Logistic Regression Model in Analyzing Bank Marketing Data

1. Introduction

1.1 Background

The data set used is collected UCI machine learning repository. It gives information about direct marketing campaigns of a Portuguese banking institution. The most important part is to access if the product (bank term deposit) would be ('yes') or would not be('no') subscribed.

The following categories of information are included: Consumer data,Campaign activities, Social and economic environment data, Outcome.

1.2 Choice of the Data Set

The data set we picked here to conduct the analysis is "bank.csv", for 2 reasons. Firstly, the this data set is smaller than the original one, which greatly improve the efficiency. Secondly, this set is randomly selected from 3 earlier dataset. It slightly ameliorate the problem of unbalanced response by randomization.

1.3 Statistical questions of interest and Analysis Plan

The goal here is to predict if a consumer will continue to subscribe or not. The response here is binary and to answer this question the approperate statistical model should be used.The model we are using here is logistic regression model.

We use logistic regression model because of the following reasons. Firstly, the data is roughly linear separable, therefore linear models have good performance. Besides, before the calculation, we have removed any outliers that may have a negative effect on the accuracy of our model. Moreover, consider that the dimension is high, we expect that nonparametric models perform worse due to a common phenomenon called the curse of dimensionality. Based on the above analysis, the logistic regression model is suitable for our analysis.

2. Statistical Analysis

2.1 Descriptive Analysis

The purpose of this project is to fully explore the needs of customers, summarize the user profiles, and provide constructive suggestions for the development of marketing activities, so as to truly promote the development of banking business.

The marketing scenario of this dataset is to recommend a term deposit business to customers, and the promotion method is limited by telephone promotion, so in the description we will be more inclined to target our expected group. There are 4521 data and 17 variables, where $\boldsymbol{y}$ is the dependent variable and $\boldsymbol{y}$ = 'yes' represents marketing success. From the describe function we could summarize the following information:

- Blue collar, management and technician occupations are the most common among customers(Fig 1.3);
- The majority of customers receiving products are **married**, indicating that these customers have more demand than single customers(Fig 1.2);

- The average **age** of customers is about 41 years old. The oldest is 87 and the youngest is 19. For all clients, the first quartile of age is 33 and the third quartile is 49, indicating that half of the customers are in the 33-49 range. It means the majority of customers receiving products are middle-aged Fig 1.3).

