Business Analytics - Final Project

Critical Thinking Group 5
Arindam Barman
Mohamed Elmoudni
Shazia Khan
Kishore Prasad

Contents

Overview

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).

1 Data Exploration Analysis

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

1.1 Variable identification

First let's display and examine the data dictionary or the data columns as shown in table 1

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 number of days that passed by after the client was last contacted
previous	Numeric	Predictor	number of contacts performed before this campaign and for this client
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?

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 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.

1.2 Missing values and Unique Values

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 3.

Table 2: Missing Values

	Missing Values
age	0
job	0
marital	0
education	0
default	0
housing	0
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 3: Unique Values

	Unique Values
age	78
job	12
marital	4
education	8
default	3
housing	3
loan	3
contact	2
month	10
day_of_week	5
duration	1544
campaign	42

	Unique Values
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

2 Data Preparation

-Convert Binary to 0 and 1
-Create dummy variables
-Data Summary Analysis
-Correlation of Variables with y

2.1 Convert Binary yes and no to 0 and 1

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

2.2 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.

2.3 Data Summary Analysis

Table 4: Data Summary

	vars	n	mean	sd	median	trimmed	mad
age	1	41188	40.0240604	10.4212500	38.000	39.3033807	10.3782000
duration	2	41188	258.2850102	259.2792488	180.000	210.6102513	139.3644000
campaign	3	41188	2.5675925	2.7700135	2.000	1.9914118	1.4826000
pdays	4	41188	962.4754540	186.9109073	999.000	999.0000000	0.0000000
previous	5	41188	0.1729630	0.4949011	0.000	0.0457332	0.0000000
emp.var.rate	6	41188	0.0818855	1.5709597	1.100	0.2661204	0.4447800
cons.price.idx	7	41188	93.5756644	0.5788400	93.749	93.5807666	0.5633880
cons.conf.idx	8	41188	-40.5026003	4.6281979	-41.800	-40.6015356	6.5234400
euribor3m	9	41188	3.6212908	1.7344474	4.857	3.8055852	0.1601208
nr.employed	10	41188	5167.0359109	72.2515277	5191.000	5178.4253338	55.0044600
y	11	41188	0.1126542	0.3161734	0.000	0.0158412	0.0000000
job__housemaid	12	41188	0.0257357	0.1583475	0.000	0.0000000	0.0000000
job__services	13	41188	0.0963630	0.2950920	0.000	0.0000000	0.0000000

