1 . Output variable -> y

y -> Whether the client has subscribed a term deposit or not

Binomial ("yes" or "no")

Solution:

**Business Problem:** To predict whether the client has subscribed a term deposit or not

**Datasets:**

Independent Variable(x):

x1: age

x2: job

x3: marital

x4:education

x5:default

x6:balance

x7:housing

x8:loan

x9:contact

x10:day

x11:month

x12:duration

x13:campaign

x14:pdays

x15:previous

x16: poutcome

Dependent Variable(y): y

**EDA :**

1.) Initailly summary and structure of the entire datasets is analysed .

The output variable “y”has two class namely, “Yes” and “No”

Input variables:- “age”, “balance” , “day”,”duration”,”campaign”,”pdays”,”prevoius” are numerical variables

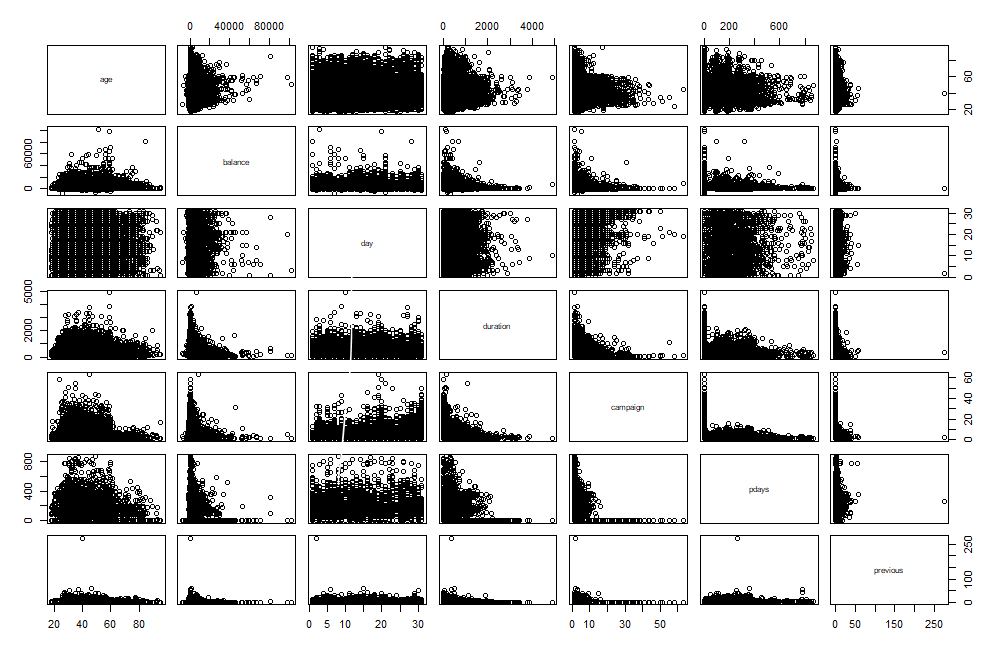
The categorical variables are : “job”,”maritial”,”education”,”default”,” housing”,”loan”,”contact”,”month”,”poutcome”

. The above numerical variables are in class(integer) and categorical variables are in class(factor), so any further conversion is not required.

2.) Analysis of missing values

There is no missing values in the datasets for any of the variables.

