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

use 250 words or less to summarize your problem, methodology, and major outcomes. Even though direct marketing is a standard method for banks to utilize in the face of competition and financial unstability, it has, however, been shown to exhibit poor performance. The telemarketing calls are simply not answered or answered and immediately disconnected. It is however welcomed by the right person who is in need of financial relief. The aim of this exercise is to target clients more effectively and efficiently based on the data from a Portuguese bank telemarketing effort. We first used logistic regression to predict the binary response variable. The outcomes. . . .

Keywords: select a few key words (up to five) related to your work. . . .logistic regression model, linear discriminant analysis (LDA), predictive modeling, bank telemarketing, direct marketing, Data Mining

How to do Effective and Sucessful Bank Telemarketing

Introduction

describe the background and motivation of your problem–

After looking at various options, we settled for this project for our final since it met all the requirements.

“Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences. Its main objective is to explore the relationship between a dependent variable and one or more independent variables (which are also called predictor or explanatory variables).” This is the definition provided by www.unesco.org for Regression Analysis

The most successful direct marketing is to predict the customers that have a higher probability to do business. Data exploration technique, is crucial to understand customer behavior. Many banks and services are moving to adopt the predictive technique based on the data mining to predict the customer profile before targeting them. The prediction or classification is the most important task in the data exploration and model building that is usually applied to classify the group of data. In classification, the outcome is a categorical variable and several combinations of input variable are used to build a model and the model that gives a better prediction with the best accuracy is chosen to target the prospective customers.

The data set contains approximately 41188 obs. of 21 variables.

This dataset is based on “Bank Marketing” UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb./

The binary classification goal is to predict if the client will subscribe a bank term deposit (variable y).

This dependent variable tells whether the client will subscribe a bank term deposit or not. This is a binary variable and as such we will be using a Logistic Regression Model.

Literature Review

discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the-art. Please do not discuss paper one at a time, instead, identify key characteristics of your topic, and discuss them in a whole. Please cite the relevant papers where appropriate.

We will be reviewing three papers addresseing the same problem of bank telemarketing.

1. http://bru-unide.iscte.pt/RePEc/pdfs/13-06.pdf
2. http://www.ijmbs.com/Vol6/1/4-vaidehi-r.pdf
3. http://www.columbia.edu/~jc4133/ADA-Project.pdf

Methodology

discuss the key aspects of your problem, data set and regression model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for regression modeling, your process for building regression models, and your model selection. .

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 infomation

- 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 varable and the unique values contained in each variable and what they represented. Wecan divide the variables in the follwing 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.

Experimentation and Results

describe the specifics of what you did (data exploration, data preparation, model building, selection, evalutation) and what you found out (statistical analysis, inter pretation and discussion of the results)

Data Exploration

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

-Variable identification

-Missing values and Unique Values

-Variables relationship to y

Table 1

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
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

Variable	Data.Type	Type	Description
nr.employed	Numeric	Predictor	Quarterly number of employees
y	Binary	Response	Has the client subscribed a term deposit?

We notice that the variables are numerical, categorical and binary. The responce 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 most 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 2 shows us that there are no missing values per say, since they are all have the values of either "unknown" or "999" in our dataset as shown in table 2 and graph format.

Table 2

Missing Values

	Missing Values
age	0
job	0
marital	0
education	0
default	0
housing	0
loan	0
contact	0
month	0

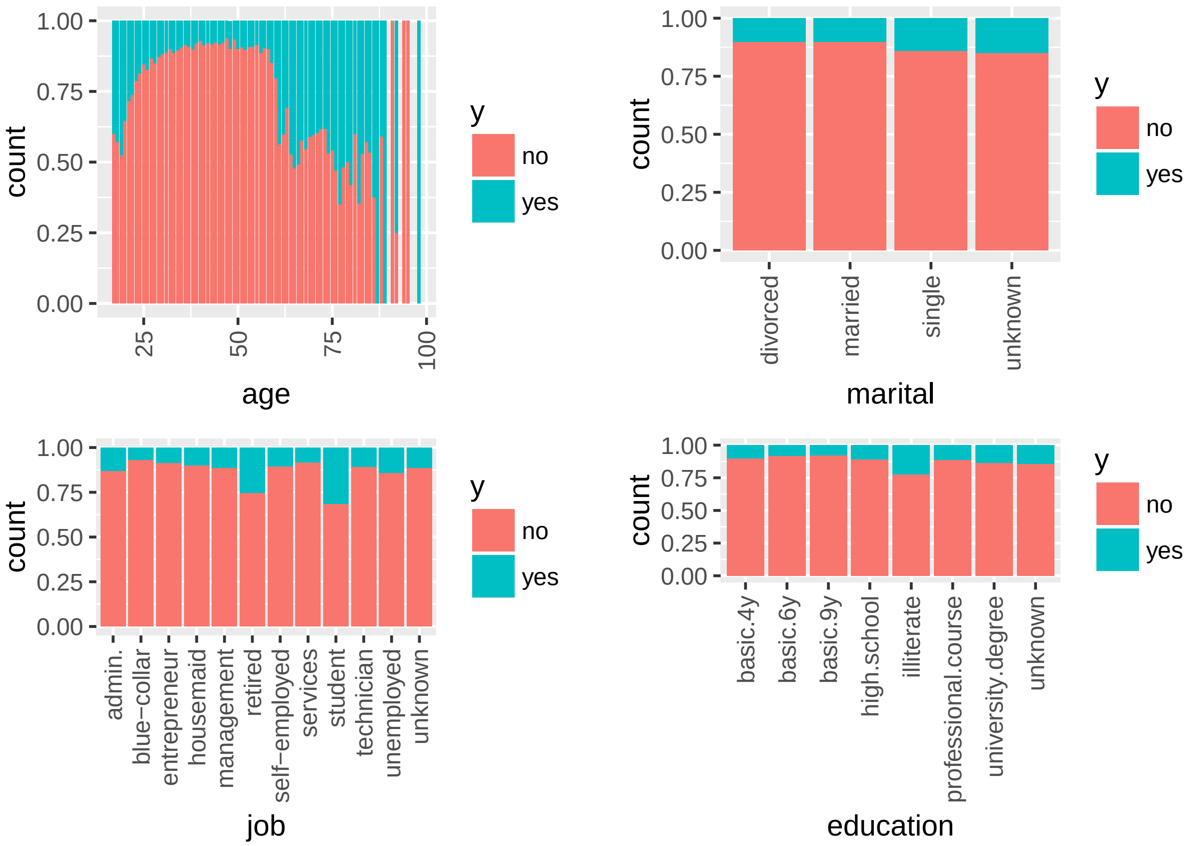
	Missing Values
day_of_week	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	0
emp.var.rate	0
cons.price.idx	0
cons.conf.idx	0
euribor3m	0
nr.employed	0
y	0