	vars	n	mean	sd	median	trimmed	mad
job__admin.	14	41188	0.2530349	0.4347560	0.000	0.1913086	0.0000000
job__blue-collar	15	41188	0.2246771	0.4173746	0.000	0.1558631	0.0000000
job__technician	16	41188	0.1637127	0.3700192	0.000	0.0796613	0.0000000
job__retired	17	41188	0.0417597	0.2000421	0.000	0.0000000	0.0000000
job__management	18	41188	0.0709916	0.2568138	0.000	0.0000000	0.0000000
job__unemployed	19	41188	0.0246188	0.1549623	0.000	0.0000000	0.0000000
job__self-employed	20	41188	0.0345003	0.1825127	0.000	0.0000000	0.0000000
job__unknown	21	41188	0.0080120	0.0891518	0.000	0.0000000	0.0000000
job__entrepreneur	22	41188	0.0353501	0.1846654	0.000	0.0000000	0.0000000
job__student	23	41188	0.0212441	0.1441986	0.000	0.0000000	0.0000000
marital__married	24	41188	0.6052248	0.4888083	1.000	0.6315246	0.0000000
marital__single	25	41188	0.2808585	0.4494240	0.000	0.2260864	0.0000000
marital__divorced	26	41188	0.1119744	0.3153387	0.000	0.0149915	0.0000000
marital__unknown	27	41188	0.0019423	0.0440294	0.000	0.0000000	0.0000000
education__basic.4y	28	41188	0.1013888	0.3018465	0.000	0.0017601	0.0000000
education__high.school	29	41188	0.2310139	0.4214864	0.000	0.1637837	0.0000000
education__basic.6y	30	41188	0.0556473	0.2292421	0.000	0.0000000	0.0000000
education__basic.9y	31	41188	0.1467660	0.3538768	0.000	0.0584790	0.0000000
education__professional.course	32	41188	0.1272944	0.3333065	0.000	0.0341406	0.0000000
education__unknown	33	41188	0.0420268	0.2006528	0.000	0.0000000	0.0000000
education__university.degree	34	41188	0.2954259	0.4562395	0.000	0.2442947	0.0000000
education__illiterate	35	41188	0.0004370	0.0209007	0.000	0.0000000	0.0000000
contact__telephone	36	41188	0.3652520	0.4815066	0.000	0.3315732	0.0000000
contact__cellular	37	41188	0.6347480	0.4815066	1.000	0.6684268	0.0000000
default__no	38	41188	0.7912013	0.4064552	1.000	0.8639840	0.0000000
default__unknown	39	41188	0.2087258	0.4064030	0.000	0.1359250	0.0000000
default__yes	40	41188	0.0000728	0.0085342	0.000	0.0000000	0.0000000
housing__no	41	41188	0.4521220	0.4977085	0.000	0.4401554	0.0000000
housing__yes	42	41188	0.5238419	0.4994373	1.000	0.5298009	0.0000000
housing__unknown	43	41188	0.0240361	0.1531632	0.000	0.0000000	0.0000000
loan__no	44	41188	0.8242692	0.3805956	1.000	0.9053168	0.0000000
loan__yes	45	41188	0.1516947	0.3587290	0.000	0.0646395	0.0000000
loan__unknown	46	41188	0.0240361	0.1531632	0.000	0.0000000	0.0000000
poutcome__nonexistent	47	41188	0.8634311	0.3433958	1.000	0.9542668	0.0000000
poutcome__failure	48	41188	0.1032340	0.3042679	0.000	0.0040665	0.0000000
poutcome__success	49	41188	0.0333350	0.1795119	0.000	0.0000000	0.0000000
month__may	50	41188	0.3342964	0.4717496	0.000	0.2928806	0.0000000
month__jun	51	41188	0.1291153	0.3353316	0.000	0.0364166	0.0000000
month__jul	52	41188	0.1741769	0.3792662	0.000	0.0927410	0.0000000
month__aug	53	41188	0.1499951	0.3570710	0.000	0.0625152	0.0000000
month__oct	54	41188	0.0174323	0.1308770	0.000	0.0000000	0.0000000
month__nov	55	41188	0.0995678	0.2994265	0.000	0.0000000	0.0000000
month__dec	56	41188	0.0044188	0.0663276	0.000	0.0000000	0.0000000
month__mar	57	41188	0.0132563	0.1143717	0.000	0.0000000	0.0000000
month__apr	58	41188	0.0639021	0.2445814	0.000	0.0000000	0.0000000
month__sep	59	41188	0.0138390	0.1168238	0.000	0.0000000	0.0000000
day__of__week__mon	60	41188	0.2067107	0.4049511	0.000	0.1334062	0.0000000
day__of__week__tue	61	41188	0.1964164	0.3972919	0.000	0.1205390	0.0000000
day__of__week__wed	62	41188	0.1974847	0.3981059	0.000	0.1218742	0.0000000
day__of__week__thu	63	41188	0.2093571	0.4068547	0.000	0.1367140	0.0000000
day__of__week__fri	64	41188	0.1900311	0.3923302	0.000	0.1125577	0.0000000
previous__contact	65	41188	0.0367826	0.1882298	0.000	0.0000000	0.0000000

Table 5: Data Summary (Cont)