3.) Pair plots of numerical data



4.) Co-relation

age balance day duration campaign

age 1.000000000 0.097782739 -0.009120046 -0.004648428 0.004760312

balance 0.097782739 1.000000000 0.004502585 0.021560380 -0.014578279

day -0.009120046 0.004502585 1.000000000 -0.030206341 0.162490216

duration -0.004648428 0.021560380 -0.030206341 1.000000000 -0.084569503

campaign 0.004760312 -0.014578279 0.162490216 -0.084569503 1.000000000

pdays -0.023758014 0.003435322 -0.093044074 -0.001564770 -0.088627668

previous 0.001288319 0.016673637 -0.051710497 0.001203057 -0.032855290

pdays previous

age -0.023758014 0.001288319

balance 0.003435322 0.016673637

day -0.093044074 -0.051710497

duration -0.001564770 0.001203057

campaign -0.088627668 -0.032855290

pdays 1.000000000 0.454819635

previous 0.454819635 1.000000000

The co-relation coefficient value is poor among the variables

4.) Changing response variable into binary format

Replaced “Yes” with 1 and “No” with 0

5.) Data Partitioning

Dividing the datasets into train and test data set

Training datasets = 70% of the bank\_data

Train datasets = Remaining 30 % of the bank\_data

6.) Handling Imbalanced datasets

As there is unequal amount of data in two categories “yes” and “no”, for accuracy it is required to equalize both the category. Here the process used is Oversampling where the minority class category datasets is increased to match the data size of majority class category.

**Model Building**

Model Summary

Call:

glm(formula = y ~ ., family = "binomial", data = up\_train)

Deviance Residuals:

Min 1Q Median 3Q Max

-7.2710 -0.5842 -0.0069 0.5985 2.9830

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.773814147 0.121110831 -6.389 0.0000000001666404 \*\*\*