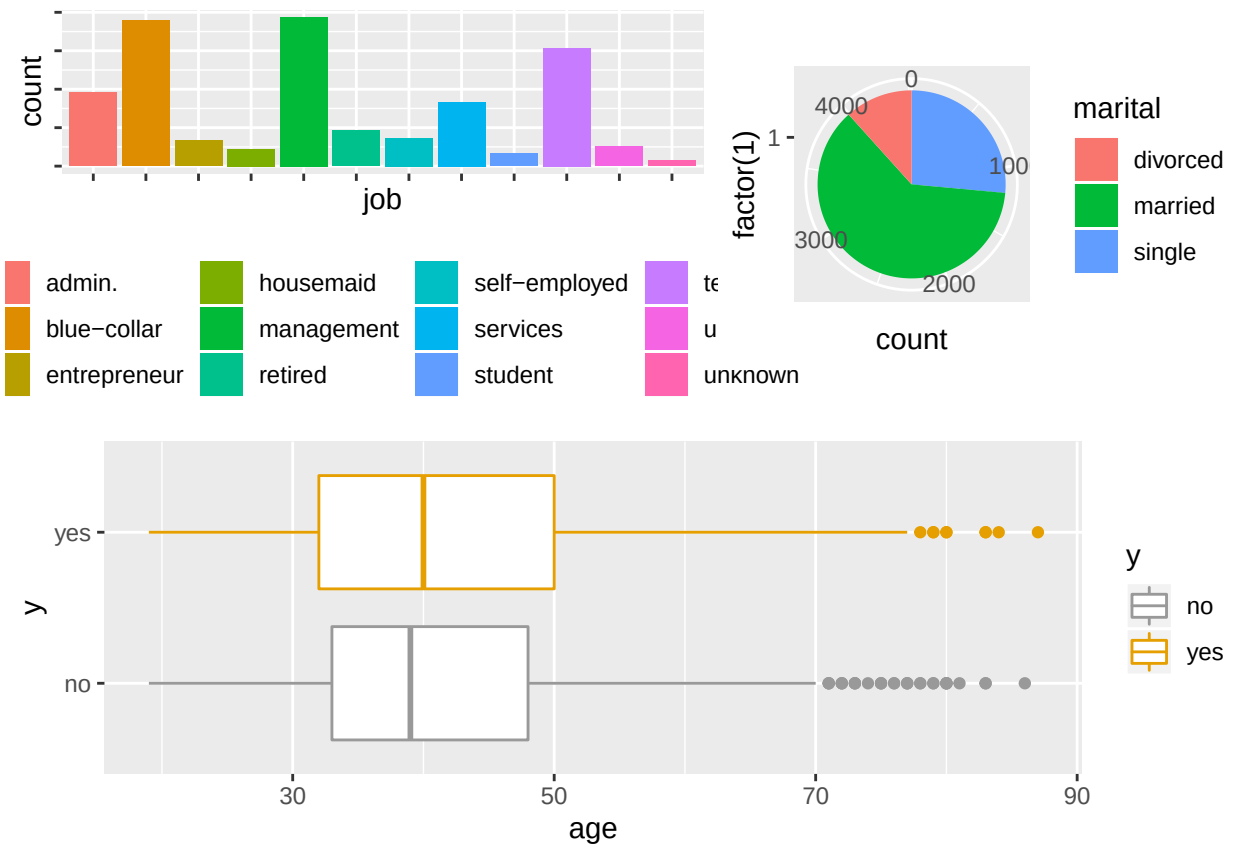


Figure 1: Top Left(1.1), Bar plot of 'job'; Top Right(1.2), Pie chart of 'marital'; Bottom(1.3): Boxplot of 'Age'.

- The average customer **balance** is 1423, but the standard deviation is large, indicating that the distribution of this data is scattered.
- The call duration ranges from 4 to 3025 seconds (almost an hour). Is it the last call time or the accumulated call time? Is it pure talk time or does it include waiting time? It's not described clearly. However, it is certain that the longer the call time, the greater the potential of the customer, and the corresponding deposit will be more.
- The number of contacts performed during this campaign ranged from 1 to 50 times. The more the corresponding number of contacts, the more likely that the customer participated in the previous event.

*You could check the Appendix 1 for complete output of the describe function.

