can predict likelihood of response from Customer. Following evaluation criteria we have used for model evaluation:

(1) 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.

(2) AUC along with Model Accuracy will be used for model evaluation. Accuracy is calculated based on certain threshold where as AUC is overall performance evaluation of model as various points.AUC criteria will be given more weight age for model evaluation in this case.

Experimentations and Results:

Data Exploration

The data is available on website for UC Irvine Machine Learning Repository. There are two different data sets available. The “bank” data has 45,211 records with 16 attributes and 1 response variable. The “bank-additional” data has 41,188 records with additional attributes added to “bank” data, it has 20 attributes and 1 response variable. We chose to use the data with additional attributes.

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:

1 - Binary values of “yes” and “no” wit null values given as “unknown”.

2 - Categorical values with “unknown” as missing values. The categorical variable require dummy variables to be created for each unique value. We included “unknown” as one of the dummy variable.

3 - 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

We notice that the variables are numerical, categorical and binary. The response variable y is binary.

Based on the original dataset, our predictor input has 21 variables. And our response variable is 1 variable called y.

Binomial Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more metric (interval or ratio scale) independent variables.

Table 1

Variable Analysis

Variable	Data.Type	Analysis
age	Numeric	No significant trend with responses variable, better response with age grp<3
job	Catagorical	12 levels, proportion of responses from admin and blue collar job profiles ar
marital	Catagorical	4 levels, % response from marital status from single is greater compare to o
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

Variable	Data.Type	Analysis
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 respons
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 varia
previous	Numeric	previous contacts seems to have influence on the positive response of the ca
poutcome	Catagorical	have relationship with campaign outcome, earlier success has better respons
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
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

Data Preparation

-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

Prepare test data. We will treat the test data the same way as the train data, and then apply models on the test data created using the train data.

Analysis the link function for given variables

In this section, we will investigate how our 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).

Now let's investigate how our initial data model aligns with the above criteria. In other words, we will plot regression model plots for each variable and compare it to a typical logistic model plot:

The main objective in the transformations is to achieve linear relationships with the dependent variable or, really, with its logit.

Model Building

In this section experimentation will be carried out with the data by formulating three different types of models with three different approaches. 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.

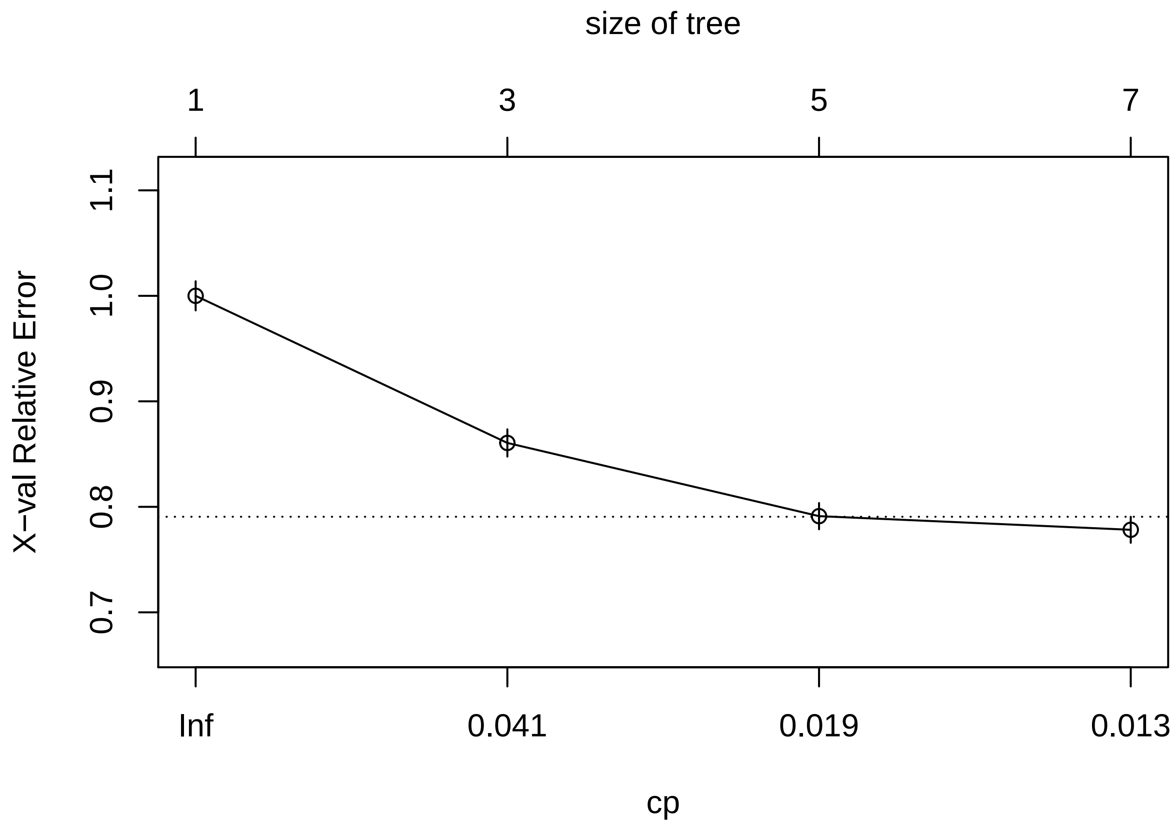
-Model 3- This model will be created by using classification technique RandomForests model.

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.

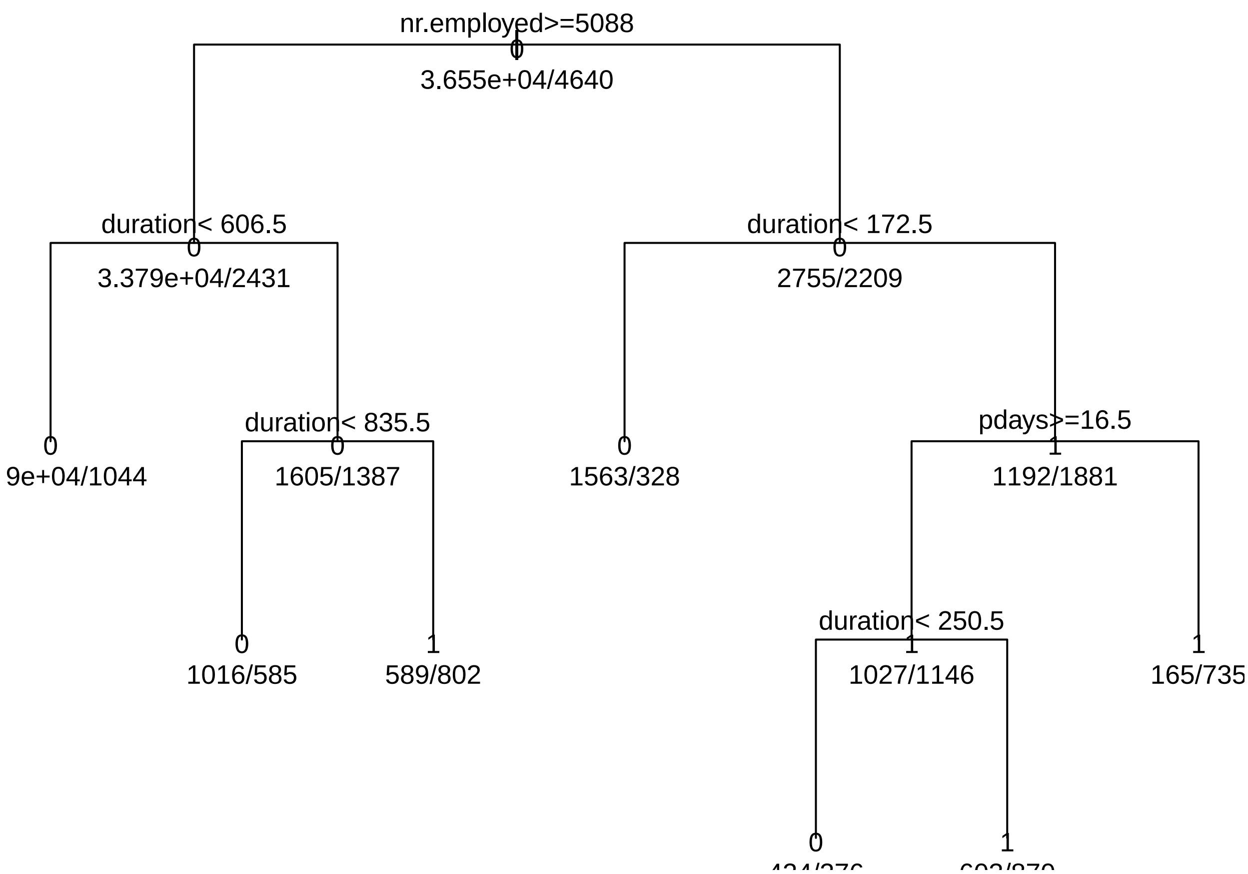
Logistics Regression- Model 1. Logistics regression function GLM has been used to classify the campaign response variable. Basic model generated by using GLM function has been enhanced by making necessary adjustments to non associated predictor variables shown as “NA” in basic model output. Next the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

There were 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.

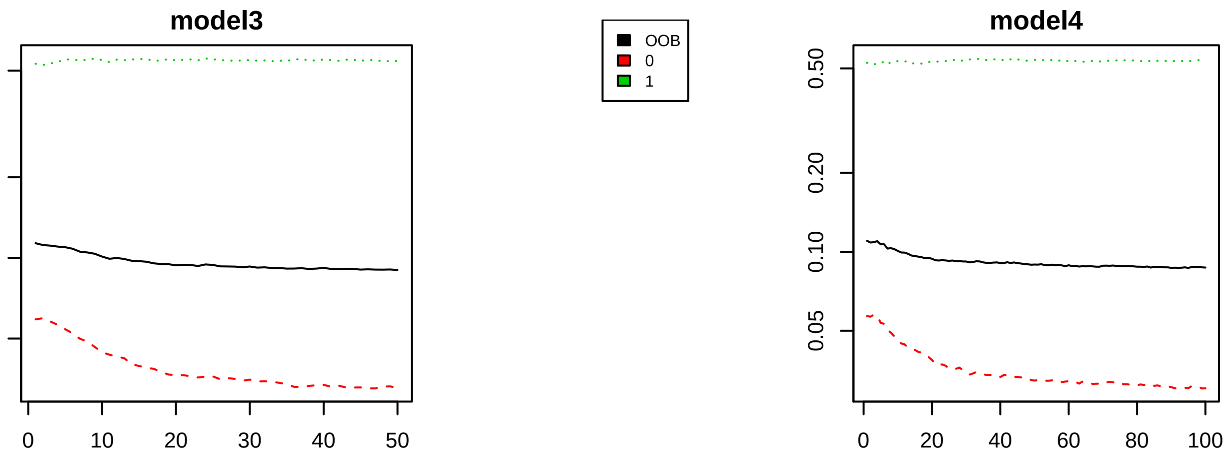
Classification Tree- Model 2. The basic idea of classification tree model is to predict a response variable y for the campaign from predictor variables. Model does this by growing a binary tree. At each node in the tree, a test is applied to one of the inputs. Depending on the outcome of the test two routes to be followed left or right. 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.