age 0.000600369 0.001446630 0.415 0.678133

jobblue-collar -0.356159947 0.047381410 -7.517 0.0000000000000561 \*\*\*

jobentrepreneur -0.443189415 0.079639744 -5.565 0.0000000262261179 \*\*\*

jobhousemaid -0.635333458 0.091912945 -6.912 0.0000000000047673 \*\*\*

jobmanagement -0.196161436 0.048868832 -4.014 0.0000596882730190 \*\*\*

jobretired 0.214363785 0.067241513 3.188 0.001433 \*\*

jobself-employed-0.400797443 0.074079408 -5.410 0.0000000628926079 \*\*\*

jobservices -0.235260931 0.054405099 -4.324 0.0000153055721567 \*\*\*

jobstudent 0.305102347 0.080312124 3.799 0.000145 \*\*\*

jobtechnician -0.174235470 0.045363677 -3.841 0.000123 \*\*\*

jobunemployed -0.250863091 0.077037302 -3.256 0.001128 \*\*

jobunknown -0.745323315 0.168357499 -4.427 0.0000095540457685 \*\*\*

maritalmarried -0.131529194 0.039012859 -3.371 0.000748 \*\*\*

maritalsingle 0.215449176 0.044720971 4.818 0.0000014527207783 \*\*\*

educationsecondary 0.243853598 0.042525472 5.734 0.0000000097918829 \*\*\*

educationtertiary 0.514688750 0.050041238 10.285 < 0.0000000000000002 \*\*\*

educationunknown 0.461798075 0.069088910 6.684 0.0000000000232328 \*\*\*

defaultyes -0.080807140 0.102956321 -0.785 0.432531

balance 0.000021840 0.000004004 5.455 0.0000000490146951 \*\*\*

housingyes -0.879573554 0.028247825 -31.138 < 0.0000000000000002 \*\*\*

loanyes -0.603953465 0.038344194 -15.751 < 0.0000000000000002 \*\*\*

contacttelephone -0.125726077 0.049895396 -2.520 0.011742 \*

contactunknown -1.680145137 0.043567648 -38.564 < 0.0000000000000002 \*\*\*

day 0.003774392 0.001615352 2.337 0.019461 \*

monthaug -0.953759235 0.050315176 -18.956 < 0.0000000000000002 \*\*\*

monthdec 0.753034740 0.142770910 5.274 0.0000001331715189 \*\*\*

monthfeb -0.258575982 0.058738552 -4.402 0.0000107182841220 \*\*\*

monthjan -1.394400482 0.077188580 -18.065 < 0.0000000000000002 \*\*\*

monthjul -1.117605700 0.050831013 -21.987 < 0.0000000000000002 \*\*\*

monthj 0.143054244 0.059852755 2.390 0.016844 \*

monthmar 1.695802053 0.096276748 17.614 < 0.0000000000000002 \*\*\*

monthmay -0.645518206 0.048369589 -13.346 < 0.0000000000000002 \*\*\*

monthnov -0.993699198 0.055213341 -17.997 < 0.0000000000000002 \*\*\*

monthoct 1.274921029 0.085510131 14.910 < 0.0000000000000002 \*\*\*

monthsep 1.045241073 0.095765449 10.915 < 0.0000000000000002 \*\*\*

duration 0.005921851 0.000059775 99.069 < 0.0000000000000002 \*\*\*

campaign -0.102695614 0.006149953 -16.699 < 0.0000000000000002 \*\*\*

pdays -0.000293829 0.000200875 -1.463 0.143537

previous 0.023259916 0.006737995 3.452 0.000556 \*\*\*

poutcomeother 0.220154307 0.060560205 3.635 0.000278 \*\*\*

poutcomesuccess 2.503480742 0.069239887 36.157 < 0.0000000000000002 \*\*\*

poutcomeunknown -0.240142651 0.063384158 -3.789 0.000151 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 77483 on 55891 degrees of freedom

Residual deviance: 44167 on 55849 degrees of freedom

AIC: 44253

Number of Fisher Scoring iterations: 6

From the model summary, most of the variables p-values <0.05 therefore they are significant except age and pdays.

As the value recorded in summary is in log form , exponential is required to get the actual values

After doing exponential of co-efficients, following is the output

Model is

Y = 0.4612504 + 1.0006005 age + 0.7003606 jobblue-collar + 0.6419856 jobentrepreneur + 0.5297588 jobhousemaid + 0.8218795 jobmanagement + 1.2390733 jobretired + 0.6697857 jobself-employed + 0.7903646 jobservices + 1.3567639 jobstudent + 0.8400991 jobtechnician + 0.7781289 jobunemployed + 0.4745808 jobunknown + 0.8767537 maritalmarried + 1.2404189 maritalsingle + 1.2761575 educationsecondary + 1.6731177 educationtertiary + 1.5869248 educationunknown + 0.9223716 defaultyes + 1.0000218balance + 0.4149598housingyes + 0.5466462loanyes + 0.8818564contacttelephone + 0.1863469 contactunknown + 1.0037815 day + 0.3852899 monthaug + 2.1234343monthdec + 0.7721504monthfeb + 0.2479817monthjan + 0.3270619monthjul + 1.1537924monthjun + 5.4510162 monthmar + 0.5243907monthmay+ 0.3702047 monthnov+ 3.5784188monthoct+ 2.8440841monthsep+ 1.0059394duration+ 0.9024016campaign+ 0.9997062pdays+ 1.0235325 previous + 1.2462690 poutcomeother + 12.2249720 poutcomesuccess + 0.7865157 poutcomeunknown