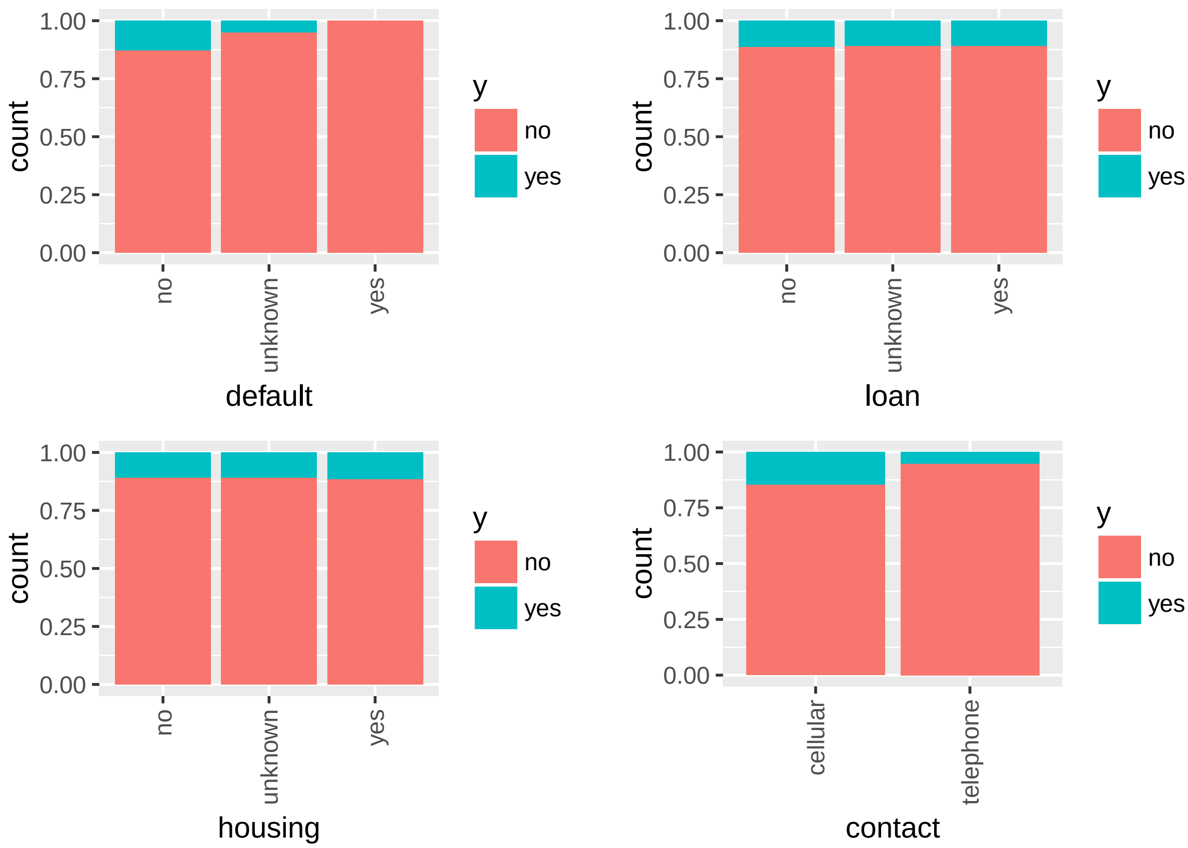
Table 3

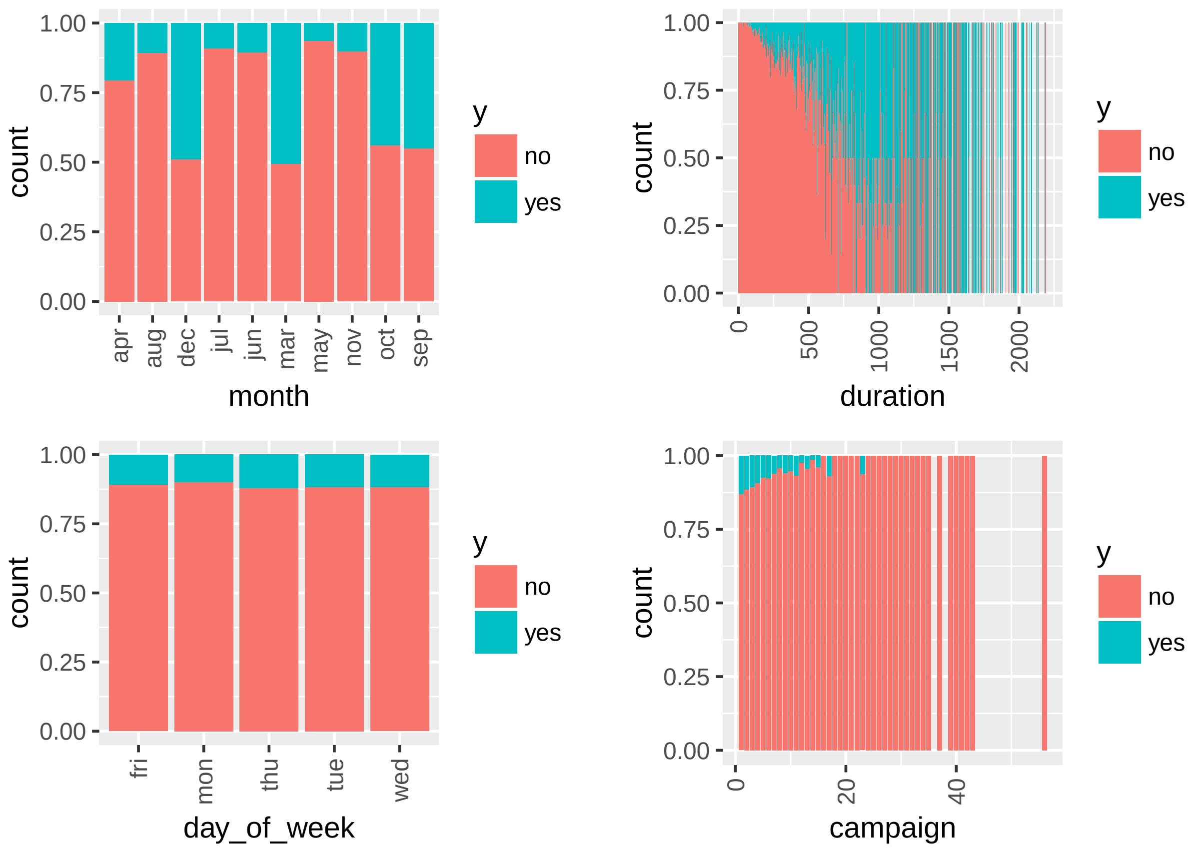
Unique Values

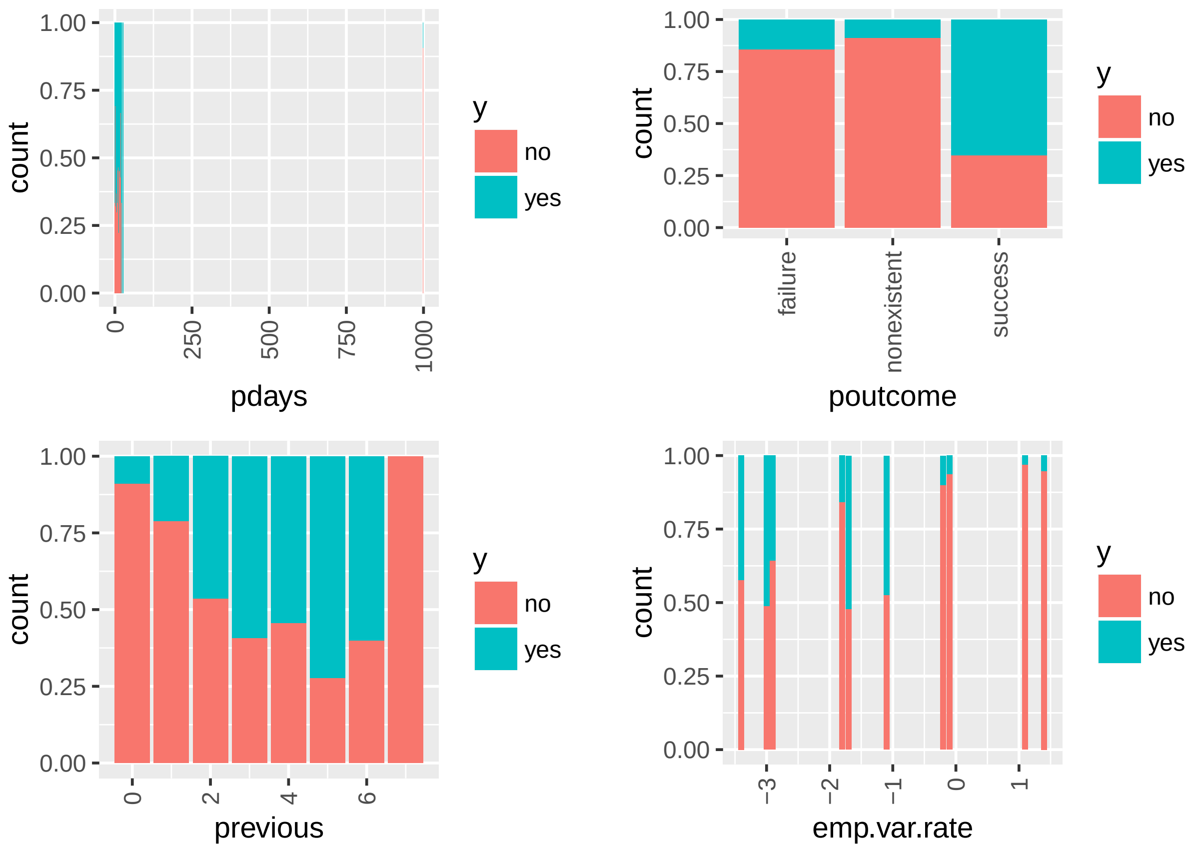
	Unique Values
age	78
job	12
marital	4
education	8
default	3
housing	3
loan	3
contact	2
month	10

	Unique Values
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


count
1.00
0.75
0.50
0.25
0.00
25
50
75
100
age
y
no
yes
marital
divorced
married
single
unknown
job
admin.
blue−collar
entrepreneur
housemaid
management
retired
self−employed
services
student
technician
unemployed
unknown
education
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown


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


count
1.00
0.75
0.50
0.25
0.00
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month
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
fri
mon
thu
tue
wed
day_of_week