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

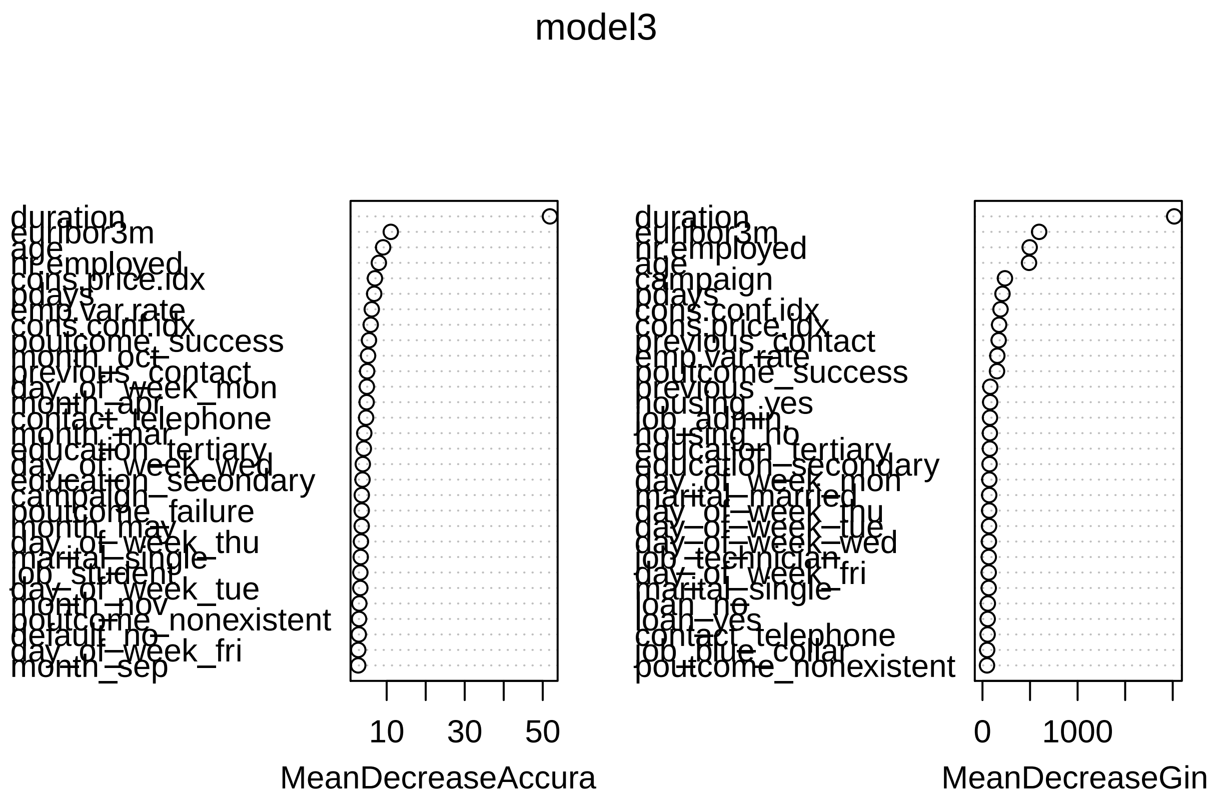


RandomForest- Model 3. In Random Forests many classification trees are formed to classify campaign response variable y. Each tree creates separate set of classification, each tree is voted for performance for that classification. The forest chooses the classification having the most votes (over all the 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 the chart above it can be seen that classification error rate to classify negative responses reduces with the increase in number of trees but there is no significant change in error rate for positive response.There is only slight reduction in error rate for negative responses when tree size is increased to 100 from 50.Number of variables tried at each split are 7 with negative classification rate of 0.03 and positive classification error rate of 0.51.

Below chart provides importance of various variables used in the model.



Results from Models:

Results from Regression Model: Logistics Regression model has a very high accuracy rate of 91.42% when model was evaluated using the validation data set. Though the AUC value for this model was comparatively lower 0.702 which indicates not good fitment 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. Regression model passed this test.

```
Hosmer and Lemeshow goodness of fit (GOF) test
```

data: model1_update$y, fitted(m) X-squared = 14.926, df = 8, p-value = 0.0606

Results from Classification Tree Model-this model has also very high accuracy rate of 91.81% which is very good.This model has AUC value of 0.865 which seem to be inline with given high accuracy.

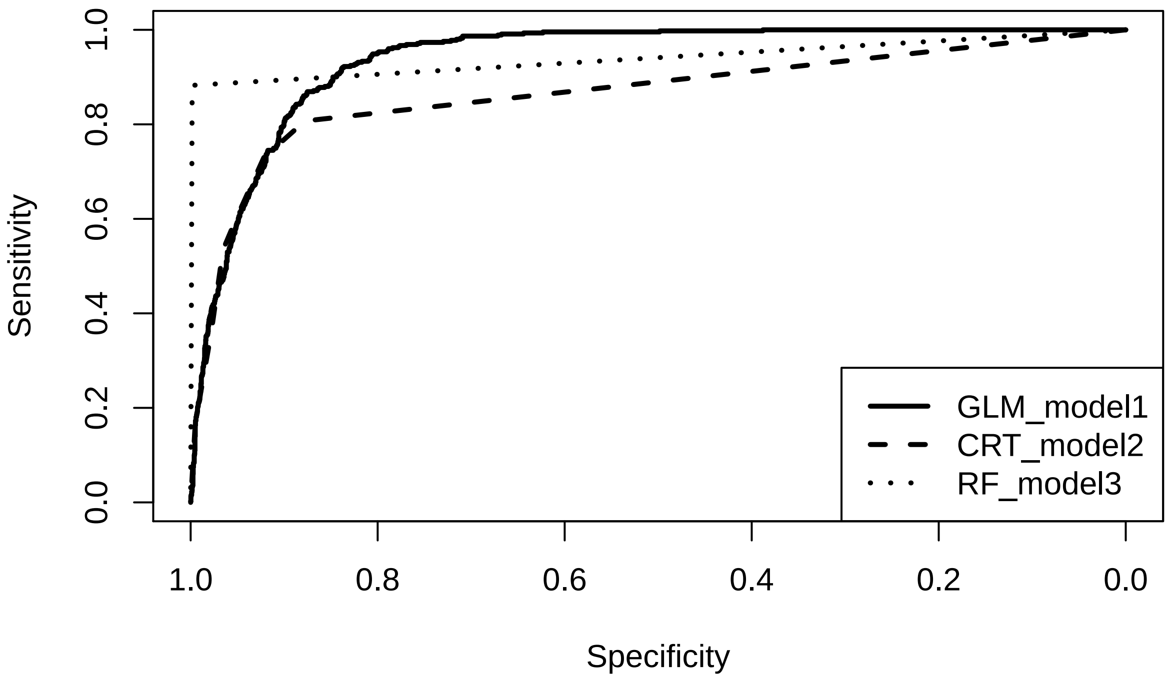
Results from RandomForest Model-The model created using Randomforest has accuracy of 98.64% which is extraordinary results and give rise to suspicion model is able to separate out the classification based on certain variable. When we looked at the importance of variable “duration” it becomes apparent that this variable is being used in a big way to classify response accurately. It can be seen that this model also shows the 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 2

Comparison of 3 Model3

	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	0.9856761	0.0143239	0.8824834	0.9851485	0.9857335	0.8822214	0.9404238



- 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%. But Accuracy is not always the key criteria for a model as Accuracy is calculated based on a defined threshold. Also due to imbalance of data o 10% to 90% distribution of response variable forced to choose the model based on other criteria. 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 most important variables by far. This variable has positive impact in campaign outcome. This could be due to the fact that longer the Customer stays on phone more productive conversation is taking place to get the Customer start their term deposit Account. "euribor3m" is most important variable which denotes inter bank interest rate in Eurozone. Term deposit interest rates are generally interlinked and tends to go up together. This variable has positive impact on response variable. Predictor "nr.employed" denotes number of employees for the bank. This variable also has positive impact on campaign response. More the number of employees 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 response. As 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 that time.

- Shortcomings :

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

- Final Recommendation :

In conclusion it can be suggested to the bank management that focus should be given in hiring more people, doing more quality phone calls. Also to time the campaign in a stable macroeconomic environment to get better return on investment from this campaign.

References

be sure to cite all references used in the report (APA format). We used R (3.2.5, R Core Team, 2016) and the R-packages *papaja* (0.1.0.9054, Aust & Barth, 2015), *papaja* (0.1.0.9054, Aust & Barth, 2015), *Amelia* (1.7.4, Honaker, King, & Blackwell, 2011), *aod* (1.3, Lesnoff, M., Lancelot, & R., 2012), *AUC* (0.3.0, Ballings & Poel, 2013), *dplyr* (0.4.3, H. Wickham & Francois, 2015), *faraway* (1.0.7, Faraway, 2016), *gdata* (2.17.0, Warnes et al., 2015), *ggplot2* (2.1.0, H. Wickham, 2009), *gplots* (3.0.1, Warnes et al., 2016), *gridExtra*

(2.2.1, Auguie, 2016), *ISLR* (1.0, James, Witten, Hastie, & Tibshirani, 2013), *knitr* (1.12, Xie, 2015), *leaps* (2.9, Fortran code by Alan Miller, 2009), *MASS* (7.3.45, W. N. Venables & Ripley, 2002), *popbio* (2.4.3, Stubben & Milligan, 2007), *psych* (1.6.4, Revelle, 2016), *Rcpp* (0.12.3, Eddelbuettel & François, 2011), *reshape* (0.8.5, Wickham & Hadley, 2007), *ROCR* (1.0.7, Sing, Sander, Beerenwinkel, & Lengauer, 2005), *stringr* (1.0.0, H. Wickham, 2015), *xtable* (1.8.2, Dahl, 2016), *lattice* (0.20.33, Sarkar, 2008), *pscl* (1.4.9, Zeileis, Kleiber, & Jackman, 2008), *randomForest* (4.6.12, A. Liaw & Wiener, 2002), *rpart* (4.1.10, Therneau, Atkinson, & Ripley, 2015), *boot* (1.3.18, Davison & Hinkley, 1997), and *ResourceSelection* (0.2.6, Lele, Keim, & Solymos, 2016) for all our analyses.

Appendix

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

```
{r
code=readLines(knitr::purl('https://raw.githubusercontent.com/kishkp/data621-ctg5/mast
documentation = 0)), eval = FALSE} #
```