**Predicting variables :** Passing the testdata in predict function w.r.t. model , we get predicted values.

**Confusion Matrix**

0 1

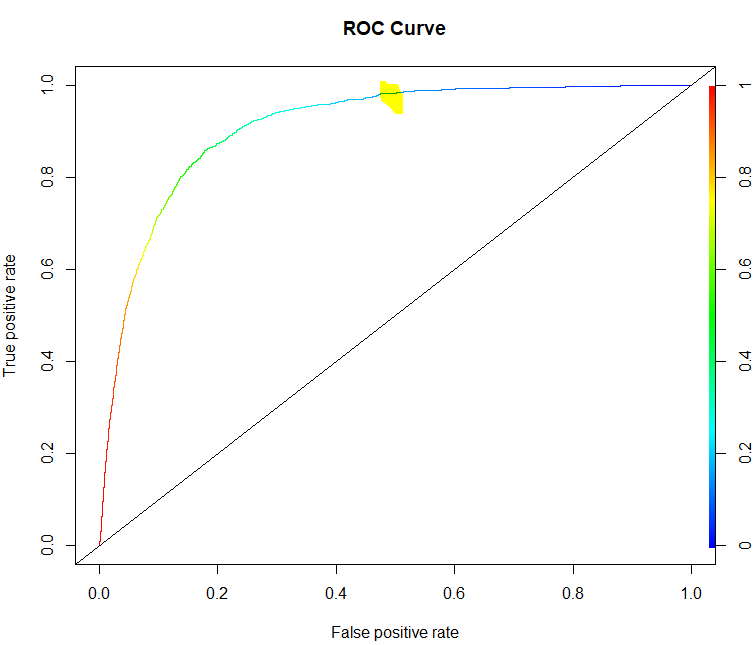
FALSE 10073 267

TRUE 1903 1319

**Accuracy** = 83.99 =84%

Error = 1-0.84 =16 %

**AOC curve**



2.) Classify whether application accepted or not using Logistic regression

Card Factor: Was the application for a credit card accepted?

Reports : Number of major derogatory reports.

Age : Age in years plus twelfths of a year.

Income :Yearly income (in USD 10,000).

Share : Ratio of monthly credit card expenditure to yearly income.

Expenditure : Average monthly credit card expenditure.

Owner : Factor. Does the individual own their home?

Selfemp : Factor. Is the individual self-employed?

Dependents :Number of dependents.

Months: Months living at current address.

Majorcards : Number of major credit cards held.

Active :Number of active credit accounts.

Solution:

**Business Problem:** To check if the application for a credit card accepted or not.

**Datasets:**

Independent Variables:

x1: reports

x2: age

x3: income

x4 : share

x5 : expenditure

x6 : owner

x7: selfemp

x8 :dependents

x9 :months

x10: majorcards

x11: active

Dependent Variable (y): card (discrete)

**EDA :**

1.) Initailly summary and structure of the entire datasets is analysed .

The output variable “card”has two class namely, “Yes” and “No”

Input variables:- “reports”, “age” , “income”, ”share”, ”expenditure”, ”dependents”, ”months”, “majorcards”,”active”are numerical variables

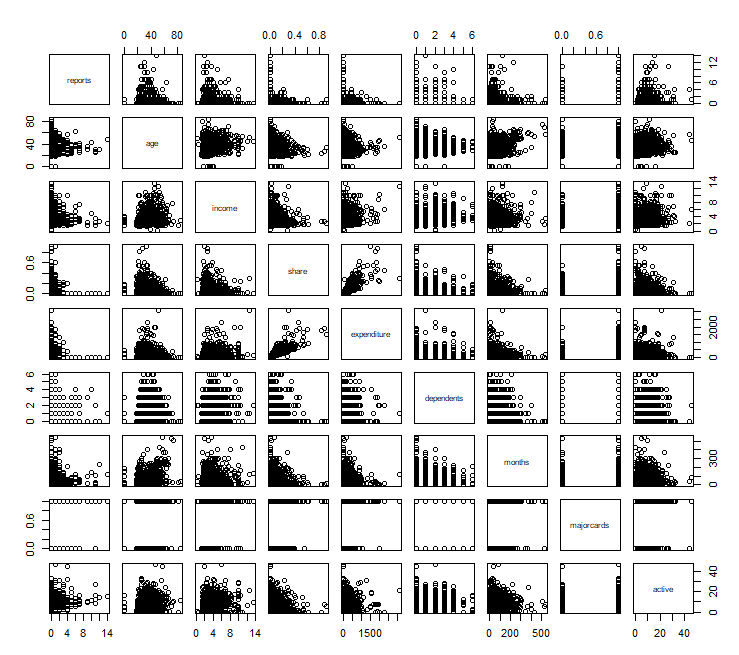
The categorical variables are : “owner”,”selfemp”.

The above numerical variables are in class(integer) or (number)and categorical variables are in class(factor), so any further conversion is not required.

2.) Analysis of missing values

There is no missing values in the datasets for any of the variables.