y
no
yes
count
1.00
0.75
0.50
0.25
0.00
0
20
40
campaign
y
no
yes


count
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
pdays
y
no
yes
count
failure
nonexistent
success
poutcome
count
0
2
4
6
previous
count
−3
−2
−1
0
1
emp.var.rate

107

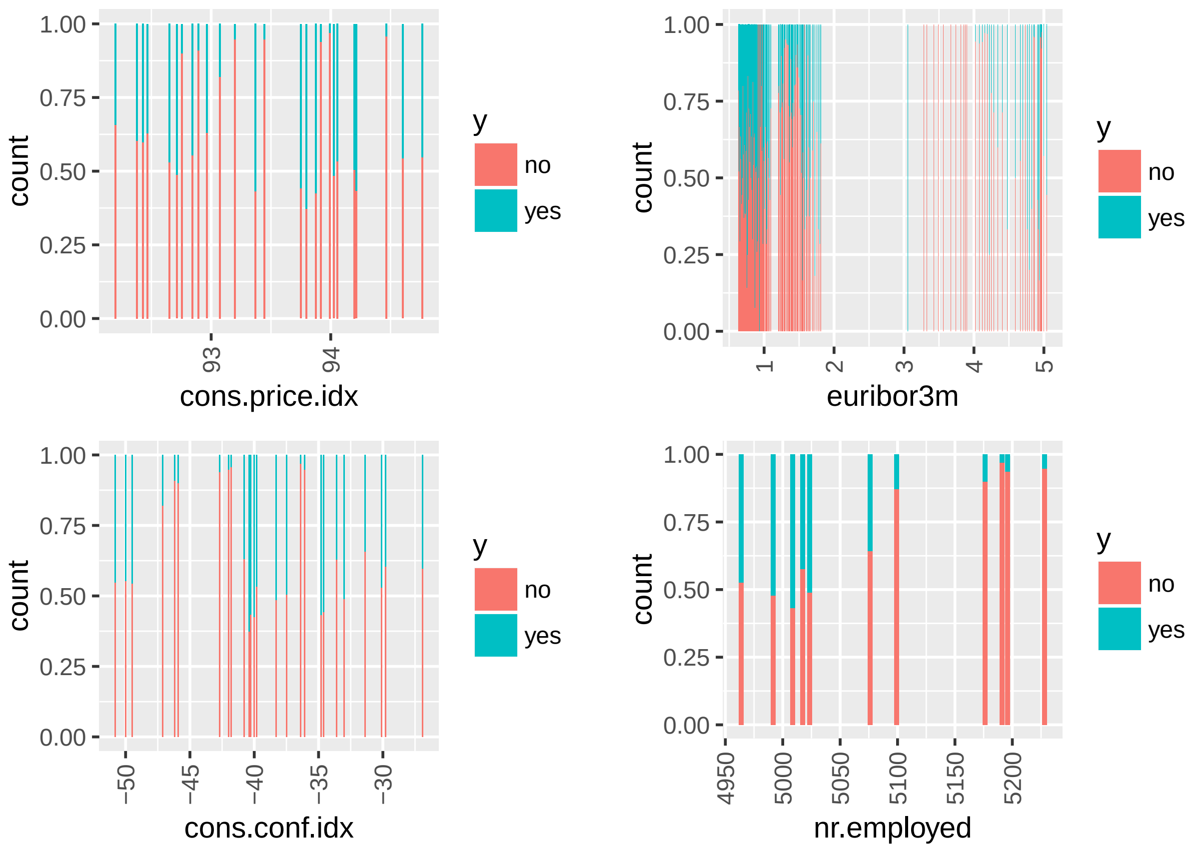


Table 4 shows the analysis of variables after data exploration.

Table 4

Variable Description

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
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

Variable	Data.Type	Analysis
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 varial
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 to 0 and 1

-Create dummy variables

-Data Summary Analysis

-Correlation of Variables with y

Convert to Binary. Now in order to prepare the data for modeling, we need to update Yes = 1 and No = 0.