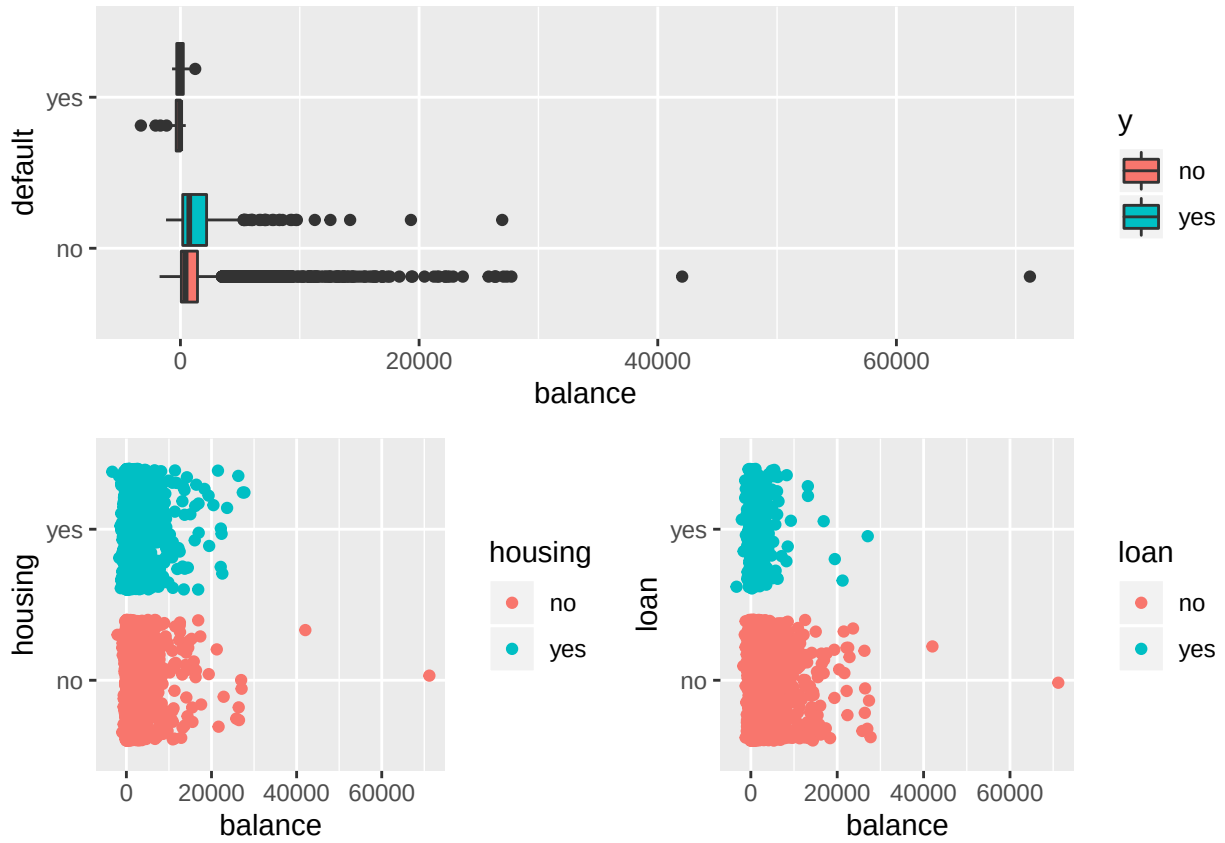


Figure 2: Top: Boxolot of default and balance grouped by whether the client has a term deposite or not. Bottom: Scatter plot of the relationship between balance and housing or personal loan.

- Significantly low balance of persons with default records, indicating that their financial situation is indeed not good;
- The existence of home loans and personal loans will directly affect the balance, and those without home loans and personal loans will have more surpluses.

Besides, we found that there seems to be no significant deviation in the earnings of people with different educational levels; there is no correlation between marital status and earnings, because the distribution of the surplus segments is relatively similar for divorced, single, and married persons.

2.2 Analysis Using Logistic Regression

Logistic Regression Defination

The logistic regression model is shown in the following equation:

$$\ln \frac{P(Y=1|x)}{1-P(Y=1|x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

in which x_i, $i = 1, \ldots, p$ are p independent variables. Some variables are actually categorical variables. For these variables, one hot encoding is conducted to convert the categorical variables to numerical, based on their alphabet sizes.

Assumptions

1, Linearity: linear relationship between predictors and the logit of the outcome.

2, No influential points and outliers: This assumption requires that the outliers are all elimnated cince the performance of the model will be effected

3, No Multicollinearity inside the data: the predictors should be uncorrelated.

Model fitting and some results

The following table shows the result of the analysis.

	education	housing	loan	contact	day	month	duration	campaign	poutcomr	marital
Est. Coef.	0.12	0.34	-0.643	0.07	0.016	-0.04	0.381	-0.31	0.48	-0.51

From the above table, we observe that the p value are not small, therefore the relationship between the mandatory jail sentence and the fatality rate is unclear.

Evaluation of the performance

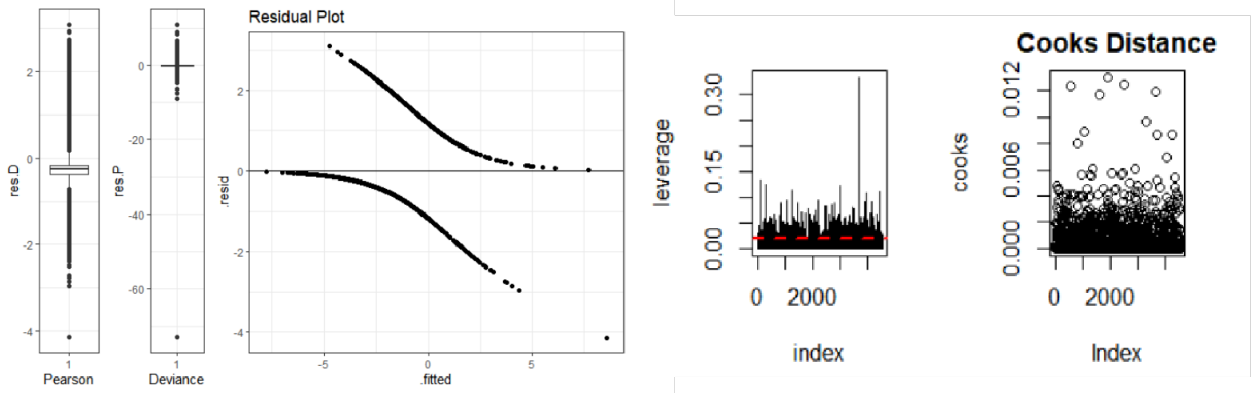


Figure 3: Left(3.1): Boxplots of two residuals and the Fitted value vs residual plot. Right(3.2): Leverage and Cook's distance.

