Methodology:

We will build a Logistic regression model to predict the response variable “y” (whether the client subscribed to a term deposit or No.)

Step 1: Imputing NA values in the datasets.

Step 2:Data Preparation: Grouping similar category variables and making dummies.

Step 3: Model Building( LOGISTIC REGRESSION )

Step 4. Finding Cutoff value and Perfomance measurements of the model.(Sensitivity, Specificity, Accuracy)

Step 5.Predict the final output on test dataset.(whether the client subscribe or no to term deposit)

Step 6:Creating confusion matrix and finding how good our model is. (by predicting on test\_25 dataset)

Initial setup

loading library dplyr

**library**(dplyr)

setwd("C:\\Users\\INS15R\\Documents\\R latest\\R EDVANCER\\Industry Based Projects\\Industry-Based-Projects-Edvancer-Eduventures")

getwd()

## [1] "C:/Users/INS15R/Documents/R latest/R EDVANCER/Industry Based Projects/Industry-Based-Projects-Edvancer-Eduventures"

Reading train and test datasets:

train=read.csv("bank-full\_train.csv",stringsAsFactors = FALSE,header = T ) *#31647,18*

test=read.csv("bank-full\_test.csv",stringsAsFactors = FALSE,header = T ) *#13564,17*

Step 1: Imputing NA values in the datasets.

apply(train,2,**function**(x)sum(is.na(x)))

## age job marital education default balance housing

## 0 0 0 0 0 0 0

## loan contact day month duration campaign pdays

## 0 0 0 0 0 0 0

## previous poutcome ID y

## 0 0 0 0

There exist no NA values in train dataset.

apply(test,2,**function**(x)sum(is.na(x)))

## age job marital education default balance housing

## 0 0 0 0 0 0 0

## loan contact day month duration campaign pdays

## 0 0 0 0 0 0 0

## previous poutcome ID

## 0 0 0

There exist no NA values in test dataset.

Step 2:Data Preparation

Combining both train n test datasets prior to data preparation.

test$y=NA

train$data='train'

test$data='test'

all\_data=rbind(train,test)

apply(all\_data,2,**function**(x)sum(is.na(x)))

## age job marital education default balance housing

## 0 0 0 0 0 0 0

## loan contact day month duration campaign pdays

## 0 0 0 0 0 0 0

## previous poutcome ID y data

## 0 0 0 13564 0

Lets see the structure and datatypes of the combined dataset.

glimpse(all\_data) *#45211,19var*

## Observations: 45,211

## Variables: 19

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37, ...

## $ job <chr> "blue-collar", "admin.", "technician", "self-employe...

## $ marital <chr> "married", "divorced", "divorced", "married", "marri...

## $ education <chr> "secondary", "secondary", "secondary", "tertiary", "...

## $ default <chr> "no", "no", "no", "no", "no", "no", "yes", "no", "no...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 387,...

## $ housing <chr> "no", "no", "no", "no", "yes", "yes", "yes", "no", "...

## $ loan <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "no...

## $ contact <chr> "cellular", "cellular", "cellular", "cellular", "unk...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 21,...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "ju...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 158...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1, ...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, -1...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 0...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unknown...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 1231...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", "n...

## $ data <chr> "train", "train", "train", "train", "train", "train"...

Creating dummy variables by combining similar categories for variable job(char type)

t=table(all\_data$job)

sort(t)

##

## unknown student housemaid unemployed entrepreneur

## 288 938 1240 1303 1487

## self-employed retired services admin. technician

## 1579 2264 4154 5171 7597

## management blue-collar

## 9458 9732

final=round(prop.table(table(all\_data$job,all\_data$y),1)\*100,1)

final

##

## no yes

## admin. 87.9 12.1

## blue-collar 92.8 7.2

## entrepreneur 90.7 9.3

## housemaid 92.4 7.6

## management 86.3 13.7

## retired 76.8 23.2

## self-employed 87.9 12.1

## services 91.0 9.0

## student 70.8 29.2

## technician 88.6 11.4

## unemployed 85.1 14.9

## unknown 88.2 11.8

s=addmargins(final,2) *#add margin across Y*

sort(s[,1])

## student retired unemployed management admin.

## 70.8 76.8 85.1 86.3 87.9

## self-employed unknown technician entrepreneur services

## 87.9 88.2 88.6 90.7 91.0

## housemaid blue-collar

## 92.4 92.8

View(s)

all\_data=all\_data %>%

mutate(job\_1=as.numeric(job %**in**% c("self-employed","unknown","technician")),

job\_2=as.numeric(job %**in**% c("services","housemaid","entrepreneur")),

job\_3=as.numeric(job %**in**% c("management","admin")),

job\_4=as.numeric(job=="student"),

job\_5=as.numeric(job=="retired"),

job\_6=as.numeric(job=="unemployed")) %>%

select(-job)

glimpse(all\_data)

## Observations: 45,211

## Variables: 24

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37, ...

## $ marital <chr> "married", "divorced", "divorced", "married", "marri...

## $ education <chr> "secondary", "secondary", "secondary", "tertiary", "...

## $ default <chr> "no", "no", "no", "no", "no", "no", "yes", "no", "no...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 387,...

## $ housing <chr> "no", "no", "no", "no", "yes", "yes", "yes", "no", "...

## $ loan <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "no...

## $ contact <chr> "cellular", "cellular", "cellular", "cellular", "unk...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 21,...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "ju...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 158...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1, ...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, -1...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 0...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unknown...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 1231...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", "n...