Create dummy variables. Now we need to create dummy variables to find out the relationship between y variables and dependent variables, for all categorical variables.

Table 5

Data Summary (Part 1/3

	vars	n	mean	sd	median
age	1	41188	40.0240604	10.4212500	38.000
duration	2	41188	258.2850102	259.2792488	180.000
campaign	3	41188	2.5675925	2.7700135	2.000
pdays	4	41188	962.4754540	186.9109073	999.000
previous	5	41188	0.1729630	0.4949011	0.000
emp.var.rate	6	41188	0.0818855	1.5709597	1.100
cons.price.idx	7	41188	93.5756644	0.5788400	93.749
cons.conf.idx	8	41188	-40.5026003	4.6281979	-41.800
euribor3m	9	41188	3.6212908	1.7344474	4.857
nr.employed	10	41188	5167.0359109	72.2515277	5191.000
y	11	41188	0.1126542	0.3161734	0.000
job_housemaid	12	41188	0.0257357	0.1583475	0.000
job_services	13	41188	0.0963630	0.2950920	0.000
job_admin.	14	41188	0.2530349	0.4347560	0.000
job_blue-collar	15	41188	0.2246771	0.4173746	0.000
job_technician	16	41188	0.1637127	0.3700192	0.000
job_retired	17	41188	0.0417597	0.2000421	0.000
job_management	18	41188	0.0709916	0.2568138	0.000
job_unemployed	19	41188	0.0246188	0.1549623	0.000
job_self-employed	20	41188	0.0345003	0.1825127	0.000
job_unknown	21	41188	0.0080120	0.0891518	0.000
job_entrepreneur	22	41188	0.0353501	0.1846654	0.000
job_student	23	41188	0.0212441	0.1441986	0.000

	vars	n	mean	sd	median
marital_married	24	41188	0.6052248	0.4888083	1.000
marital_single	25	41188	0.2808585	0.4494240	0.000
marital_divorced	26	41188	0.1119744	0.3153387	0.000
marital_unknown	27	41188	0.0019423	0.0440294	0.000
education_illiterate	28	41188	0.0004370	0.0209007	0.000
education_unknown	29	41188	0.0420268	0.2006528	0.000
education_primary	30	41188	0.1570360	0.3638392	0.000
education_secondary	31	41188	0.3777799	0.4848381	0.000
education_tertiary	32	41188	0.4227202	0.4939977	0.000
default_no	33	41188	0.7912013	0.4064552	1.000
default_unknown	34	41188	0.2087258	0.4064030	0.000
default_yes	35	41188	0.0000728	0.0085342	0.000
housing_no	36	41188	0.4521220	0.4977085	0.000
housing_yes	37	41188	0.5238419	0.4994373	1.000
housing_unknown	38	41188	0.0240361	0.1531632	0.000
loan_no	39	41188	0.8242692	0.3805956	1.000
loan_yes	40	41188	0.1516947	0.3587290	0.000
loan_unknown	41	41188	0.0240361	0.1531632	0.000
contact_telephone	42	41188	0.3652520	0.4815066	0.000
contact_cellular	43	41188	0.6347480	0.4815066	1.000
month_may	44	41188	0.3342964	0.4717496	0.000
month_jun	45	41188	0.1291153	0.3353316	0.000
month_jul	46	41188	0.1741769	0.3792662	0.000
month_aug	47	41188	0.1499951	0.3570710	0.000
month_oct	48	41188	0.0174323	0.1308770	0.000

	vars	n	mean	sd	median
month_nov	49	41188	0.0995678	0.2994265	0.000
month_dec	50	41188	0.0044188	0.0663276	0.000
month_mar	51	41188	0.0132563	0.1143717	0.000
month_apr	52	41188	0.0639021	0.2445814	0.000
month_sep	53	41188	0.0138390	0.1168238	0.000
day_of_week_mon	54	41188	0.2067107	0.4049511	0.000
day_of_week_tue	55	41188	0.1964164	0.3972919	0.000
day_of_week_wed	56	41188	0.1974847	0.3981059	0.000
day_of_week_thu	57	41188	0.2093571	0.4068547	0.000
day_of_week_fri	58	41188	0.1900311	0.3923302	0.000
previous_contact	59	41188	0.0367826	0.1882298	0.000
poutcome_nonexistent	60	41188	0.8634311	0.3433958	1.000
poutcome_failure	61	41188	0.1032340	0.3042679	0.000
poutcome_success	62	41188	0.0333350	0.1795119	0.000

Table 6

Data Summary (Part 2/3)

	trimmed	mad	min	max	range
age	39.3033807	10.3782000	17.000	98.000	81.000
duration	210.6102513	139.3644000	0.000	4918.000	4918.000
campaign	1.9914118	1.4826000	1.000	56.000	55.000
pdays	999.0000000	0.0000000	0.000	999.000	999.000
previous	0.0457332	0.0000000	0.000	7.000	7.000
emp.var.rate	0.2661204	0.4447800	-3.400	1.400	4.800
cons.price.idx	93.5807666	0.5633880	92.201	94.767	2.566

	trimmed	mad	min	max	range
cons.conf.idx	-40.6015356	6.5234400	-50.800	-26.900	23.900
euribor3m	3.8055852	0.1601208	0.634	5.045	4.411
nr.employed	5178.4253338	55.0044600	4963.600	5228.100	264.500
y	0.0158412	0.0000000	0.000	1.000	1.000
job__housemaid	0.0000000	0.0000000	0.000	1.000	1.000
job__services	0.0000000	0.0000000	0.000	1.000	1.000
job__admin.	0.1913086	0.0000000	0.000	1.000	1.000
job__blue-collar	0.1558631	0.0000000	0.000	1.000	1.000
job__technician	0.0796613	0.0000000	0.000	1.000	1.000
job__retired	0.0000000	0.0000000	0.000	1.000	1.000
job__management	0.0000000	0.0000000	0.000	1.000	1.000
job__unemployed	0.0000000	0.0000000	0.000	1.000	1.000
job__self-employed	0.0000000	0.0000000	0.000	1.000	1.000
job__unknown	0.0000000	0.0000000	0.000	1.000	1.000
job__entrepreneur	0.0000000	0.0000000	0.000	1.000	1.000
job__student	0.0000000	0.0000000	0.000	1.000	1.000
marital__married	0.6315246	0.0000000	0.000	1.000	1.000
marital__single	0.2260864	0.0000000	0.000	1.000	1.000
marital__divorced	0.0149915	0.0000000	0.000	1.000	1.000
marital__unknown	0.0000000	0.0000000	0.000	1.000	1.000
education__illiterate	0.0000000	0.0000000	0.000	1.000	1.000
education__unknown	0.0000000	0.0000000	0.000	1.000	1.000
education__primary	0.0713159	0.0000000	0.000	1.000	1.000
education__secondary	0.3472323	0.0000000	0.000	1.000	1.000
education__tertiary	0.4034050	0.0000000	0.000	1.000	1.000