3.) Pair plots of numerical data



4.) Co-relation

reports age income share expenditure dependents

reports 1.000000000 0.044088513 0.01102287 -0.15901079 -0.13653760 0.01973090

age 0.044088513 1.000000000 0.32465320 -0.11569704 0.01494770 0.21214643

income 0.011022871 0.324653199 1.00000000 -0.05442926 0.28110402 0.31760130

share -0.159010789 -0.115697038 -0.05442926 1.00000000 0.83877932 -0.08261776

expenditure -0.136537597 0.014947698 0.28110402 0.83877932 1.00000000 0.05266406

dependents 0.019730896 0.212146432 0.31760130 -0.08261776 0.05266406 1.00000000

months 0.048967618 0.436425540 0.13034627 -0.05534756 -0.02900660 0.04651197

majorcards -0.007303561 0.009776687 0.10713778 0.05146956 0.07751381 0.01028454

active 0.207755016 0.181069715 0.18054026 -0.02347440 0.05472424 0.10713276

months majorcards active

reports 0.04896762 -0.007303561 0.20775502

age 0.43642554 0.009776687 0.18106971

income 0.13034627 0.107137782 0.18054026

share -0.05534756 0.051469560 -0.02347440

expenditure -0.02900660 0.077513810 0.05472424

dependents 0.04651197 0.010284541 0.10713276

months 1.00000000 -0.041446883 0.10002764

majorcards -0.04144688 1.000000000 0.11960278

active 0.10002764 0.119602777 1.00000000

From the above variable, we could say that there is no strong co-relation among the variables.

5.) Changing response variable into binary format

Replaced “Yes” with 1 and “No” with 0

6.) Data Partitioning

Dividing the datasets into train and test data set

Training datasets = 70% of the credit data

Train datasets = Remaining 30 % of the credit data

7.) Handling Imbalanced datasets

As there is unequal amount of data in two categories “yes” and “no”, for accuracy it is required to equalize both the category. Here the process used is Oversampling where the minority class category datasets is increased to match the data size of majority class category.

**Model Building**

Model Summary

Call:

glm(formula = card ~ ., family = "binomial", data = up\_train)

Deviance Residuals:

Min 1Q Median 3Q Max

-8.49 0.00 0.00 0.00 8.49

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -900353663034658 7576562 -118834073 <0.0000000000000002 \*\*\*

reports -355892193504459 991870 -358809333 <0.0000000000000002 \*\*\*

age 11768868009317 214171 54950749 <0.0000000000000002 \*\*\*

income 51574849909611 1375421 37497499 <0.0000000000000002 \*\*\*

share 22165631002023120 46694463 474695065 <0.0000000000000002 \*\*\*

expenditure 956580967706 17069 56040776 <0.0000000000000002 \*\*\*

owneryes 258478999955193 4221276 61232428 <0.0000000000000002 \*\*\*

selfempyes -288750799102017 6817829 -42352311 <0.0000000000000002 \*\*\*

dependents -40271855105383 1560839 -25801423 <0.0000000000000002 \*\*\*

months -225815193996 30485 -7407520 <0.0000000000000002 \*\*\*

majorcards 111951026734721 4439087 25219382 <0.0000000000000002 \*\*\*

active 31175807831365 303666 102664792 <0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1987.9 on 1433 degrees of freedom

Residual deviance: 9659.7 on 1422 degrees of freedom

AIC: 9683.7

Number of Fisher Scoring iterations: 25

From the model summary, all the variables p-values <0.05 therefore they are significant .

As the value recorded in summary is in log form , exponential is required to get the actual values

**Predicting variables:** Passing the testdata in predict function w.r.t. model , we get predicted values.

**Confusion Matrix**

0 1

FALSE 78 12

TRUE 10 294

**Accuracy** = 0.944 or 94%

Error = 1-0.944 = 0.0588 = 0.6 or 6%

**AOC curve**