Data Analysis details

Variable Description

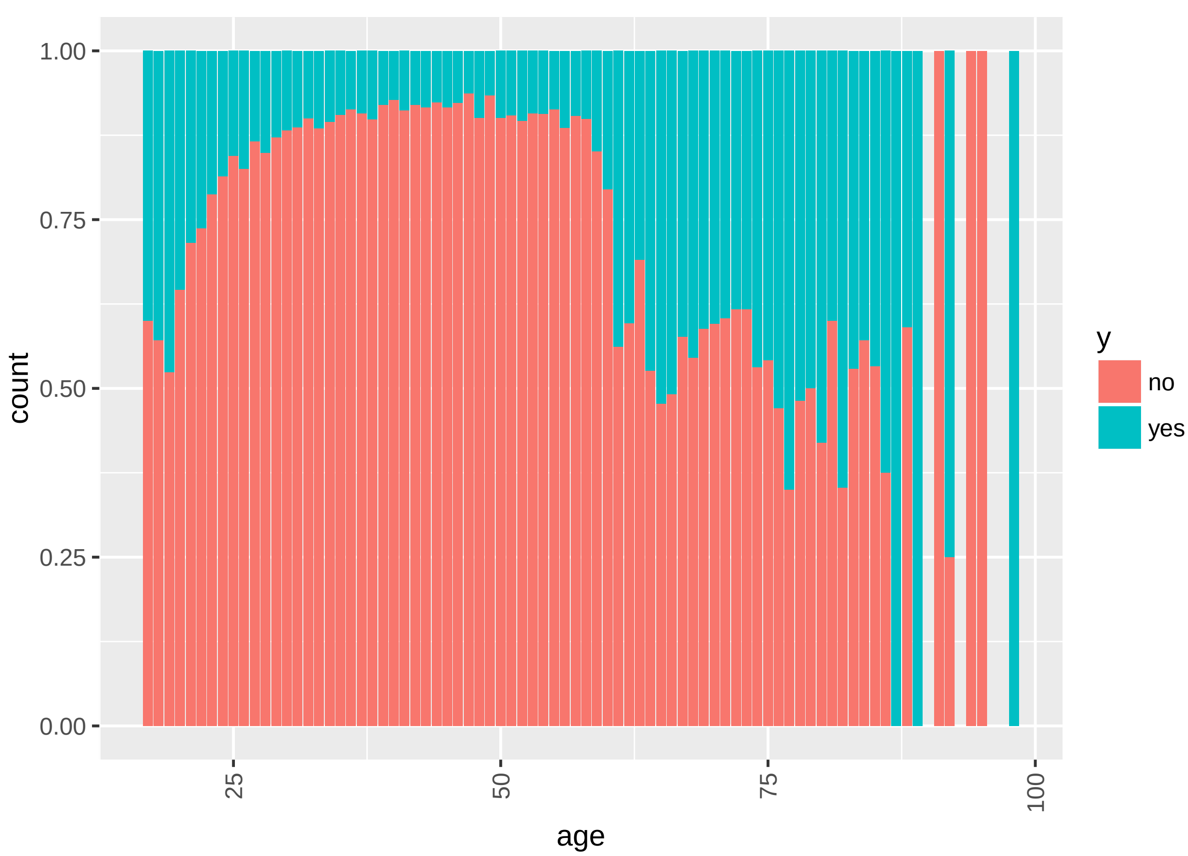
Table 3

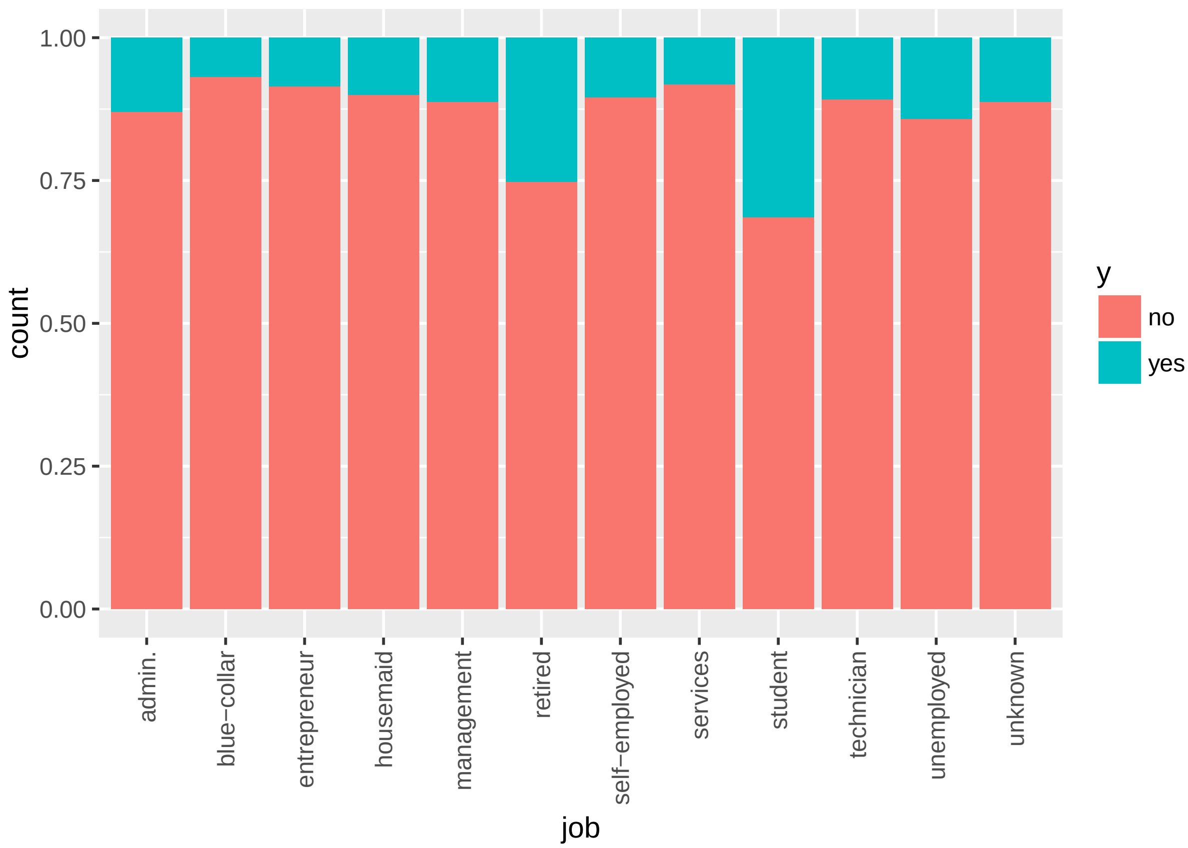
Variable Description

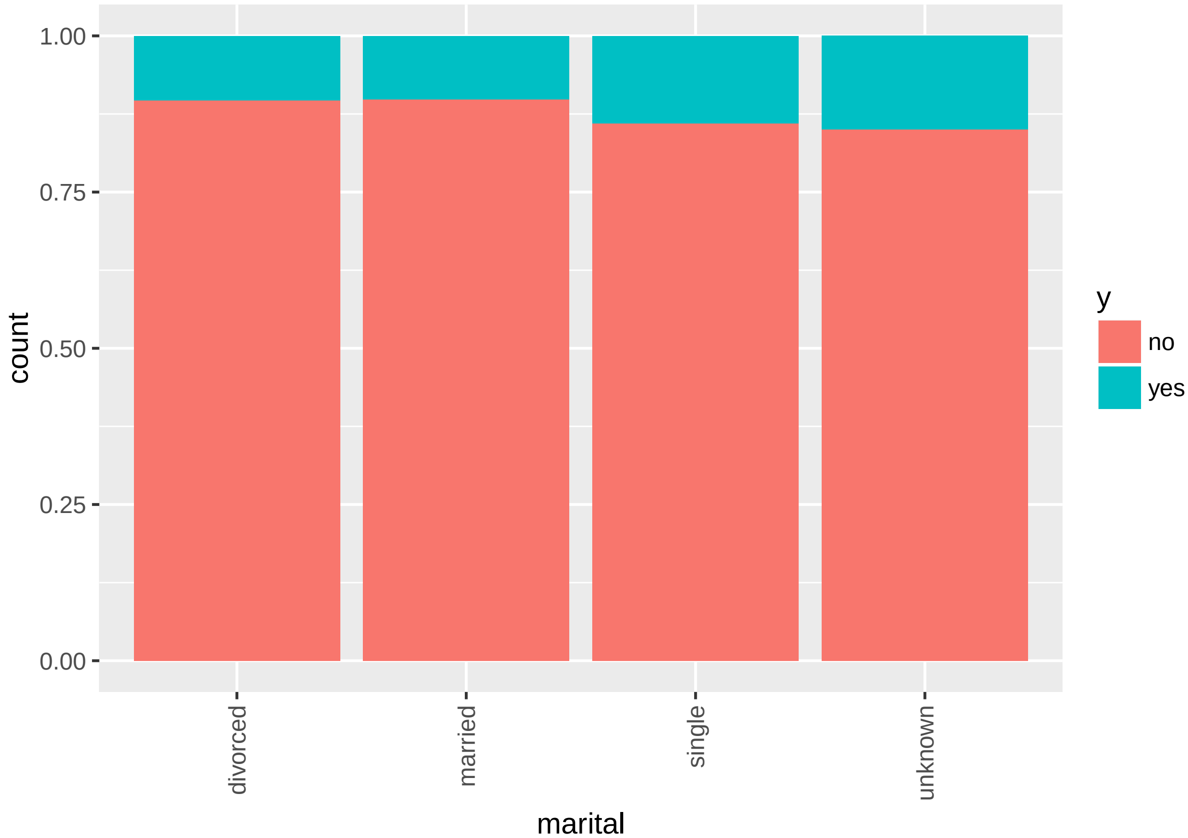
Variable	Data.Type	Type	Description
age	Numeric	Predictor	Client's age
job	Catagorical	Predictor	Client's job
marital	Catagorical	Predictor	Client's marital status
education	Catagorical	Predictor	Client's education level
default	Binary	Predictor	Credit in default?

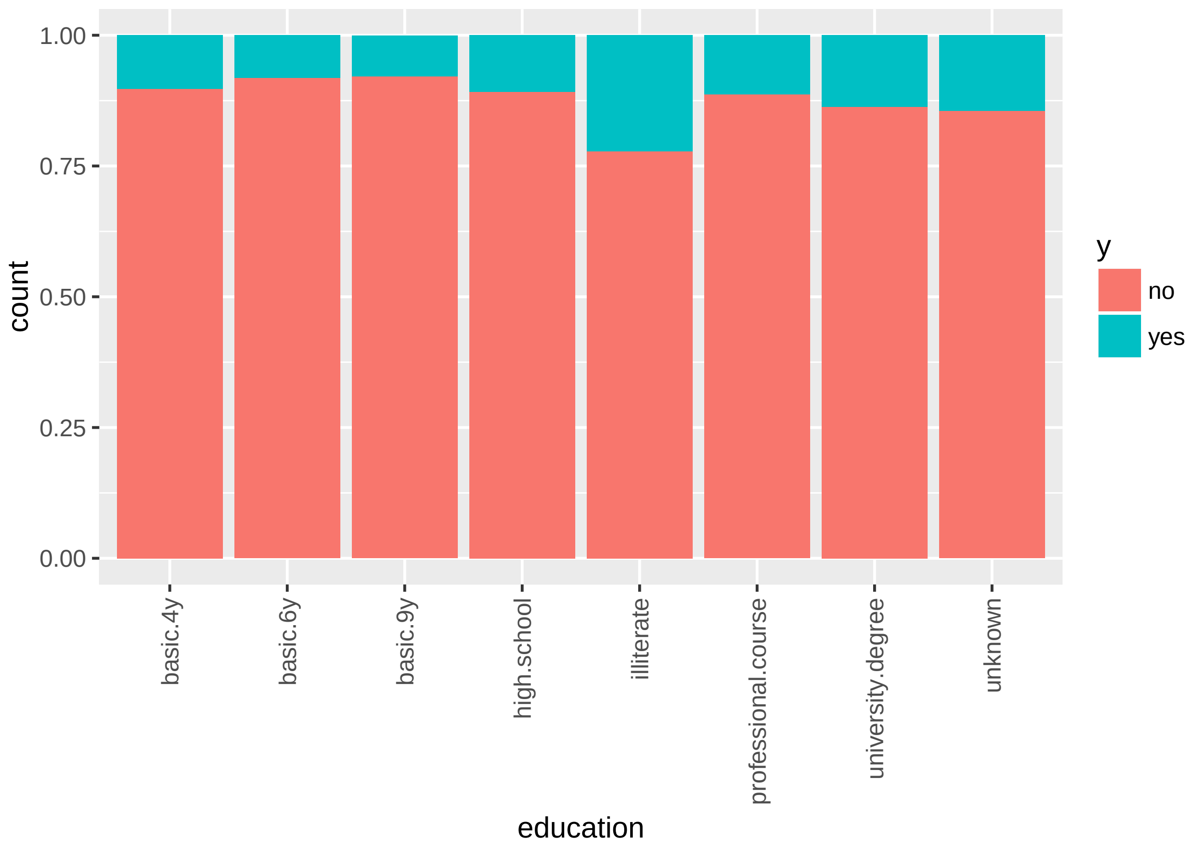
Variable	Data.Type	Type	Description
balance	Numeric	Predictor	Client's average yearly balance, in euros
housing	Binary	Predictor	Client has housing loan?
loan	Binary	Predictor	Client has personal loan?
contact	Catagorical	Predictor	Client's contact communication type
day	Catagorical	Predictor	Client 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
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?