	trimmed	mad	min	max	range
default_no	0.8639840	0.0000000	0.000	1.000	1.000
default_unknown	0.1359250	0.0000000	0.000	1.000	1.000
default_yes	0.0000000	0.0000000	0.000	1.000	1.000
housing_no	0.4401554	0.0000000	0.000	1.000	1.000
housing_yes	0.5298009	0.0000000	0.000	1.000	1.000
housing_unknown	0.0000000	0.0000000	0.000	1.000	1.000
loan_no	0.9053168	0.0000000	0.000	1.000	1.000
loan_yes	0.0646395	0.0000000	0.000	1.000	1.000
loan_unknown	0.0000000	0.0000000	0.000	1.000	1.000
contact_telephone	0.3315732	0.0000000	0.000	1.000	1.000
contact_cellular	0.6684268	0.0000000	0.000	1.000	1.000
month_may	0.2928806	0.0000000	0.000	1.000	1.000
month_jun	0.0364166	0.0000000	0.000	1.000	1.000
month_jul	0.0927410	0.0000000	0.000	1.000	1.000
month_aug	0.0625152	0.0000000	0.000	1.000	1.000
month_oct	0.0000000	0.0000000	0.000	1.000	1.000
month_nov	0.0000000	0.0000000	0.000	1.000	1.000
month_dec	0.0000000	0.0000000	0.000	1.000	1.000
month_mar	0.0000000	0.0000000	0.000	1.000	1.000
month_apr	0.0000000	0.0000000	0.000	1.000	1.000
month_sep	0.0000000	0.0000000	0.000	1.000	1.000
day_of_week_mon	0.1334062	0.0000000	0.000	1.000	1.000
day_of_week_tue	0.1205390	0.0000000	0.000	1.000	1.000
day_of_week_wed	0.1218742	0.0000000	0.000	1.000	1.000
day_of_week_thu	0.1367140	0.0000000	0.000	1.000	1.000

	trimmed	mad	min	max	range
day_of_week_fri	0.1125577	0.0000000	0.000	1.000	1.000
previous_contact	0.0000000	0.0000000	0.000	1.000	1.000
poutcome_nonexistent	0.9542668	0.0000000	0.000	1.000	1.000
poutcome_failure	0.0040665	0.0000000	0.000	1.000	1.000
poutcome_success	0.0000000	0.0000000	0.000	1.000	1.000

Table 7

Data Summary (Part 3/3)

	skew	kurtosis	se
age	0.7846397	0.7908857	0.0513493
duration	3.2629036	20.2442057	1.2775632
campaign	4.7621598	36.9732194	0.0136489
pdays	-4.9218314	22.2253936	0.9209781
previous	3.8317631	20.1051076	0.0024386
emp.var.rate	-0.7240428	-1.0627423	0.0077407
cons.price.idx	-0.2308708	-0.8299589	0.0028522
cons.conf.idx	0.3031578	-0.3587887	0.0228048
euribor3m	-0.7091363	-1.4068549	0.0085463
nr.employed	-1.0441863	-0.0040511	0.3560096
y	2.4501517	4.0033404	0.0015579
job_housemaid	5.9900255	33.8812283	0.0007802
job_services	2.7356021	5.4836522	0.0014540
job_admin.	1.1360815	-0.7093361	0.0021422
job_blue-collar	1.3192765	-0.2595158	0.0020566
job_technician	1.8176306	1.3038128	0.0018232

	skew	kurtosis	se
job__retired	4.5813276	18.9890235	0.0009857
job__management	3.3409260	9.1620092	0.0012654
job__unemployed	6.1352936	35.6426931	0.0007636
job__self-employed	5.1008881	24.0196428	0.0008993
job__unknown	11.0368168	119.8142342	0.0004393
job__entrepreneur	5.0322224	23.3238288	0.0009099
job__student	6.6400673	42.0915155	0.0007105
marital__married	-0.4305257	-1.8146917	0.0024085
marital__single	0.9751869	-1.0490361	0.0022145
marital__divorced	2.4609486	4.0563667	0.0015538
marital__unknown	22.6233213	509.8270434	0.0002169
education__illiterate	47.8022616	2283.1116468	0.0001030
education__unknown	4.5647225	18.8371487	0.0009887
education__primary	1.8852047	1.5540345	0.0017928
education__secondary	0.5041563	-1.7458688	0.0023890
education__tertiary	0.3128675	-1.9021601	0.0024341
default__no	-1.4328481	0.0530549	0.0020028
default__unknown	1.4333905	0.0546097	0.0020025
default__yes	117.1551691	13723.6668447	0.0000421
housing__no	0.1923892	-1.9630341	0.0024524
housing__yes	-0.0954727	-1.9909333	0.0024609
housing__unknown	6.2149702	36.6267442	0.0007547
loan__no	-1.7039679	0.9035286	0.0018753
loan__yes	1.9418382	1.7707787	0.0017676
loan__unknown	6.2149702	36.6267442	0.0007547

	skew	kurtosis	se
contact_telephone	0.5596796	-1.6867997	0.0023726
contact_cellular	-0.5596796	-1.6867997	0.0023726
month_may	0.7024895	-1.5065451	0.0023245
month_jun	2.2119941	2.8929884	0.0016523
month_jul	1.7181345	0.9520092	0.0018688
month_aug	1.9603741	1.8431112	0.0017594
month_oct	7.3741903	52.3799548	0.0006449
month_nov	2.6745954	5.1535859	0.0014754
month_dec	14.9430876	221.3012387	0.0003268
month_mar	8.5114073	70.4457653	0.0005636
month_apr	3.5659885	10.7165344	0.0012051
month_sep	8.3227782	67.2702700	0.0005756
day_of_week_mon	1.4484821	0.0981028	0.0019953
day_of_week_tue	1.5282275	0.3354874	0.0019576
day_of_week_wed	1.5197359	0.3096048	0.0019616
day_of_week_thu	1.4286962	0.0411737	0.0020047
day_of_week_fri	1.5801046	0.4967426	0.0019332
previous_contact	4.9217092	22.2237610	0.0009275
poutcome_nonexistent	-2.1166376	2.4802150	0.0016920
poutcome_failure	2.6079414	4.8014749	0.0014992
poutcome_success	5.1991402	25.0316666	0.0008845

Data Summary Analysis.