	min	max	range	skew	kurtosis	se
age	17.000	98.000	81.000	0.7846397	0.7908857	0.0513493
duration	0.000	4918.000	4918.000	3.2629036	20.2442057	1.2775632
campaign	1.000	56.000	55.000	4.7621598	36.9732194	0.0136489
pdays	0.000	999.000	999.000	-4.9218314	22.2253936	0.9209781
previous	0.000	7.000	7.000	3.8317631	20.1051076	0.0024386
emp.var.rate	-3.400	1.400	4.800	-0.7240428	-1.0627423	0.0077407
cons.price.idx	92.201	94.767	2.566	-0.2308708	-0.8299589	0.0028522
cons.conf.idx	-50.800	-26.900	23.900	0.3031578	-0.3587887	0.0228048
euribor3m	0.634	5.045	4.411	-0.7091363	-1.4068549	0.0085463
nr.employed	4963.600	5228.100	264.500	-1.0441863	-0.0040511	0.3560096
y	0.000	1.000	1.000	2.4501517	4.0033404	0.0015579
job__housemaid	0.000	1.000	1.000	5.9900255	33.8812283	0.0007802
job__services	0.000	1.000	1.000	2.7356021	5.4836522	0.0014540
job__admin.	0.000	1.000	1.000	1.1360815	-0.7093361	0.0021422
job__blue-collar	0.000	1.000	1.000	1.3192765	-0.2595158	0.0020566
job__technician	0.000	1.000	1.000	1.8176306	1.3038128	0.0018232
job__retired	0.000	1.000	1.000	4.5813276	18.9890235	0.0009857
job__management	0.000	1.000	1.000	3.3409260	9.1620092	0.0012654
job__unemployed	0.000	1.000	1.000	6.1352936	35.6426931	0.0007636
job__self-employed	0.000	1.000	1.000	5.1008881	24.0196428	0.0008993
job__unknown	0.000	1.000	1.000	11.0368168	119.8142342	0.0004393
job__entrepreneur	0.000	1.000	1.000	5.0322224	23.3238288	0.0009099
job__student	0.000	1.000	1.000	6.6400673	42.0915155	0.0007105
marital__married	0.000	1.000	1.000	-0.4305257	-1.8146917	0.0024085
marital__single	0.000	1.000	1.000	0.9751869	-1.0490361	0.0022145
marital__divorced	0.000	1.000	1.000	2.4609486	4.0563667	0.0015538
marital__unknown	0.000	1.000	1.000	22.6233213	509.8270434	0.0002169
education__basic.4y	0.000	1.000	1.000	2.6410882	4.9754678	0.0014873
education__high.school	0.000	1.000	1.000	1.2763381	-0.3709700	0.0020768
education__basic.6y	0.000	1.000	1.000	3.8766175	13.0284796	0.0011296
education__basic.9y	0.000	1.000	1.000	1.9963181	1.9853343	0.0017437
education__professional.course	0.000	1.000	1.000	2.2363598	3.0013781	0.0016423
education__unknown	0.000	1.000	1.000	4.5647225	18.8371487	0.0009887
education__university.degree	0.000	1.000	1.000	0.8967623	-1.1958465	0.0022481
education__illiterate	0.000	1.000	1.000	47.8022616	2283.1116468	0.0001030
contact__telephone	0.000	1.000	1.000	0.5596796	-1.6867997	0.0023726
contact__cellular	0.000	1.000	1.000	-0.5596796	-1.6867997	0.0023726
default__no	0.000	1.000	1.000	-1.4328481	0.0530549	0.0020028
default__unknown	0.000	1.000	1.000	1.4333905	0.0546097	0.0020025
default__yes	0.000	1.000	1.000	117.1551691	13723.6668447	0.0000421
housing__no	0.000	1.000	1.000	0.1923892	-1.9630341	0.0024524
housing__yes	0.000	1.000	1.000	-0.0954727	-1.9909333	0.0024609
housing__unknown	0.000	1.000	1.000	6.2149702	36.6267442	0.0007547
loan__no	0.000	1.000	1.000	-1.7039679	0.9035286	0.0018753
loan__yes	0.000	1.000	1.000	1.9418382	1.7707787	0.0017676
loan__unknown	0.000	1.000	1.000	6.2149702	36.6267442	0.0007547
poutcome__nonexistent	0.000	1.000	1.000	-2.1166376	2.4802150	0.0016920
poutcome__failure	0.000	1.000	1.000	2.6079414	4.8014749	0.0014992
poutcome__success	0.000	1.000	1.000	5.1991402	25.0316666	0.0008845
month__may	0.000	1.000	1.000	0.7024895	-1.5065451	0.0023245

	min	max	range	skew	kurtosis	se
month_jun	0.000	1.000	1.000	2.2119941	2.8929884	0.0016523
month_jul	0.000	1.000	1.000	1.7181345	0.9520092	0.0018688
month_aug	0.000	1.000	1.000	1.9603741	1.8431112	0.0017594
month_oct	0.000	1.000	1.000	7.3741903	52.3799548	0.0006449
month_nov	0.000	1.000	1.000	2.6745954	5.1535859	0.0014754
month_dec	0.000	1.000	1.000	14.9430876	221.3012387	0.0003268
month_mar	0.000	1.000	1.000	8.5114073	70.4457653	0.0005636
month_apr	0.000	1.000	1.000	3.5659885	10.7165344	0.0012051
month_sep	0.000	1.000	1.000	8.3227782	67.2702700	0.0005756
day_of_week_mon	0.000	1.000	1.000	1.4484821	0.0981028	0.0019953
day_of_week_tue	0.000	1.000	1.000	1.5282275	0.3354874	0.0019576
day_of_week_wed	0.000	1.000	1.000	1.5197359	0.3096048	0.0019616
day_of_week_thu	0.000	1.000	1.000	1.4286962	0.0411737	0.0020047
day_of_week_fri	0.000	1.000	1.000	1.5801046	0.4967426	0.0019332
previous_contact	0.000	1.000	1.000	4.9217092	22.2237610	0.0009275

2.4 Correlation of Variables with y

Now we will produce the correlation table between the independent variables and the dependent variable

Table 6: Variable 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_university.degree	0.0503638
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
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

education_professional.course	0.0010032
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
education_high.school	-0.0074525
month_aug	-0.0088126
month_jun	-0.0091818
marital_divorced	-0.0106080
education_basic.4y	-0.0107980
housing_no	-0.0110852
month_nov	-0.0117959
job_entrepreneur	-0.0166439
day_of_week_mon	-0.0212649
education_basic.6y	-0.0235168
month_jul	-0.0322301
job_services	-0.0323009
marital_married	-0.0433978
education_basic.9y	-0.0451351
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