Based on the above graph(Fig 3.1), two kinds of residuals gives different pattern in boxplot and the residual plot doesn't show a linear pattern, which indicates the lack of fit. So we further do RUNS test to ditermine the performance.

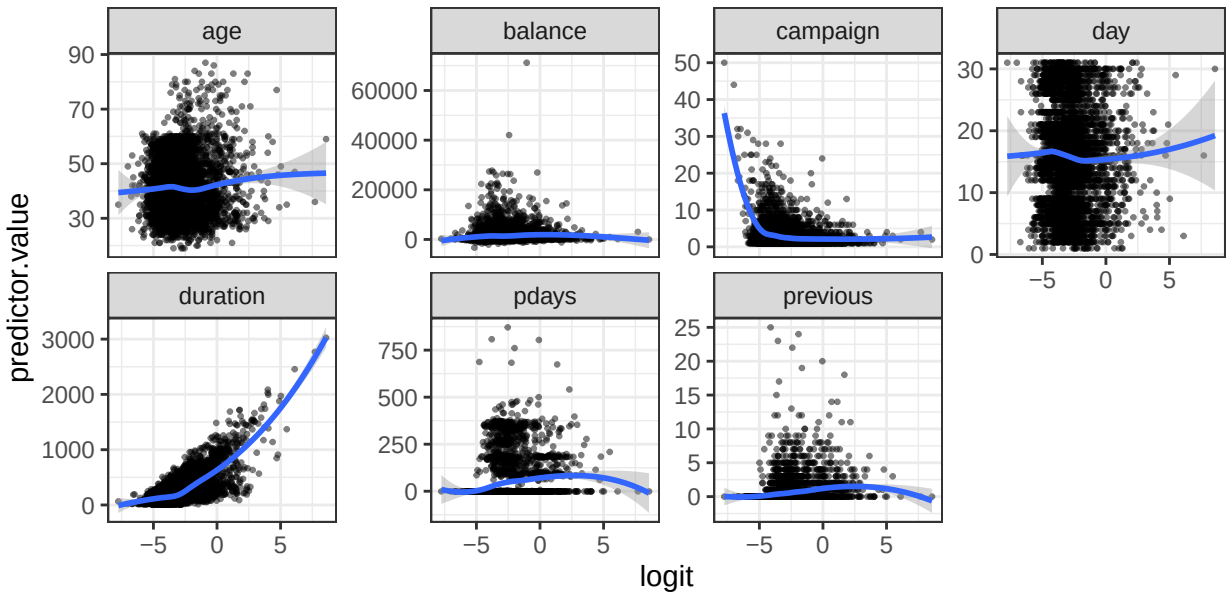
```
Standardized Runs Statistic = -0.074367, p-value = 0.9407
```

The null here suppose this is no lack-of-fit, while the P-value is large we fail to reject the null, it means furthur analysis is needed for this data and model

2.3 Model Diagnostic Analysis

Linrearity checking

Here we only check the linear relationship between the continous predictors and the response, it shows that most of predictos indeed is linear on the response while campaign, pdays and previous shows a quadratic pattern, which is one of the reason for lack-of-fit.



Influential point check

Here we check for outliers by leverage points plot and cooks distance and indeed there are influential points and the removal method is included in the coding to improve the goodness-of-fit(Fig 3.2).

Multicollinearity Check

Here we use VIF to check for multicollinearity:

	education	housing	loan	contact	day	month	duration	campaign	poutcomr	marital
GVIF	1.36	1.12	1.05	1.90	1.36	3.39	1.12	1.15	1.13	1.08

If GVIF is larger than 5 we consider it as problematic and the result shows that after proper model selsction, multicollinearity has indeed been removed.