Correlation of Variables with y. Now we will produce the correlation table between the independent variables and the dependent variable

Table 8

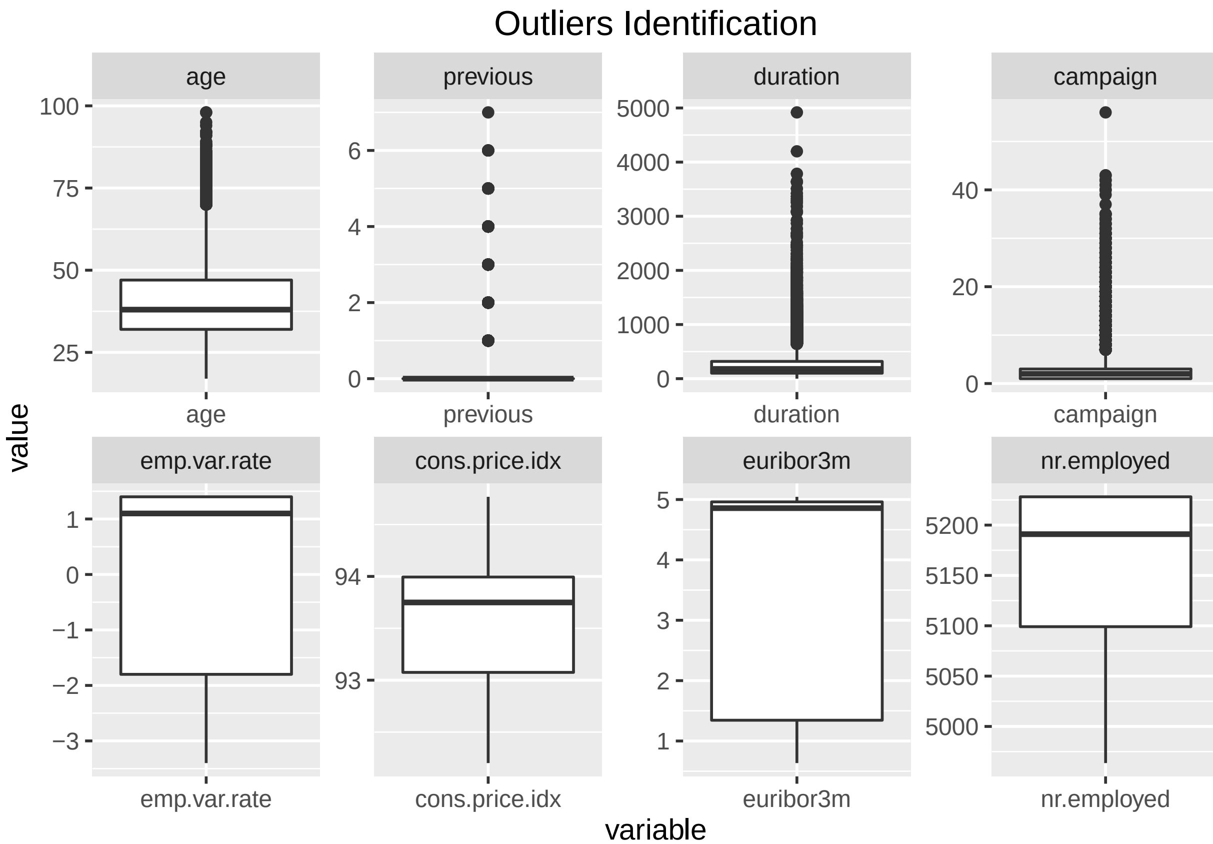
Correlation between "y" and predictor variables

	Correlation
y	1.0000000
duration	0.4052738
previous_contact	0.3248767
poutcome_success	0.3162694
previous	0.2301810
contact_cellular	0.1447731
month_mar	0.1440140
month_oct	0.1373659
month_sep	0.1260674
default_no	0.0993445
job_student	0.0939550
job_retired	0.0922208
month_dec	0.0793034
month_apr	0.0761364
cons.conf.idx	0.0548779
marital_single	0.0541335
education_tertiary	0.0471911
poutcome_failure	0.0317987
job_admin.	0.0314260
age	0.0303988
education_unknown	0.0214301
job_unemployed	0.0147519
day_of_week_thu	0.0138884

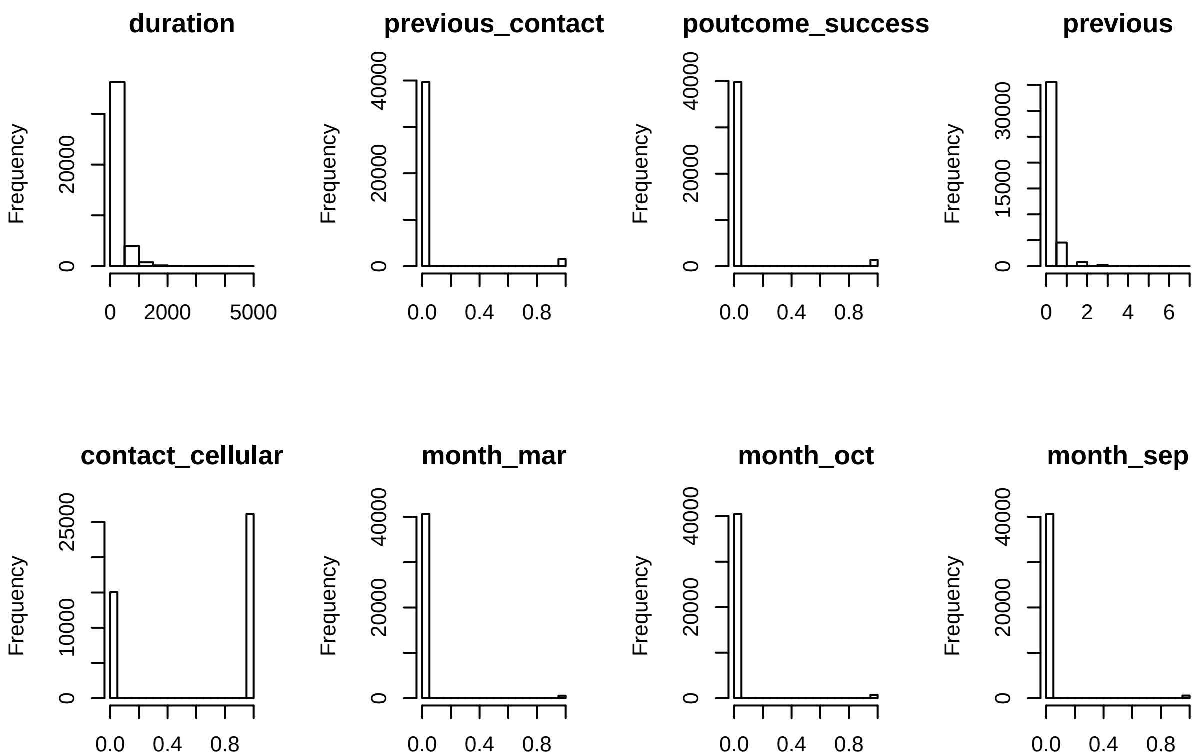
	Correlation
housing_yes	0.0117429
day_of_week_tue	0.0080461
education_illiterate	0.0072462
day_of_week_wed	0.0063020
marital_unknown	0.0052108
loan_no	0.0051231
job_unknown	-0.0001515
job_management	-0.0004189
housing_unknown	-0.0022700
loan_unknown	-0.0022700
default_yes	-0.0030410
loan_yes	-0.0044661
job_self-employed	-0.0046625
job_technician	-0.0061486
job_housemaid	-0.0065049
day_of_week_fri	-0.0069963
month_aug	-0.0088126
month_jun	-0.0091818
marital_divorced	-0.0106080
housing_no	-0.0110852
month_nov	-0.0117959
job_entrepreneur	-0.0166439
day_of_week_mon	-0.0212649
education_primary	-0.0237753
month_jul	-0.0322301

	Correlation
job_services	-0.0323009
education_secondary	-0.0394222
marital_married	-0.0433978
campaign	-0.0663574
job_blue-collar	-0.0744233
default_unknown	-0.0992934
month_may	-0.1082712
cons.price.idx	-0.1362112
contact_telephone	-0.1447731
poutcome_nonexistent	-0.1935068
emp.var.rate	-0.2983344
euribor3m	-0.3077714
pdays	-0.3249145
nr.employed	-0.3546783

Outliers.