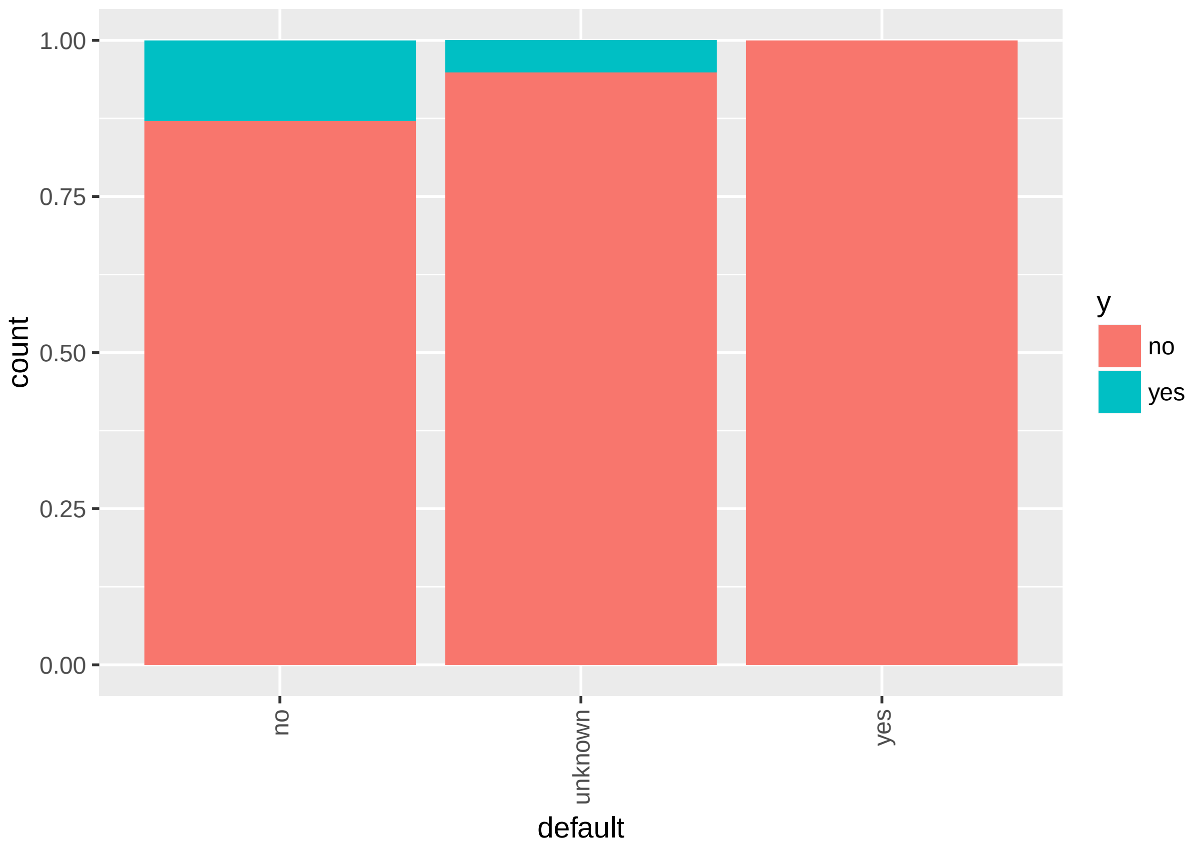
Predictor and Response variable Association

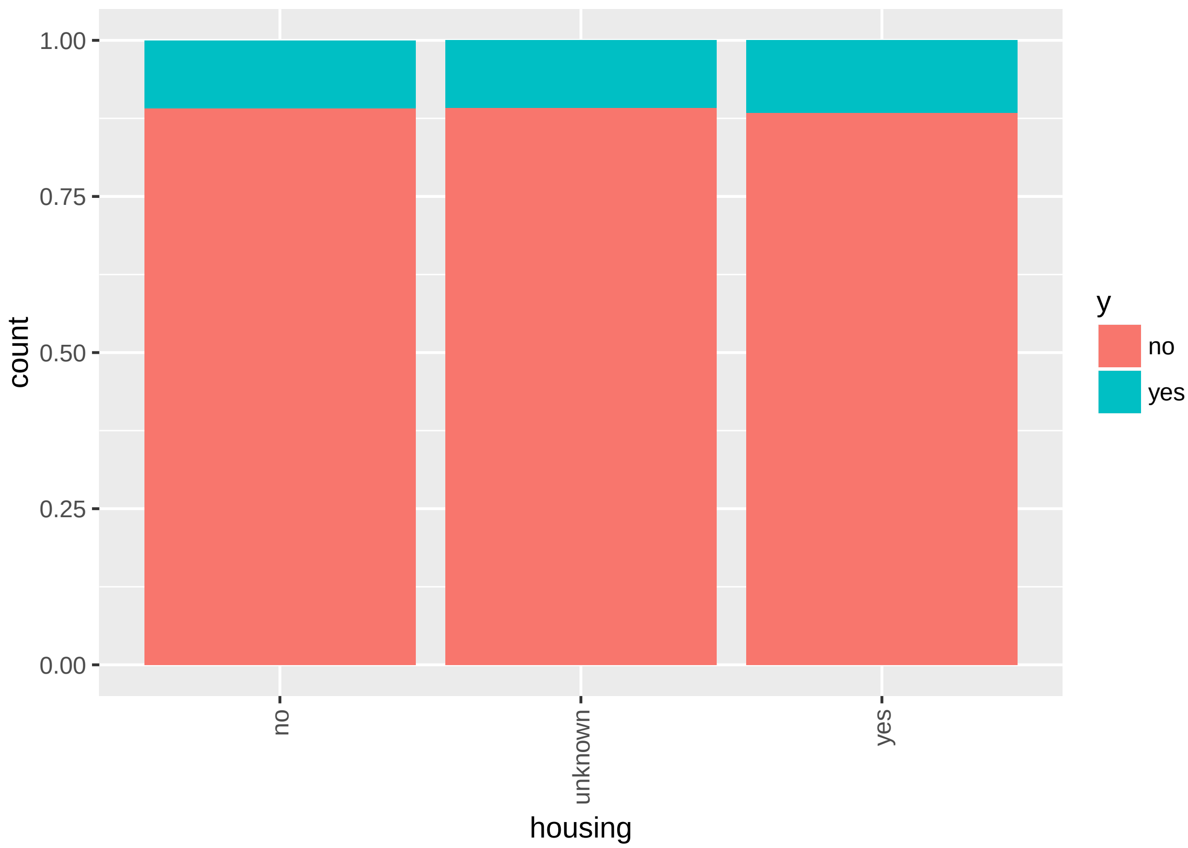


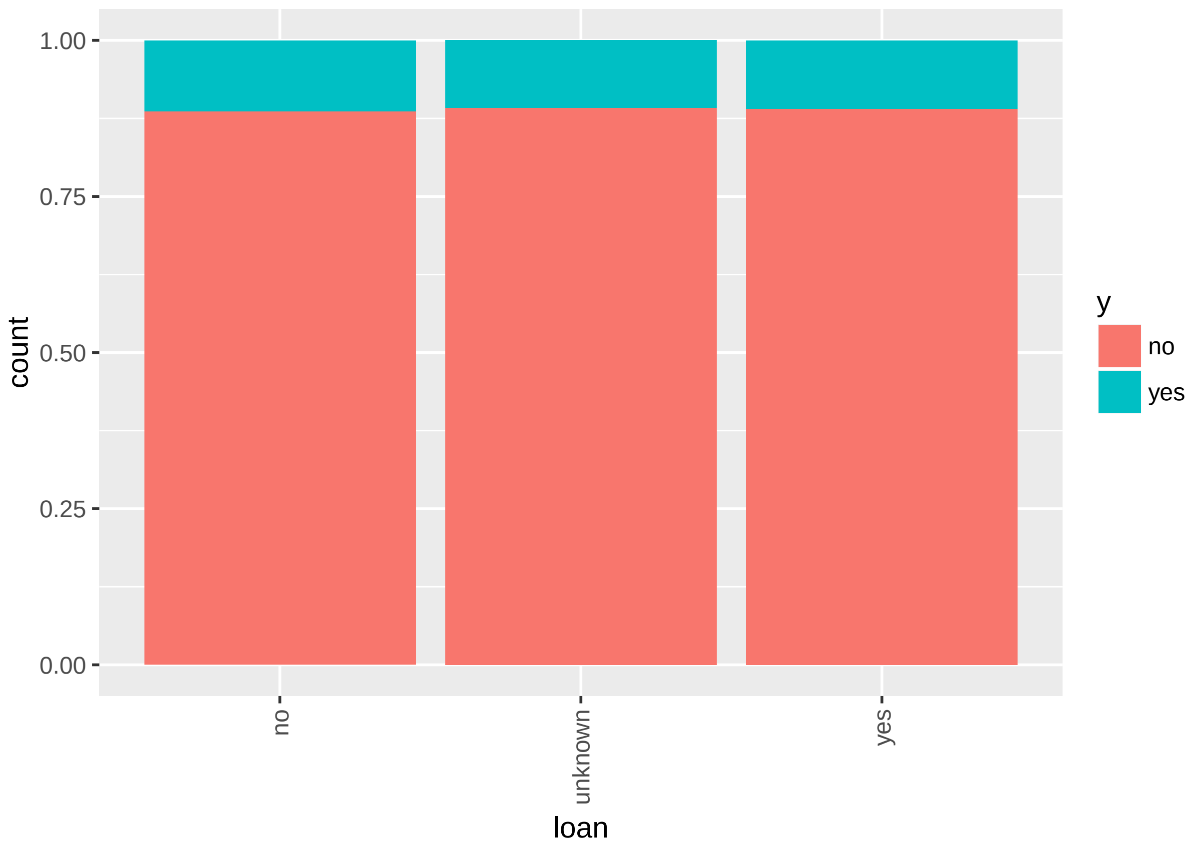

count
1.00
0.75
0.50
0.25
0.00
admin.
blue−collar
entrepreneur
housemaid
management
retired
self−employed
services
student
technician
unemployed
unknown
job
y
no
yes