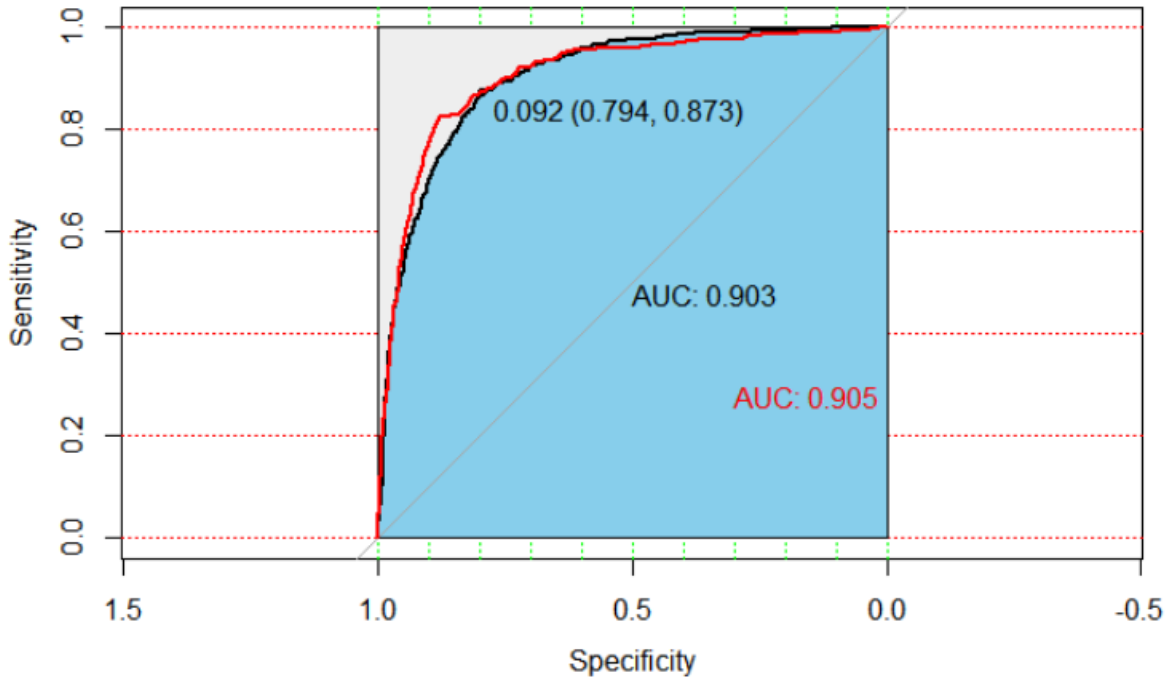
2.4 Another prediction model

We use Support Vector Machine as an alternative. SVM is expected to work here, since it also provides linear separation hyperplane, and it is suitable for processing high dimensional data. SVM model minimizes the following loss function:

$$L = \sum_{i=1}^{N} \max\{0, 1 - y_i w^T x_i\},$$

in which w is a $(p+1)$ dimensional vector that represents the weight.

Now we compare the result of logistic regression and SVM. The Receiver Operating Characteristics (ROC) curves are plotted as following:



From these figures, we observe that the ROC curves of these two models are approximately the same, and the Area Under Curve (AUC) value are both around 0.9. The result indicates that both logistic regression model and the SVM model has good performance for the classification of bank data.

3. Conclusion and Suggestion

From the overall prediction accuracy of the model, the logistic regression model and the SVM model show similar performance. Since we used the data set 'bank.cvs' which is a sample from raw dara, resampling data before modeling could lead to a better result which would reduce the bias of model prediction effects due to sample imbalance. Several stategies could conclude as following: If the time deposit business and the housing loan are negatively correlated, it can be considered that it is relatively difficult to open a time deposit business with a home loan. At the same time, the overall surplus of people with a home loan will be worse than those without a home loan. So next time the marketing center can be targeted at people with good earnings and no home loans. Judging from the marketing results, it seems that groups under 20 or over 60 are more likely to succeed in marketing, so they can be considered as the key marketing target. Students and retirees are more likely to have time deposits and succeed in marketing. It is not easy for blue-collars, entrepreneurs, service providers, and technicians to sell successfully. Sales to this type of people should be avoided as much as possible. For more accurate subsequent simulations, decision trees and neural network algorithms can be used for further optimization.