Histograms of Variables



!!!!!!!

Analysis the link function. 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:

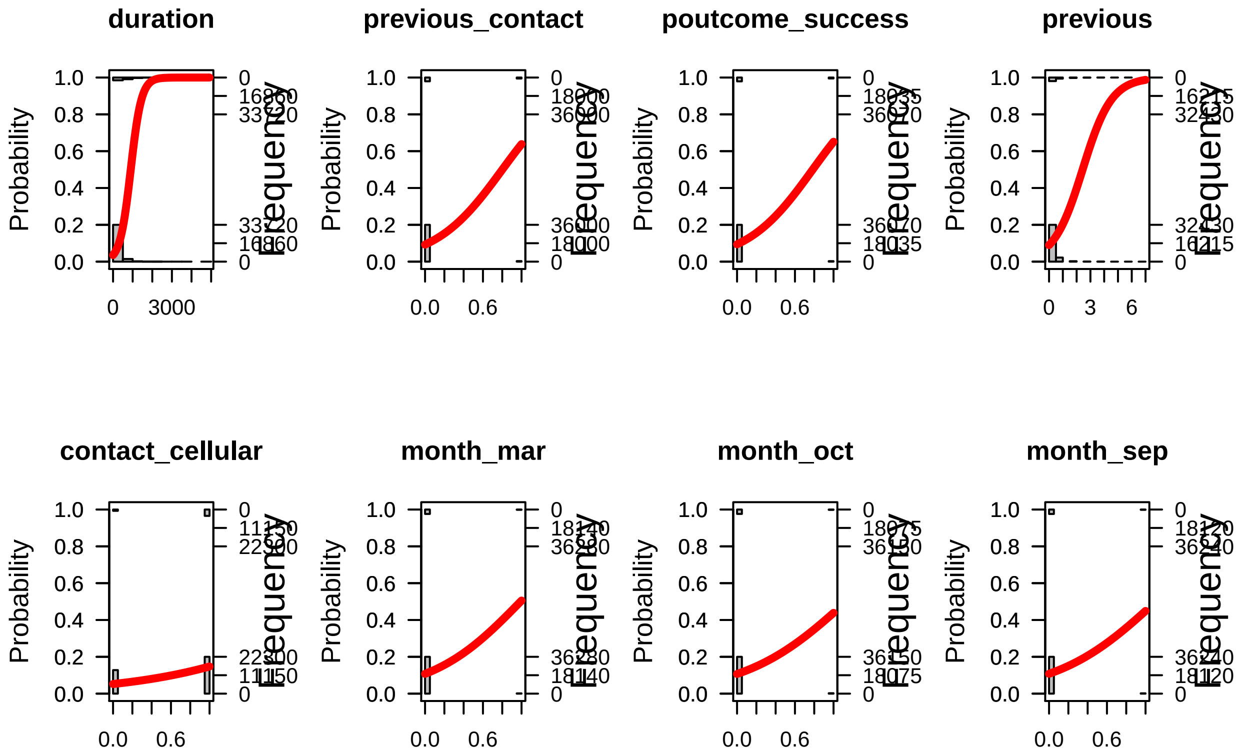
$g(E(y)) = a + Bx_1 + B_2x_2 + B_3x_3 + ...$

where, g() is the link function, E(y) is the expectation of target variable and $B_0 + B_1x_1 + B_2x_2 + B_3x_3$ is the linear predictor (B_0, B_1, B_2, B_3 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 (p) and Probability of Failure (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).



!!!!!!!

Prepare test data. Now in order to prepare the data for modeling, we need to update Yes = 1 and No = 0.

Test - Create dummy variables

Now we need to create dummy variables to find out the relationship between y variables and dependent variables, for all categorical variables.

Model Building

In this section, we will create 3 models. Aside from using original and transformed data, we will also be using different methods and functions such as Linear Discriminant

Analysis, step function, and logit function to enhance our models.

Below is our model definition:

-Model 1- This model will be created using all the variables in train data set with logit function GLM.

-Model 2: This model step function will be used to enhance the model 1.

-Model 3- This model will be created using classification and regression tree.

Model 1. Taking the treated data and splitting into 80/20 to train model and validate the data.

Call: glm(formula = y ~ ., family = binomial(link = “logit”), data = DS_TARGET_FLAG_TRAIN)

Deviance Residuals: Min 1Q Median 3Q Max

-6.0098 -0.2994 -0.1859 -0.1345 3.3659

Coefficients: (10 not defined because of singularities) Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.398e+12 8.695e+12 -0.161 0.872233

age -1.747e-04 2.427e-03 -0.072 0.942611

duration 4.708e-03 7.465e-05 63.068 < 2e-16 ***campaign -4.019e-02 1.155e-02 -3.479 0.000504*** pdays -3.252e-02 1.750e-02 -1.858 0.063152 .

previous -7.825e-02 6.010e-02 -1.302 0.192906

emp.var.rate -1.747e+00 1.421e-01 -12.297 < 2e-16 ***cons.price.idx 2.185e+00 2.523e-01 8.660 < 2e-16*** cons.conf.idx 2.062e-02 7.766e-03 2.656 0.007910 ** euribor3m 3.270e-01 1.300e-01 2.516 0.011861 *

nr.employed 5.312e-03 3.114e-03 1.706 0.088042 .

job_housemaid -2.413e-01 1.770e-01 -1.363 0.172833

job_services -3.316e-01 1.254e-01 -2.644 0.008188 ** job_admin. -1.836e-01 1.106e-01 -1.661 0.096793 .

`job_blue-collar` -4.495e-01 1.184e-01 -3.796 0.000147 ***job_technician***

-2.332e-01 1.178e-01 -1.980 0.047698

job_retired 7.275e-02 1.518e-01 0.479 0.631718

job_management -2.380e-01 1.331e-01 -1.788 0.073818 .

job_unemployed -1.877e-01 1.587e-01 -1.183 0.236887

`job_self-employed` -3.541e-01 1.538e-01 -2.302 0.021350 *

job_unknown -2.786e-01 2.559e-01 -1.089 0.276285

job_entrepreneur -3.734e-01 1.612e-01 -2.317 0.020513 *

job_student NA NA NA NA

marital_married -5.778e-02 4.107e-01 -0.141 0.888116

marital_single 2.694e-03 4.118e-01 0.007 0.994780

marital_divorced -5.583e-02 4.148e-01 -0.135 0.892934

marital_unknown NA NA NA NA

education_illiterate 9.115e-01 7.523e-01 1.212 0.225701

education_unknown -1.318e-02 1.009e-01 -0.131 0.896102

education_primary -1.153e-01 7.656e-02 -1.507 0.131920

education_secondary -1.390e-01 5.215e-02 -2.666 0.007677

education_tertiary NA NA NA NA

default_no 1.398e+12 8.695e+12 0.161 0.872233

default_unknown 1.398e+12 8.695e+12 0.161 0.872233

default_yes 1.398e+12 8.695e+12 0.161 0.872233

housing_no 4.444e-02 1.483e-01 0.300 0.764409

housing_yes 3.820e-02 1.472e-01 0.260 0.795179

housing_unknown NA NA NA NA

loan_no 5.304e-02 5.746e-02 0.923 0.355961

loan_yes NA NA NA NA

loan_unknown NA NA NA NA

contact_telephone -6.462e-01 7.684e-02 -8.410 $<$ 2e-16 ***contact_cellular NA NA***