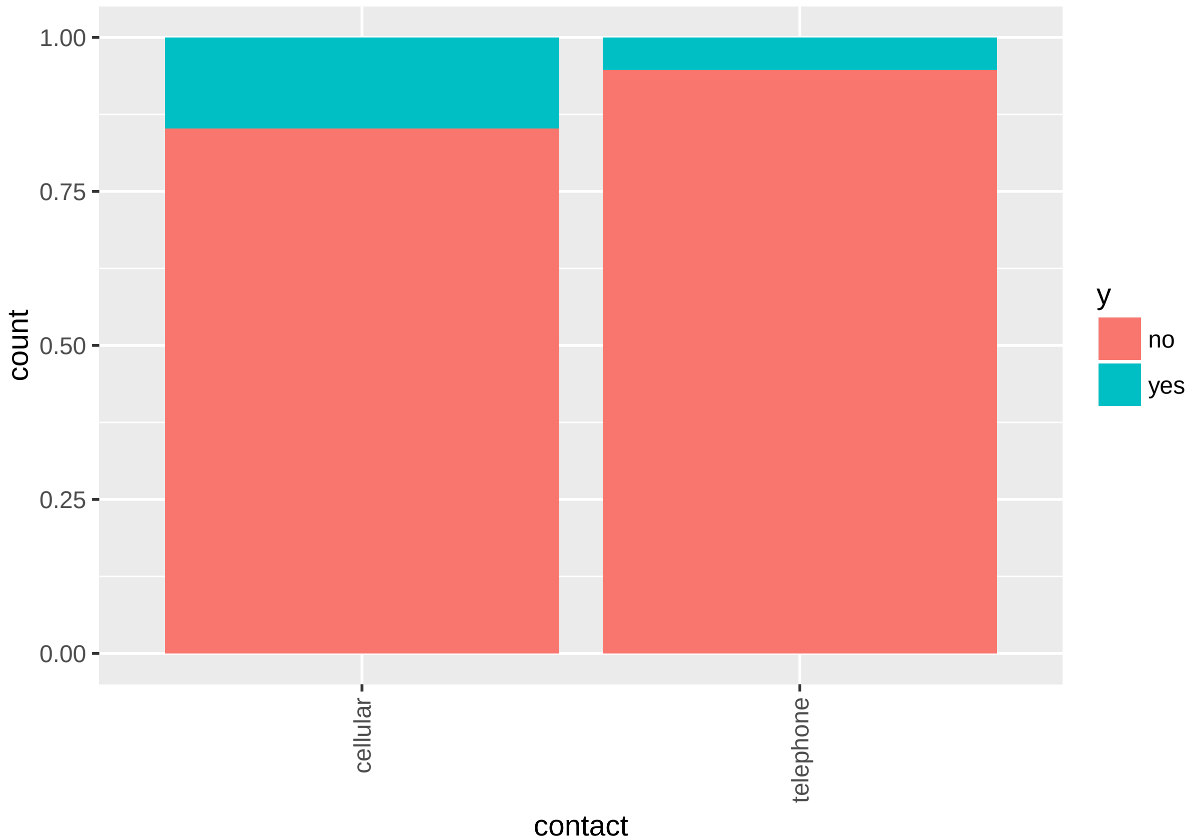

1.00
0.75
0.50
0.25
0.00
count
divorced
married
single
unknown
marital
y
no
yes


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

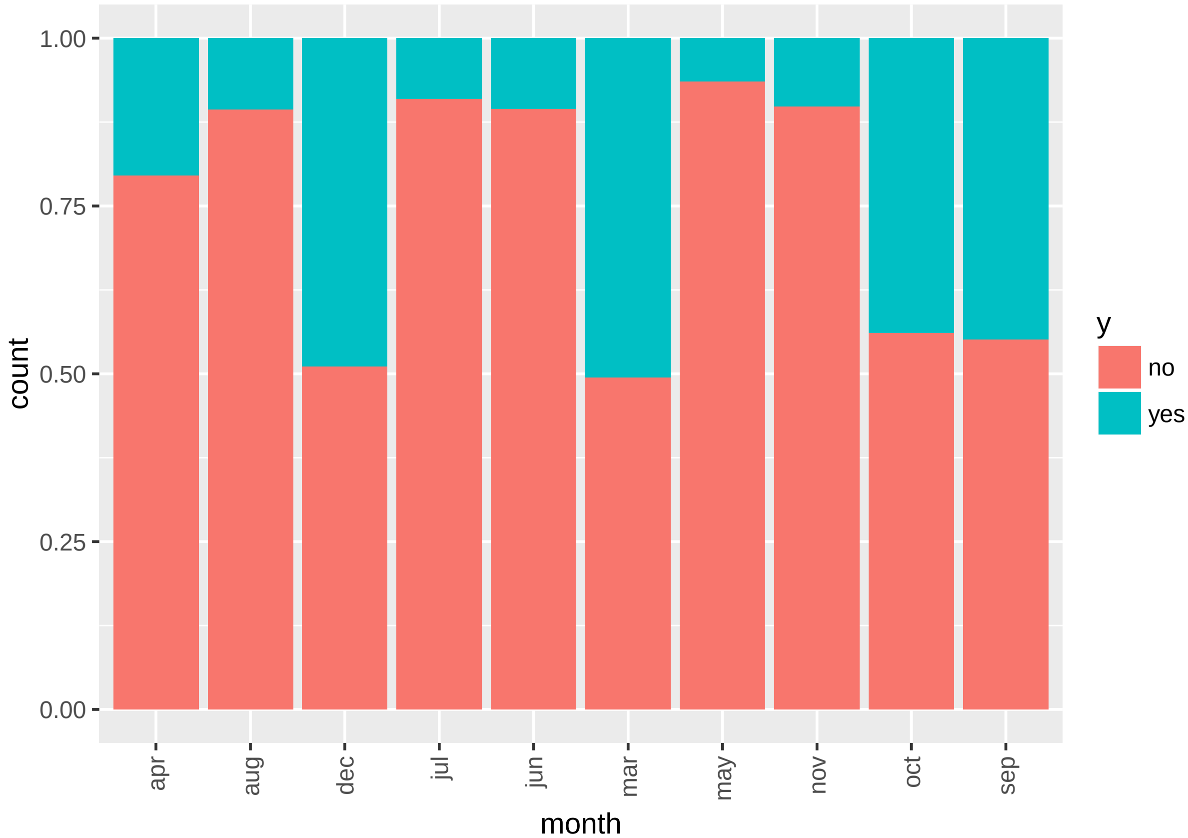

1.00
0.75
0.50
0.25
0.00
count
no
unknown
yes
default
y
no
yes

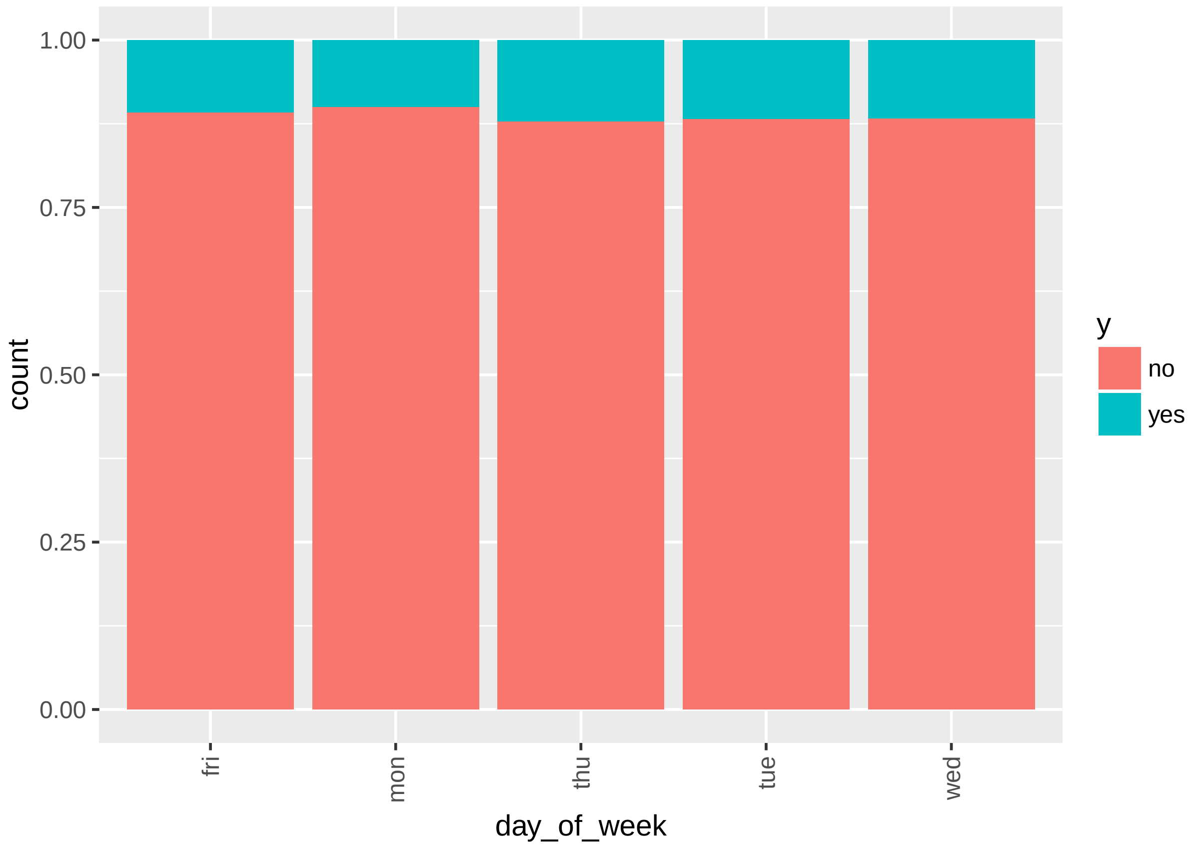

1.00
0.75
0.50
0.25
0.00
count
no
unknown
yes
housing
y
no
yes

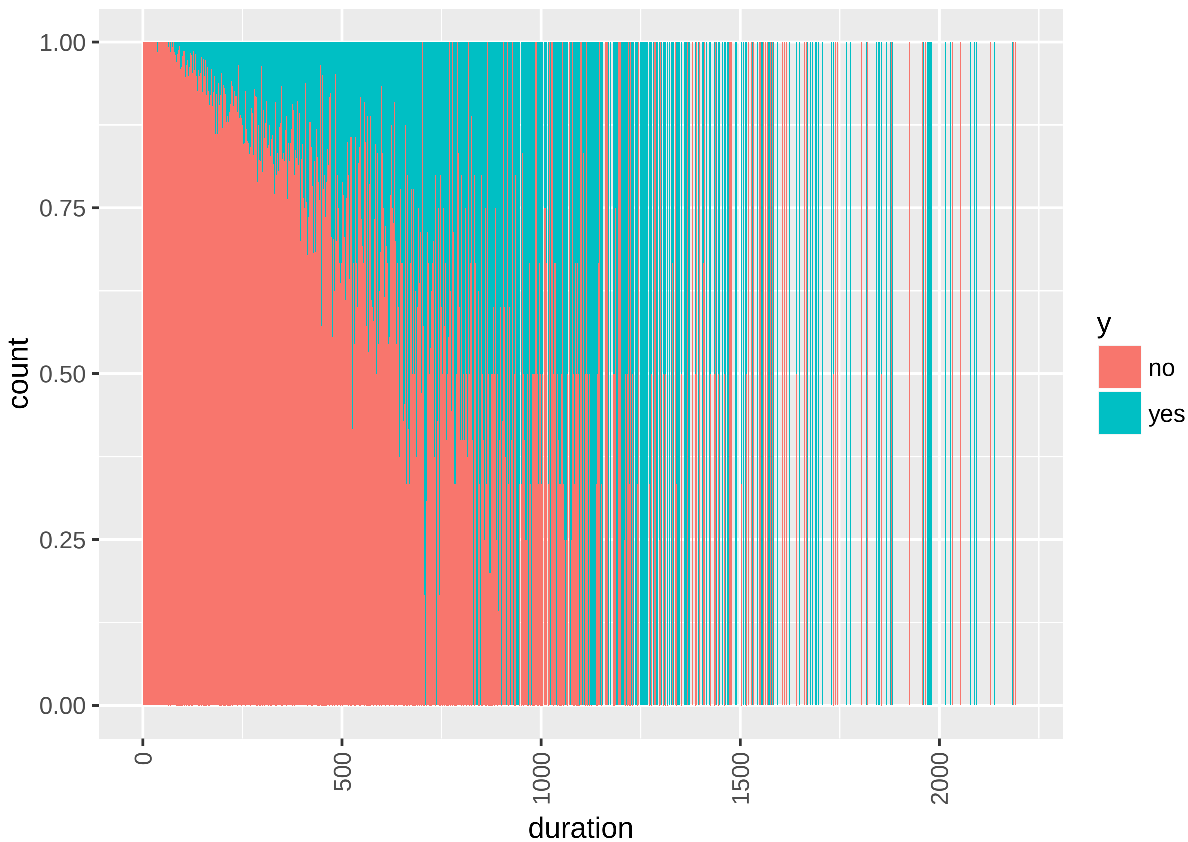

1.00
0.75
0.50
0.25
0.00
count
no
unknown
yes
loan
y
no
yes


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

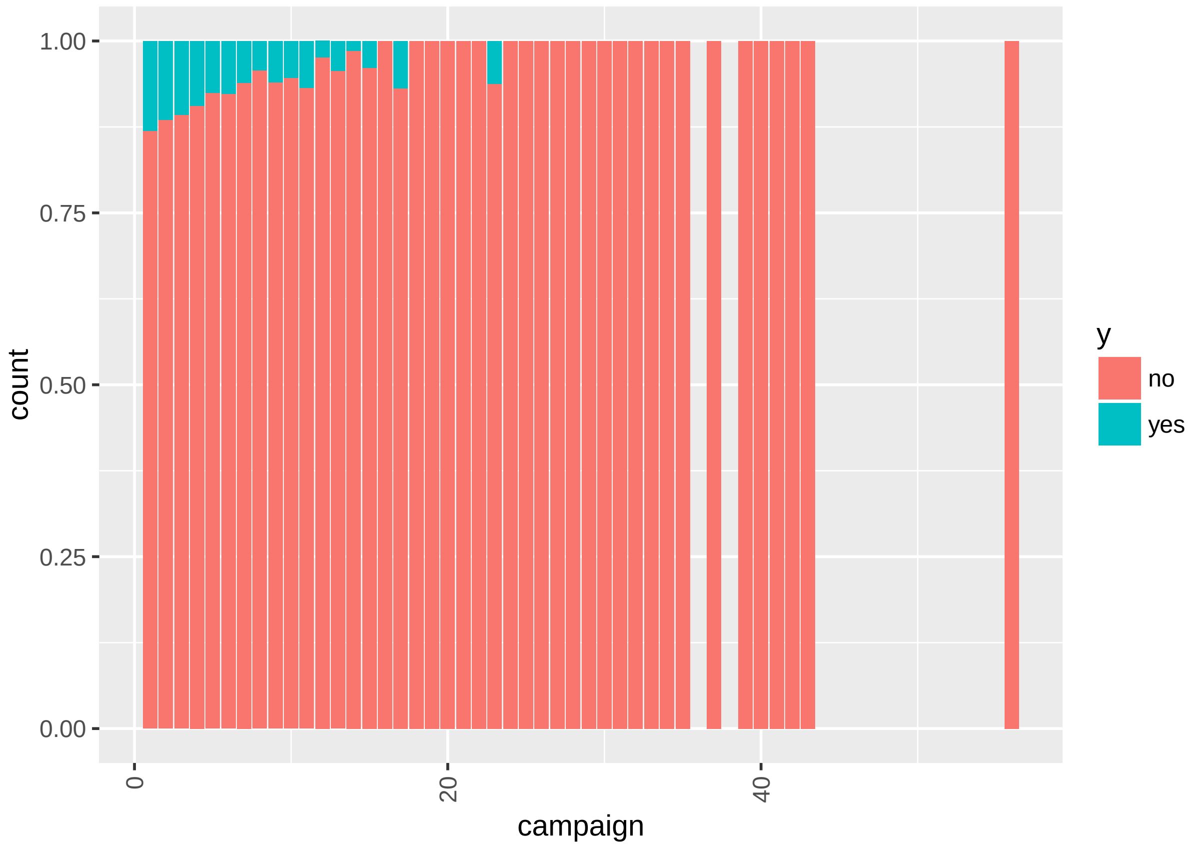
331


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

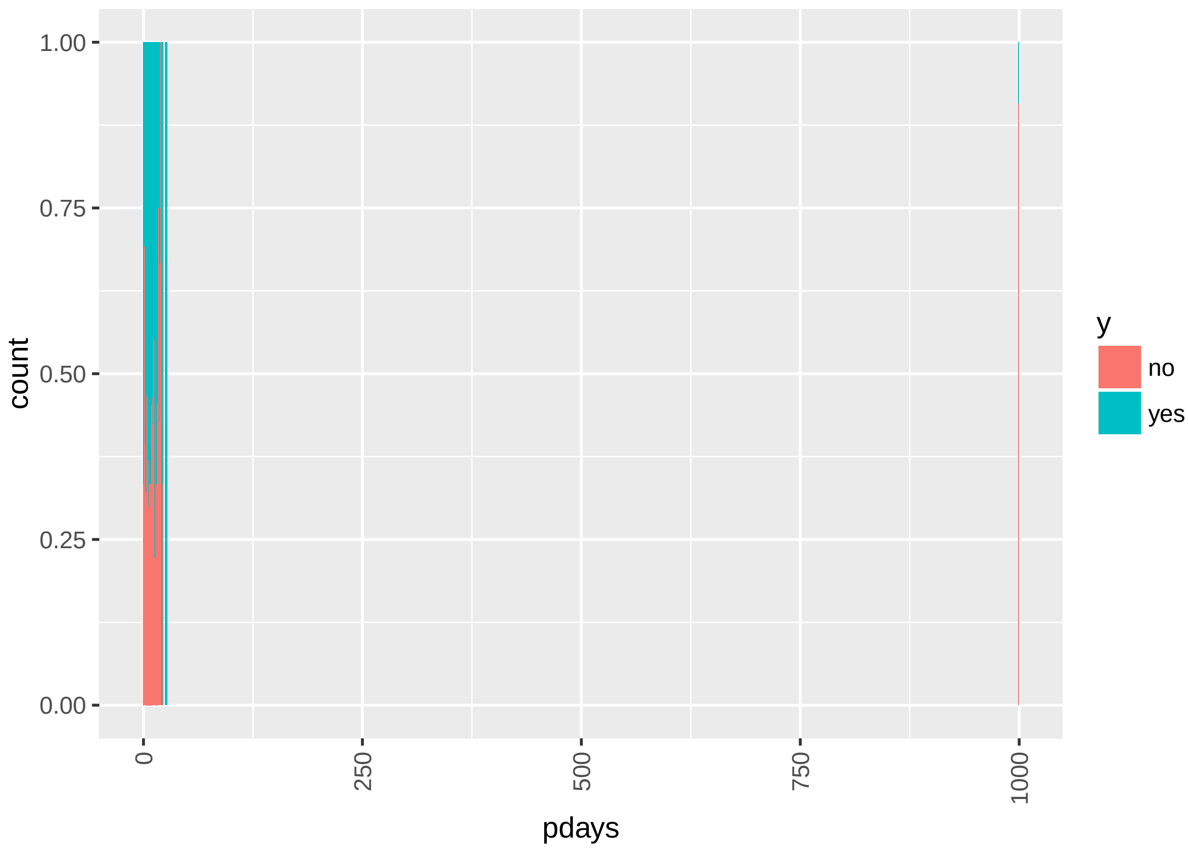

1.00
0.75
0.50
0.25
0.00
count
fri
mon
thu
tue
wed
day_of_week
y
no
yes


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

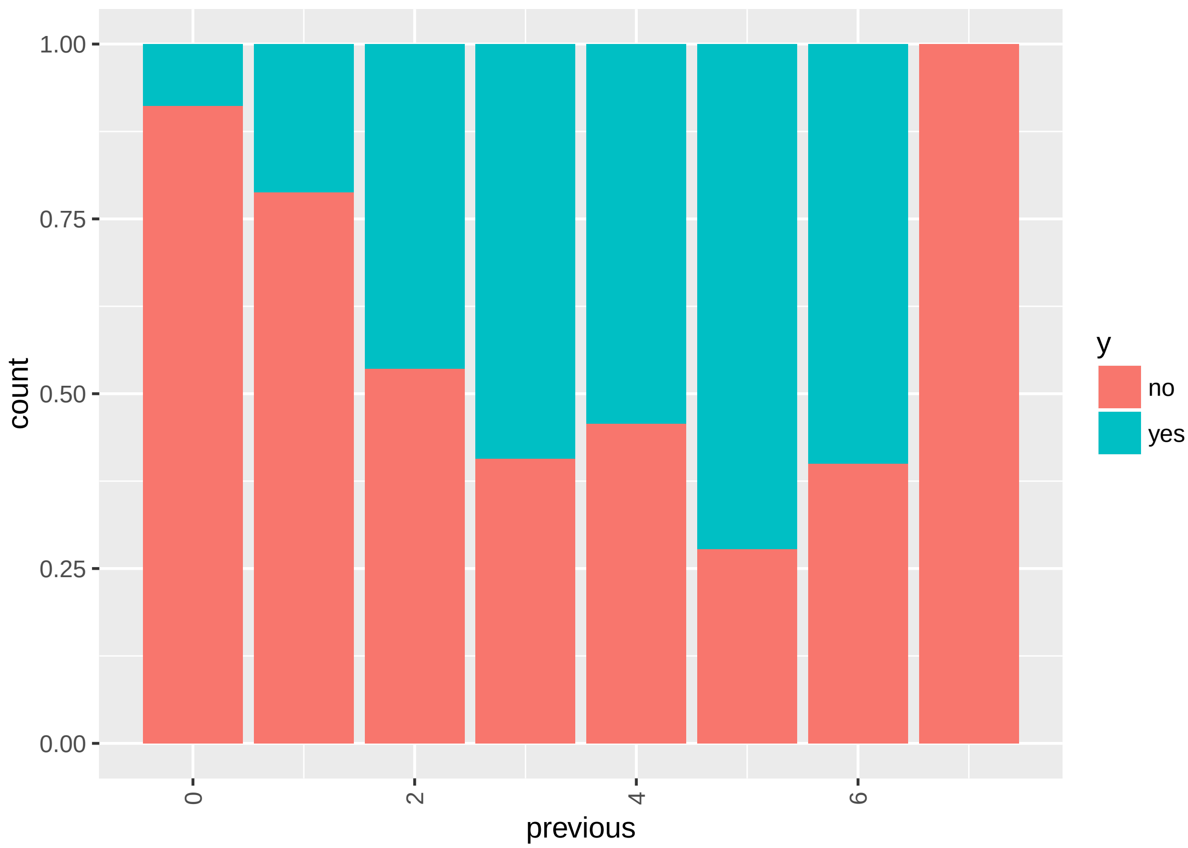
334

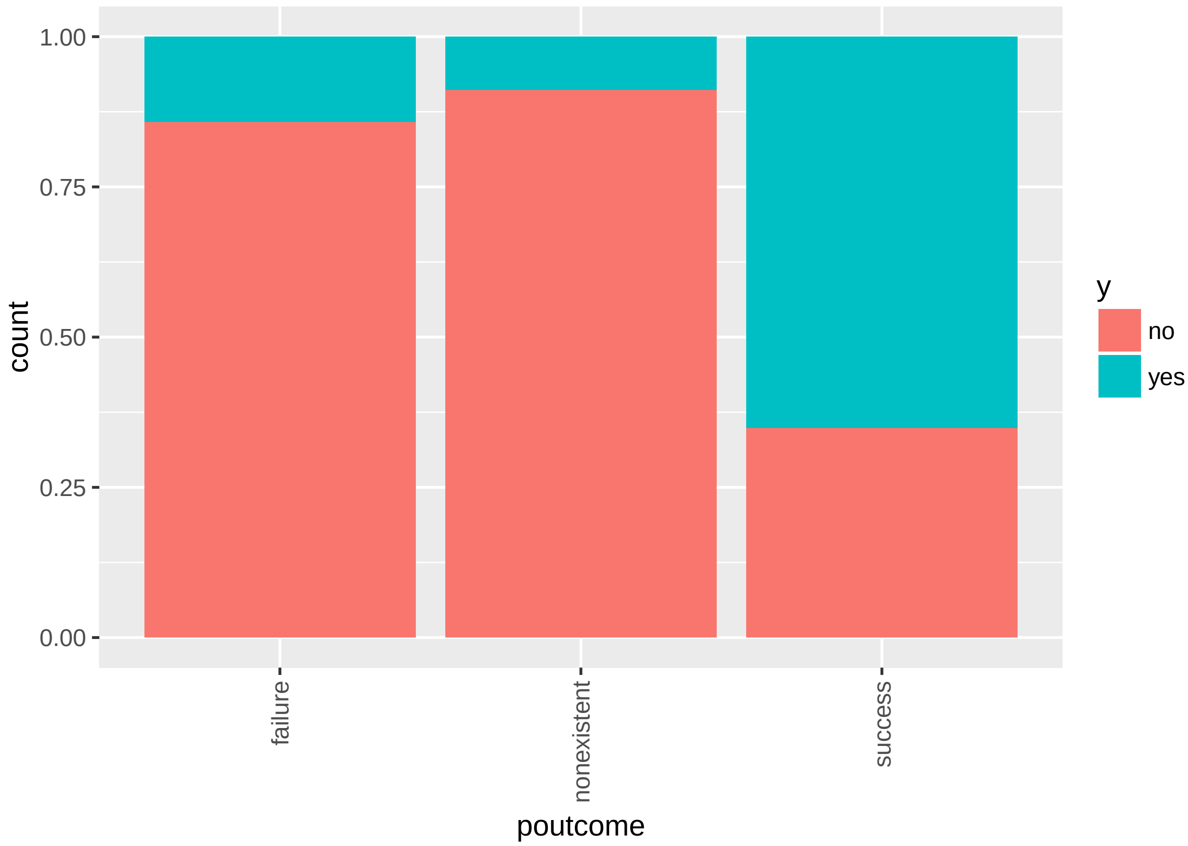


335

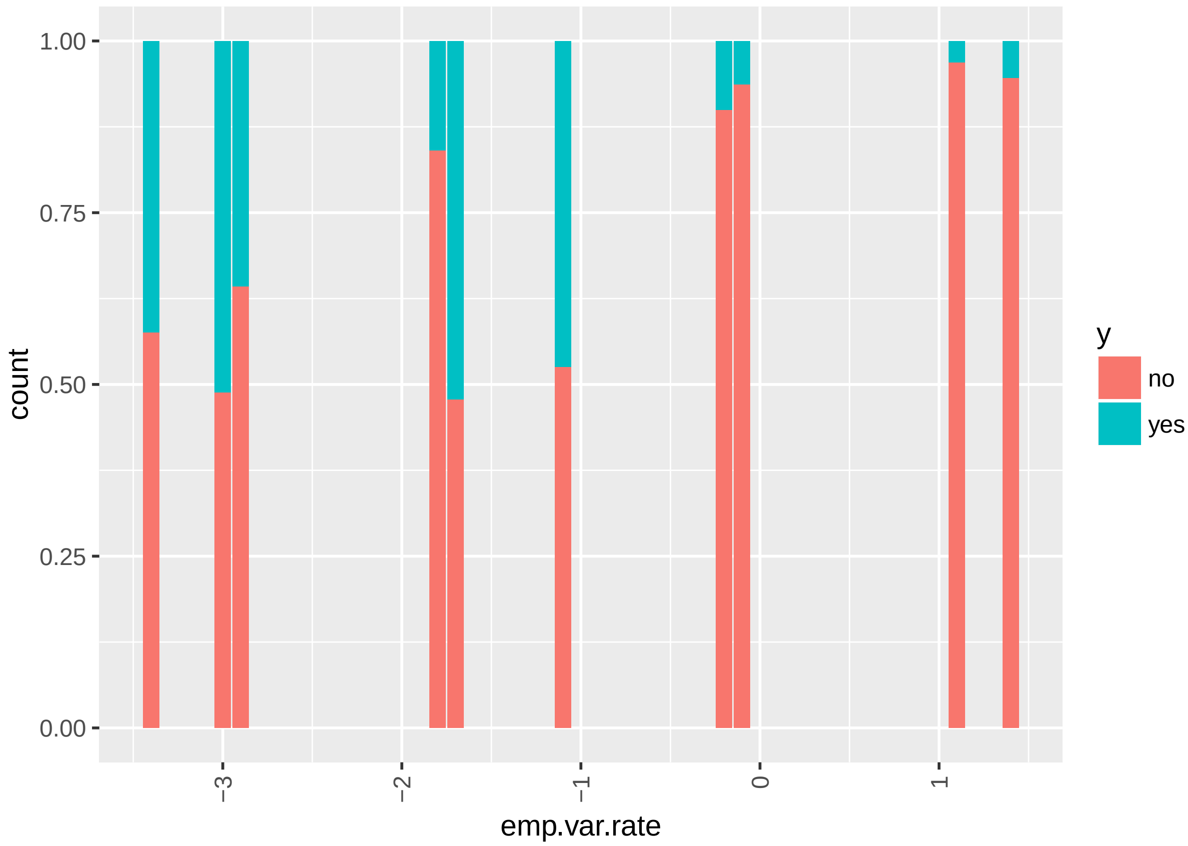


336

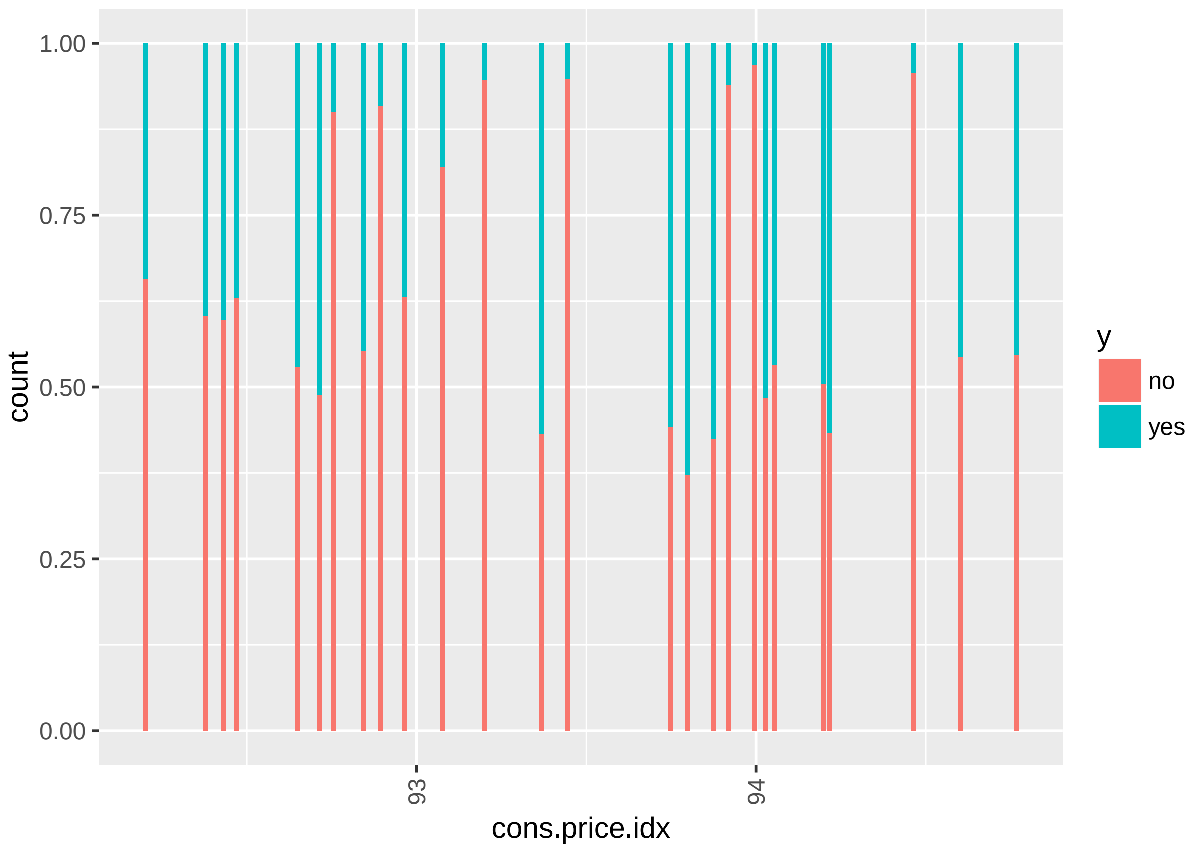

1.00
0.75
0.50
0.25
0.00
count
0
2
4
6
previous
y
no
yes

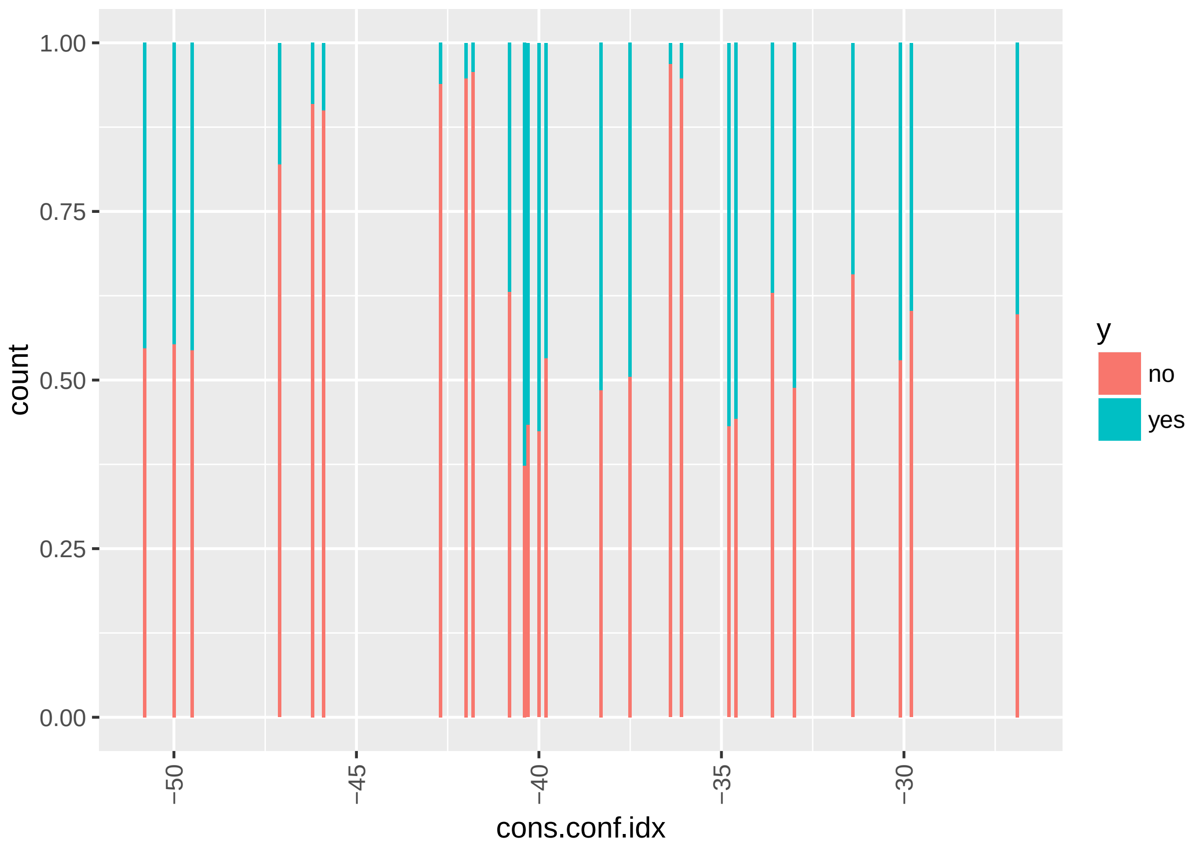

1.00
0.75
0.50
0.25
0.00
count
failure
nonexistent