4. Reference

[1][Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2]Decision tree: http://f.dataguru.cn/thread-657436-1-1.html

Appendix 1

```
describe(data2)
```

```
## data2
##
##  17  Variables      4521  Observations
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0       67    0.999    41.17    11.81       27       29
##      .25      .50      .75      .90      .95
##       33       39       49       56       59
##
## lowest : 19 20 21 22 23, highest: 81 83 84 86 87
## ---------------------------------------------------------------------------
## job
##        n  missing distinct
##     4521        0       12
##
## lowest : admin.       blue-collar  entrepreneur housemaid    management
## highest: services     student      technician   unemployed   unknown
##
## admin. (478, 0.106), blue-collar (946, 0.209), entrepreneur (168, 0.037),
## housemaid (112, 0.025), management (969, 0.214), retired (230, 0.051),
## self-employed (183, 0.040), services (417, 0.092), student (84, 0.019),
## technician (768, 0.170), unemployed (128, 0.028), unknown (38, 0.008)
## ---------------------------------------------------------------------------
## marital
##        n  missing distinct
##     4521        0        3
##
## Value      divorced  married   single
## Frequency       528     2797     1196
## Proportion    0.117    0.619    0.265
## ---------------------------------------------------------------------------
## education
##        n  missing distinct
##     4521        0        4
##
## Value        primary secondary  tertiary   unknown
## Frequency        678      2306      1350       187
## Proportion     0.150     0.510     0.299     0.041
## ---------------------------------------------------------------------------
## default
##        n  missing distinct
##     4521        0        2
##
## Value         no   yes
## Frequency   4445    76
## Proportion 0.983 0.017
## ---------------------------------------------------------------------------
```

```
## balance
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0     2353        1     1423     2150     -162        0
##      .25      .50      .75      .90      .95
##       69      444     1480     3913     6102
##
## lowest : -3313 -2082 -1746 -1680 -1400, highest: 27069 27359 27733 42045 71188
## ------------------------------------------------------------------------
## housing
##        n  missing distinct
##     4521        0        2
##
## Value         no   yes
## Frequency   1962  2559
## Proportion 0.434 0.566
## ------------------------------------------------------------------------
## loan
##        n  missing distinct
##     4521        0        2
##
## Value         no   yes
## Frequency   3830   691
## Proportion 0.847 0.153
## ------------------------------------------------------------------------
## contact
##        n  missing distinct
##     4521        0        3
##
## Value       cellular telephone   unknown
## Frequency       2896       301      1324
## Proportion     0.641     0.067     0.293
## ------------------------------------------------------------------------
## day
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0       31    0.999    15.92    9.487        3        5
##      .25      .50      .75      .90      .95
##        9       16       21       28       30
##
## lowest :  1  2  3  4  5, highest: 27 28 29 30 31
## ------------------------------------------------------------------------
## month
##        n  missing distinct
##     4521        0       12
##
## lowest : apr aug dec feb jan, highest: mar may nov oct sep
##
## Value        apr   aug   dec   feb   jan   jul   jun   mar   may   nov
## Frequency    293   633    20   222   148   706   531    49  1398   389
## Proportion 0.065 0.140 0.004 0.049 0.033 0.156 0.117 0.011 0.309 0.086
##
## Value        oct   sep
## Frequency     80    52
## Proportion 0.018 0.012
## ------------------------------------------------------------------------
```

```
## duration
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0      875        1      264    242.4       31       58
##      .25      .50      .75      .90      .95
##      104      185      329      579      763
##
## lowest :    4    5    6    7    8, highest: 2029 2087 2456 2769 3025
## ------------------------------------------------------------------------
## campaign
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0       32    0.919    2.794    2.421        1        1
##      .25      .50      .75      .90      .95
##        1        2        3        6        8
##
## lowest :  1  2  3  4  5, highest: 30 31 32 44 50
## ------------------------------------------------------------------------
## pdays
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0      292     0.45    39.77    71.01       -1       -1
##      .25      .50      .75      .90      .95
##       -1       -1       -1      183      317
##
## lowest :  -1   1   2   3   5, highest: 687 761 804 808 871
## ------------------------------------------------------------------------
## previous
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4521        0       24    0.449   0.5426   0.9726        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        2        3
##
## lowest :  0  1  2  3  4, highest: 20 22 23 24 25
## ------------------------------------------------------------------------
## poutcome
##        n  missing distinct
##     4521        0        4
##
## Value      failure   other success unknown
## Frequency      490     197     129    3705
## Proportion   0.108   0.044   0.029   0.820
## ------------------------------------------------------------------------
## y
##        n  missing distinct     Info      Sum     Mean      Gmd
##     4521        0        2    0.306      521   0.1152    0.204
##
## ------------------------------------------------------------------------
```