NA NA

month_may -8.159e-01 1.525e-01 -5.351 8.75e-08 month_jun -8.924e-01 2.359e-01 -3.782 0.000156 ***month_jul -2.331e-01 1.755e-01 -1.329 0.183973***

month_aug 4.954e-01 1.418e-01 3.493 0.000478 month_oct -1.759e-01 1.423e-01 -1.236 0.216519

month_nov -7.888e-01 1.522e-01 -5.183 2.18e-07 ***month_dec -4.478e-02 2.118e-01 -0.211 0.832574***

month_mar 1.640e+00 1.546e-01 10.608 < 2e-16 month_apr -3.739e-01 1.795e-01 -2.083 0.037227 *

month_sep NA NA NA NA

day_of_week_mon -1.180e-01 6.609e-02 -1.785 0.074295 .

day_of_week_tue 9.455e-02 6.584e-02 1.436 0.150970

day_of_week_wed 1.742e-01 6.566e-02 2.653 0.007988 ** day_of_week_thu 5.491e-02 6.406e-02 0.857 0.391323

day_of_week_fri NA NA NA NA

previous_contact -3.122e+01 1.730e+01 -1.805 0.071107 .

poutcome_nonexistent -3.940e-01 2.274e-01 -1.733 0.083118 .

poutcome_failure -8.014e-01 2.294e-01 -3.493 0.000477 *** poutcome_success NA NA NA NA

— Signif. codes: 0 “***’ 0.001 ’’*** *0.01 ’*” 0.05 “.” 0.1 “” 1

(Dispersion parameter for binomial family taken to be 1)

`Null deviance: 28999  on 41187  degrees of freedom`

Residual deviance: 17077 on 41136 degrees of freedom AIC: 17181

Number of Fisher Scoring iterations: 25

Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

`Df Deviance Resid. Df Resid. Dev  Pr(>Chi)`

NULL 41187 28999

age 1 37.4 41186 28961 9.822e-10 ***duration 1 4904.6 41185 24057 < 2.2e-16*** campaign 1 227.8 41184 23829 < 2.2e-16 ***pdays 1 2511.1 41183 21318 < 2.2e-16*** previous 1 101.7 41182 21216 < 2.2e-16 ***emp.var.rate 1 2406.9 41181 18809 < 2.2e-16*** cons.price.idx 1 494.9 41180 18314 < 2.2e-16 ***cons.conf.idx 1 151.0 41179 18163 < 2.2e-16*** euribor3m 1 23.1 41178 18140 1.500e-06 ***nr.employed 1 24.1 41177 18116 9.241e-07*** job_housemaid 1 0.2 41176 18116 0.6320863

job_services 1 11.2 41175 18105 0.0008038 ***job_admin. 1 14.5 41174 18090 0.0001391*** `job_blue-collar` 1 94.1 41173 17996 < 2.2e-16 ***job_technician 1 0.3 41172 17996 0.6006861***

job_retired 1 20.2 41171 17976 6.965e-06 job_management 1 0.0 41170 17976 0.8657697

job_unemployed 1 0.0 41169 17976 0.9078989

`job_self-employed` 1 1.0 41168 17975 0.3229617

job_unknown 1 0.2 41167 17975 0.6957600

job_entrepreneur 1 13.5 41166 17961 0.0002383 ***job_student 0 0.0 41166 17961***

marital_married 1 5.1 41165 17956 0.0233007

marital_single 1 2.8 41164 17953 0.0946716 .

marital_divorced 1 0.1 41163 17953 0.7715131

marital_unknown 0 0.0 41163 17953

education_illiterate 1 1.7 41162 17951 0.1972689

education_unknown 1 1.1 41161 17950 0.2859474

education_primary 1 1.8 41160 17948 0.1816799

education_secondary 1 22.6 41159 17926 1.969e-06 *education_tertiary 0 0.0 41159 17926*

default_no 1 40.9 41158 17885 1.575e-10 **default_unknown 1 0.0 41157 17885 0.8273594**

default_yes 1 0.0 41156 17885 1.0000000

housing_no 1 0.0 41155 17885 0.8790128

housing_yes 1 0.3 41154 17884 0.5916891

housing_unknown 0 0.0 41154 17884

loan_no 1 1.6 41153 17883 0.2119091

loan_yes 0 0.0 41153 17883

loan_unknown 0 0.0 41153 17883

contact_telephone 1 196.8 41152 17686 < 2.2e-16 *contact_cellular 0 0.0 41152 17686*

month_may 1 225.7 41151 17460 < 2.2e-16 **month_jun 1 0.1 41150 17460 0.7043051**

month_jul 1 3.5 41149 17457 0.0627720 .

month_aug 1 24.4 41148 17432 7.654e-07 *month_oct 1 0.1 41147 17432 0.7458953*

month_nov 1 58.5 41146 17374 2.048e-14 **month_dec 1 1.2 41145 17373 0.2675276**

month_mar 1 229.1 41144 17143 < 2.2e-16 *month_apr 0 4.8 41144 17139*

month_sep 0 0.0 41144 17139

day_of_week_mon 2 15.5 41142 17123 0.0004232 **day_of_week_tue 1 0.1 41141 17123 0.7408142**

day_of_week_wed 0 6.9 41141 17116

day_of_week_thu 1 0.9 41140 17115 0.3410693

day_of_week_fri 0 0.0 41140 17115

previous_contact 2 12.6 41138 17103 0.0018813 poutcome_nonexistent 1 13.3 41137 17089 0.0002679 ***poutcome_failure 1 12.1 41136 17077 0.0005072***
poutcome_success 0 0.0 41136 17077

— Signif. codes: 0 “***’ 0.001 ’’*** *0.01 ’*” 0.05 “.” 0.1 “” 1 llh llhNull G2 McFadden r2ML -8.538582e+03 -1.449936e+04 1.192156e+04 4.111064e-01 2.513192e-01 r2CU 4.972428e-01
[1] “Accuracy 0.914056809905317”

Model 2.

Model 3.

Model Selection

Model Evaluation

Statistical analysis

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), and *pscl* (1.4.9, Zeileis, Kleiber, & Jackman, 2008) for all our analyses.

Interpretation and Disussion of Results

Discussion and Conclusions

conclude your findings, limitations, and suggest areas for future work

References

be sure to cite all references used in the report (APA format).

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} #
```

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

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

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

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).

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/