success
poutcome
y
no
yes

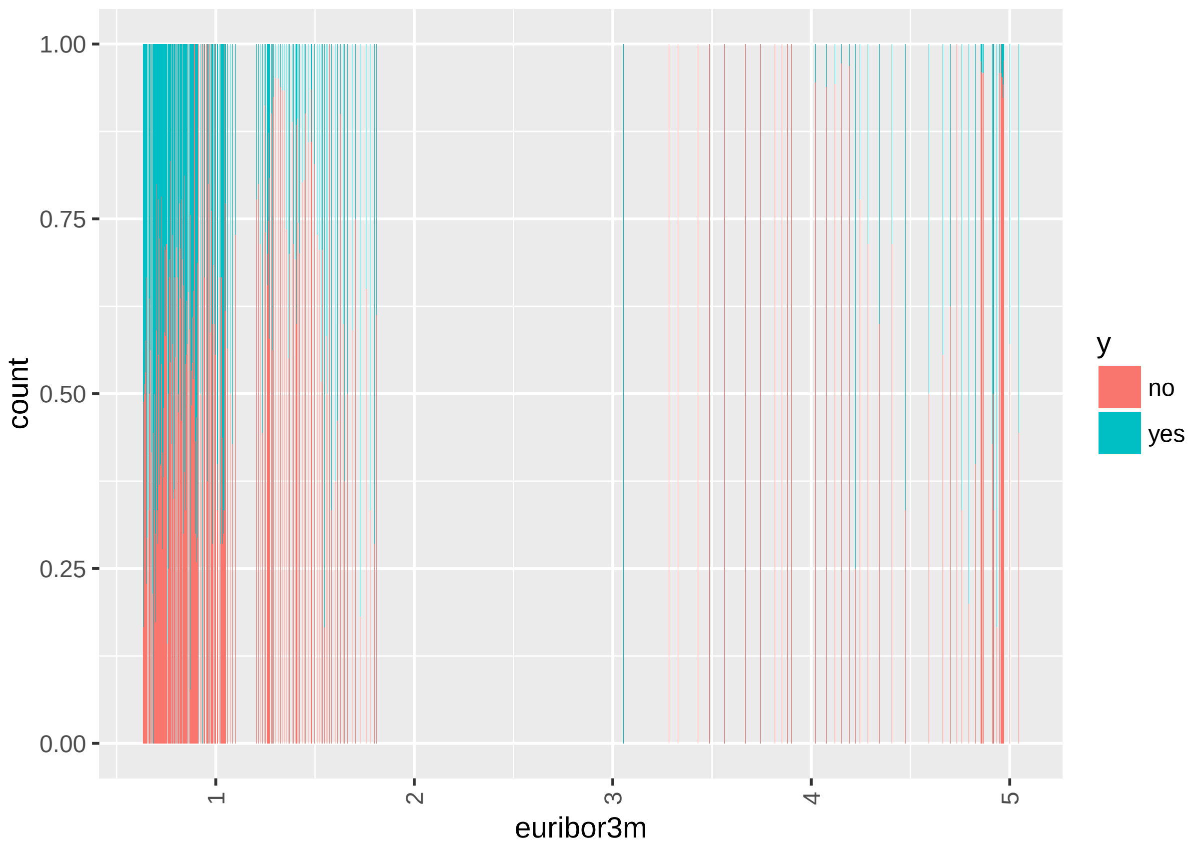
338



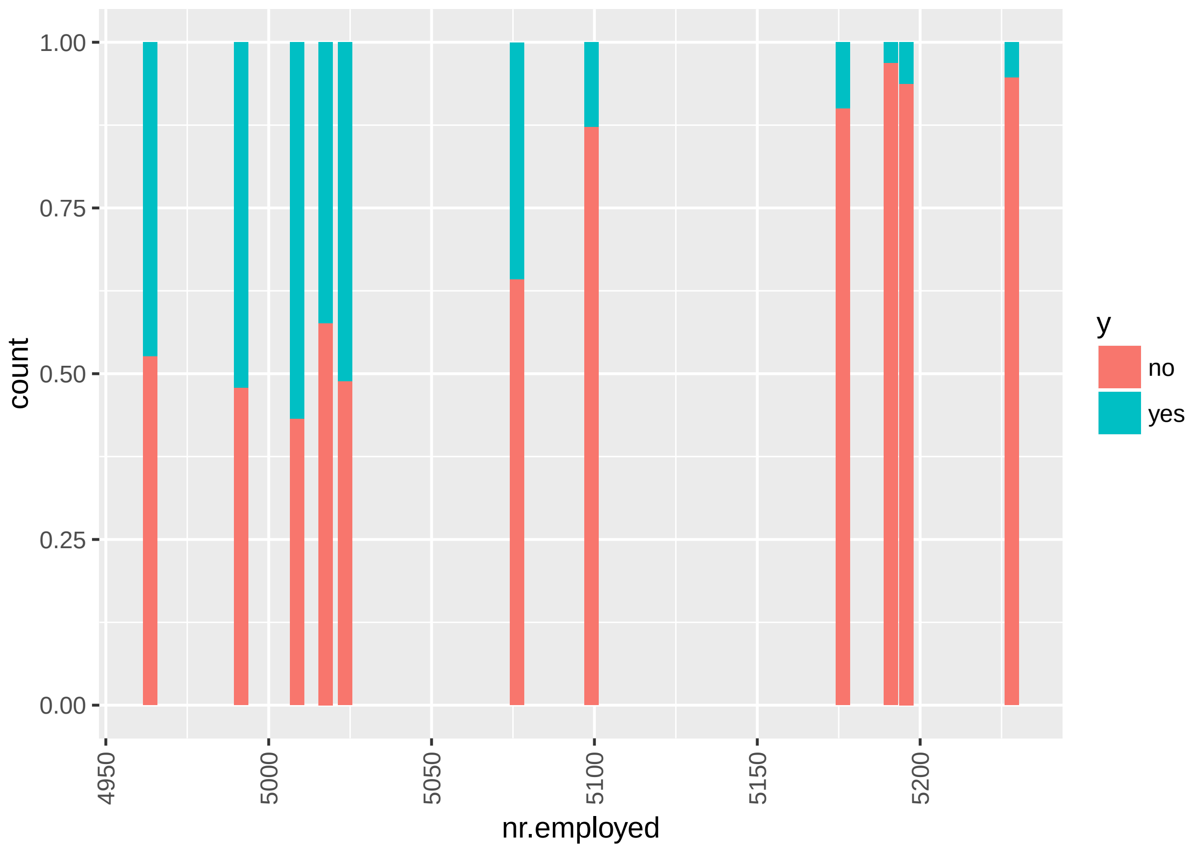
339


1.00
0.75
0.50
0.25
0.00
count
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



342



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 4

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 5

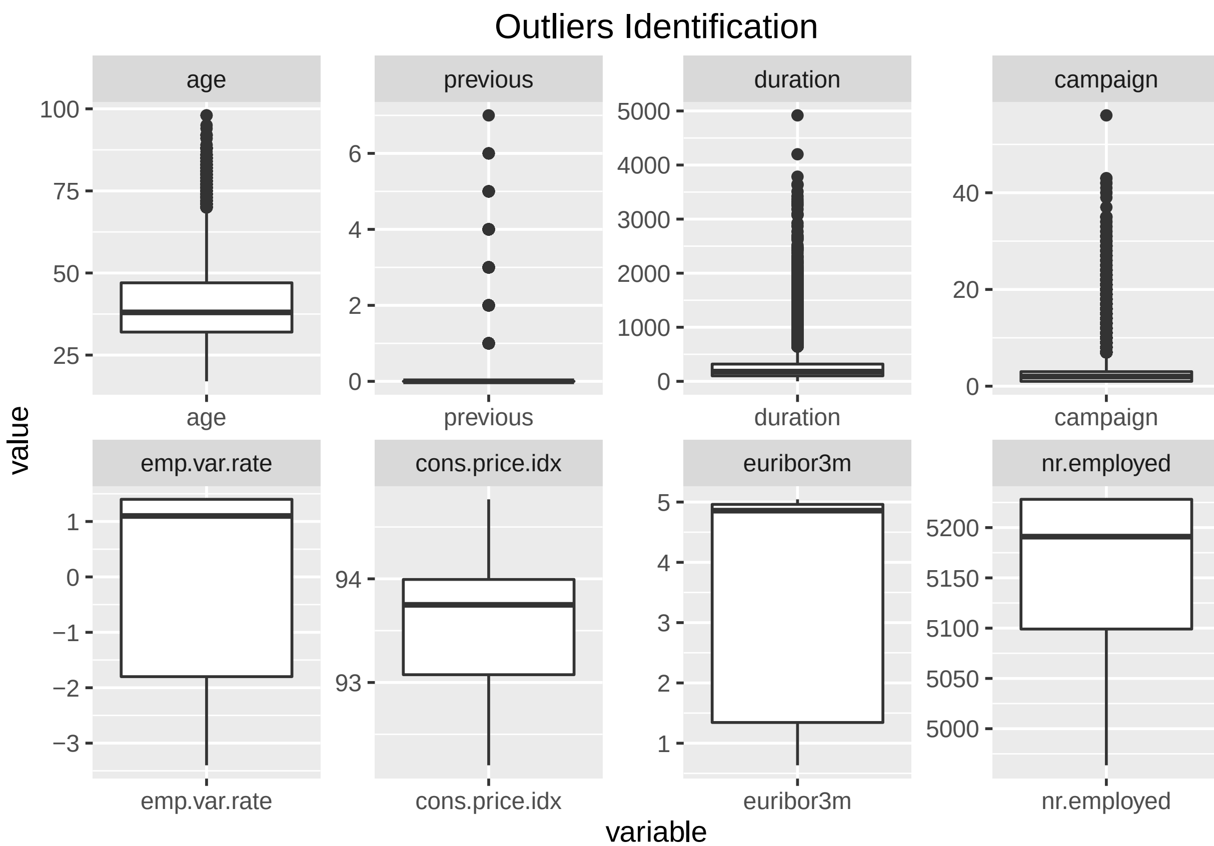
Unique Values

	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

Auguie, B. (2016). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2015). *Papaja: Create aPA manuscripts with rMarkdown*. Retrieved from https://github.com/crsh/papaja

Ballings, M., & Poel, D. V. den. (2013). *AUC: Threshold independent performance measures for probabilistic classifiers*. Retrieved from https://CRAN.R-project.org/package=AUC

Dahl, D. B. (2016). *Xtable: Export tables to laTeX or hTML*. Retrieved from https://CRAN.R-project.org/package=xtable

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press. Retrieved from http://statwww.epfl.ch/davison/BMA/

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. Retrieved from http://www.jstatsoft.org/v40/i08/

