Output variable -> y

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

Binomial ("yes" or "no")

Ans:

> bank <- read.csv("bank-full.csv", **sep = ";"**)

> head(bank)

age job marital education default balance housing loan contact day month duration campaign

1 58 management married tertiary no 2143 yes no unknown 5 may 261 1

2 44 technician single secondary no 29 yes no unknown 5 may 151 1

3 33 entrepreneur married secondary no 2 yes yes unknown 5 may 76 1

4 47 blue-collar married unknown no 1506 yes no unknown 5 may 92 1

5 33 unknown single unknown no 1 no no unknown 5 may 198 1

6 35 management married tertiary no 231 yes no unknown 5 may 139 1

pdays previous poutcome y

1 -1 0 unknown no

2 -1 0 unknown no

3 -1 0 unknown no

4 -1 0 unknown no

5 -1 0 unknown no

6 -1 0 unknown no

> sum(is.na(bank))

[1] 0

**No NA values, no need of Imputation.**

> str(bank)

'data.frame': 45211 obs. of 17 variables:

$ age : int 58 44 33 47 33 35 28 42 58 43 ...

$ job : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...

$ marital : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...

$ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...

$ default : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...

$ balance : int 2143 29 2 1506 1 231 447 2 121 593 ...

$ housing : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...

$ loan : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...

$ contact : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...

$ day : int 5 5 5 5 5 5 5 5 5 5 ...

$ month : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...

$ duration : int 261 151 76 92 198 139 217 380 50 55 ...

$ campaign : int 1 1 1 1 1 1 1 1 1 1 ...

$ pdays : int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ previous : int 0 0 0 0 0 0 0 0 0 0 ...

$ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...

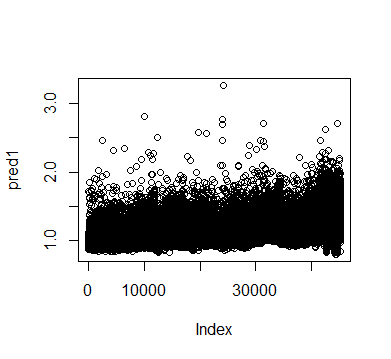
$ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

**Building LR model:**

> lmmodel.bank <- lm(y ~ ., data = bank)

> pred1 <- predict(lmmodel.bank,bank)

> plot(pred1)



**From plot LR model cannot be used to classify the data.**

**So we will build the Glm model.**

> model.bank <- glm(y~., data = bank,family = binomial)

> summary(model.bank)

Call:

glm(formula = y ~ ., family = binomial, data = bank)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.7286 -0.3744 -0.2530 -0.1502 3.4288

Coefficients:

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

(Intercept) -2.536e+00 1.837e-01 -13.803 < 2e-16 \*\*\*

age 1.127e-04 2.205e-03 0.051 0.959233

jobblue-collar -3.099e-01 7.267e-02 -4.264 2.01e-05 \*\*\*

jobentrepreneur -3.571e-01 1.256e-01 -2.844 0.004455 \*\*

jobhousemaid -5.040e-01 1.365e-01 -3.693 0.000221 \*\*\*

jobmanagement -1.653e-01 7.329e-02 -2.255 0.024130 \*

jobretired 2.524e-01 9.722e-02 2.596 0.009436 \*\*

jobself-employed -2.983e-01 1.120e-01 -2.664 0.007726 \*\*

jobservices -2.238e-01 8.406e-02 -2.662 0.007763 \*\*

jobstudent 3.821e-01 1.090e-01 3.505 0.000457 \*\*\*

jobtechnician -1.760e-01 6.893e-02 -2.554 0.010664 \*

jobunemployed -1.767e-01 1.116e-01 -1.583 0.113456

jobunknown -3.133e-01 2.335e-01 -1.342 0.179656

maritalmarried -1.795e-01 5.891e-02 -3.046 0.002318 \*\*

maritalsingle 9.250e-02 6.726e-02 1.375 0.169066

educationsecondary 1.835e-01 6.479e-02 2.833 0.004618 \*\*

educationtertiary 3.789e-01 7.532e-02 5.031 4.88e-07 \*\*\*

educationunknown 2.505e-01 1.039e-01 2.411 0.015915 \*

defaultyes -1.668e-02 1.628e-01 -0.102 0.918407

balance 1.283e-05 5.148e-06 2.493 0.012651 \*

housingyes -6.754e-01 4.387e-02 -15.395 < 2e-16 \*\*\*

loanyes -4.254e-01 5.999e-02 -7.091 1.33e-12 \*\*\*

contacttelephone -1.634e-01 7.519e-02 -2.173 0.029784 \*

contactunknown -1.623e+00 7.317e-02 -22.184 < 2e-16 \*\*\*

day 9.969e-03 2.497e-03 3.993 6.53e-05 \*\*\*

monthaug -6.939e-01 7.847e-02 -8.842 < 2e-16 \*\*\*

monthdec 6.911e-01 1.767e-01 3.912 9.17e-05 \*\*\*

monthfeb -1.473e-01 8.941e-02 -1.648 0.099427 .

monthjan -1.262e+00 1.217e-01 -10.367 < 2e-16 \*\*\*

monthjul -8.308e-01 7.740e-02 -10.733 < 2e-16 \*\*\*

monthjun 4.536e-01 9.367e-02 4.843 1.28e-06 \*\*\*

monthmar 1.590e+00 1.199e-01 13.265 < 2e-16 \*\*\*

monthmay -3.991e-01 7.229e-02 -5.521 3.36e-08 \*\*\*

monthnov -8.734e-01 8.441e-02 -10.347 < 2e-16 \*\*\*

monthoct 8.814e-01 1.080e-01 8.159 3.37e-16 \*\*\*

monthsep 8.741e-01 1.195e-01 7.314 2.58e-13 \*\*\*

duration 4.194e-03 6.453e-05 64.986 < 2e-16 \*\*\*

campaign -9.078e-02 1.014e-02 -8.955 < 2e-16 \*\*\*

pdays -1.027e-04 3.061e-04 -0.335 0.737268

previous 1.015e-02 6.503e-03 1.561 0.118476

poutcomeother 2.035e-01 8.986e-02 2.265 0.023543 \*

poutcomesuccess 2.291e+00 8.235e-02 27.821 < 2e-16 \*\*\*

poutcomeunknown -9.179e-02 9.347e-02 -0.982 0.326093

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 32631 on 45210 degrees of freedom

Residual deviance: 21562 on 45168 degrees of freedom

AIC: 21648

Number of Fisher Scoring iterations: 6

> prob <- predict(model.bank,bank,type="response")

> confusion<-table(prob>0.5,bank$y)

> confusion

no yes

FALSE 38940 3456

TRUE 982 1833

> Accuracy<-sum(diag(confusion)/sum(confusion))

**> Accuracy**

**[1] 0.901838**

**Accuracy is 90% so no need to adjust cut off values.**