Faraway, J. (2016). *Faraway: Functions and datasets for books by julian faraway*. Retrieved from https://CRAN.R-project.org/package=faraway

Fortran code by Alan Miller, T. L. using. (2009). *Leaps: Regression subset selection*. Retrieved from https://CRAN.R-project.org/package=leaps

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. Retrieved from http://www.jstatsoft.org/v45/i07/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *ISLR: Data for an introduction to statistical learning with applications in r*. Retrieved from https://CRAN.R-project.org/package=ISLR

Lele, S. R., Keim, J. L., & Solymos, P. (2016). *ResourceSelection: Resource selection (probability) functions for use-availability data*. Retrieved from https://CRAN.R-project.org/package=ResourceSelection

Lesnoff, M., Lancelot, & R. (2012). *Aod: Analysis of overdispersed data*. Retrieved from http://cran.r-project.org/package=aod

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2016). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in r. *Bioinformatics*, *21*(20), 7881. Retrieved from http://rocr.bioinf.mpi-sb.mpg.de

Stubben, C. J., & Milligan, B. G. (2007). Estimating and analyzing demographic models using the popbio package in r. *Journal of Statistical Software*, *22*(11).

Therneau, T., Atkinson, B., & Ripley, B. (2015). *Rpart: Recursive partitioning and regression trees*. Retrieved from https://CRAN.R-project.org/package=rpart

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., … Venables, B. (2016). *Gplots: Various r programming tools for plotting data*. Retrieved from https://CRAN.R-project.org/package=gplots

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., … others. (2015). *Gdata: Various r programming tools for data manipulation*. Retrieved from https://CRAN.R-project.org/package=gdata

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2015). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., & Francois, R. (2015). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, & Hadley. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12). Retrieved from http://www.jstatsoft.org/v21/i12/paper

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from http://yihui.name/knitr/

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8). Retrieved from http://www.jstatsoft.org/v27/i08/