## $ data <chr> "train", "train", "train", "train", "train", "train"...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

Making dummies for variable marital

t=table(all\_data$marital)

sort(t)

##

## divorced single married

## 5207 12790 27214

all\_data=all\_data %>%

mutate(divorced=as.numeric(marital %**in**% c("divorced")),

single=as.numeric(marital %**in**% c("single"))

) %>%

select(-marital)

glimpse(all\_data)

## Observations: 45,211

## Variables: 25

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37, ...

## $ education <chr> "secondary", "secondary", "secondary", "tertiary", "...

## $ default <chr> "no", "no", "no", "no", "no", "no", "yes", "no", "no...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 387,...

## $ housing <chr> "no", "no", "no", "no", "yes", "yes", "yes", "no", "...

## $ loan <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "no...

## $ contact <chr> "cellular", "cellular", "cellular", "cellular", "unk...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 21,...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "ju...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 158...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1, ...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, -1...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 0...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unknown...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 1231...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", "n...

## $ data <chr> "train", "train", "train", "train", "train", "train"...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0...

Making dummies for variable education

t=table(all\_data$education)

sort(t)

##

## unknown primary tertiary secondary

## 1857 6851 13301 23202

all\_data=all\_data %>%

mutate(edu\_primary=as.numeric(education %**in**% c("primary")),

edu\_sec=as.numeric(education %**in**% c("secondary")),

edu\_tert=as.numeric(education %**in**% c("tertiary"))

) %>%

select(-education)

glimpse(all\_data)

## Observations: 45,211

## Variables: 27

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <chr> "no", "no", "no", "no", "no", "no", "yes", "no", "...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <chr> "no", "no", "no", "no", "yes", "yes", "yes", "no",...

## $ loan <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "...

## $ contact <chr> "cellular", "cellular", "cellular", "cellular", "u...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unkno...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

Making dummies for varible default

table(all\_data$default)

##

## no yes

## 44396 815

all\_data$default=as.numeric(all\_data$default=="yes")

Making dummies for variable housing

table(all\_data$housing)

##

## no yes

## 20081 25130

all\_data$housing=as.numeric(all\_data$housing=="yes")

glimpse(all\_data)

## Observations: 45,211

## Variables: 27

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "...

## $ contact <chr> "cellular", "cellular", "cellular", "cellular", "u...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unkno...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

Making dummies for variable loan

table(all\_data$loan)

##

## no yes

## 37967 7244

all\_data$loan=as.numeric(all\_data$loan=="yes")

glimpse(all\_data)

## Observations: 45,211

## Variables: 27

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ contact <chr> "cellular", "cellular", "cellular", "cellular", "u...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unkno...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

Making dummies for variable contact

t=table(all\_data$contact)

sort(t)

##

## telephone unknown cellular

## 2906 13020 29285

all\_data=all\_data %>%

mutate(co\_cellular=as.numeric(contact %**in**% c("cellular")),

co\_tel=as.numeric(contact %**in**% c("telephone"))

) %>%

select(-contact)

glimpse(all\_data)

## Observations: 45,211

## Variables: 28

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ month <chr> "aug", "jul", "aug", "mar", "may", "jun", "jun", "...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unkno...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

## $ co\_cellular <dbl> 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

Making dummies for variable month

table(all\_data$month)

##

## apr aug dec feb jan jul jun mar may nov oct sep

## 2932 6247 214 2649 1403 6895 5341 477 13766 3970 738 579

*#lets convert into percentage across months.*

finalmnth=round(prop.table(table(all\_data$month,all\_data$y),1)\*100,1)

sss=addmargins(finalmnth,2) *#adding margin across Y*

sort(sss[,1])

## mar oct sep dec apr feb aug jan nov jun jul may

## 46.7 55.0 57.8 58.0 80.1 83.1 88.7 89.1 89.3 90.1 91.0 93.2

*#may taken as base var*

all\_data=all\_data %>%

mutate(month\_1=as.numeric(month %**in**% c("aug","jun","nov","jan","jul")),

month\_2=as.numeric(month %**in**% c("dec","sep")),

month\_3=as.numeric(month=="mar"),

month\_4=as.numeric(month=="oct"),

month\_5=as.numeric(month=="apr"),

month\_6=as.numeric(month=="feb")) %>%

select(-month)

glimpse(all\_data)

## Observations: 45,211

## Variables: 33

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ poutcome <chr> "unknown", "unknown", "unknown", "unknown", "unkno...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

## $ co\_cellular <dbl> 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ month\_1 <dbl> 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0,...

## $ month\_2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_3 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...

Making dummies for variable outcome

t=table(all\_data$poutcome)

sort(t)

##

## success other failure unknown

## 1511 1840 4901 36959

*#unknown as base var*

all\_data=all\_data %>%

mutate(poc\_success=as.numeric(poutcome %**in**% c("success")),

poc\_failure=as.numeric(poutcome %**in**% c("failure")),

poc\_other=as.numeric(poutcome %**in**% c("other"))

)%>%

select(-poutcome)

glimpse(all\_data)

## Observations: 45,211

## Variables: 35

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

## $ co\_cellular <dbl> 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ month\_1 <dbl> 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0,...

## $ month\_2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_3 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...

## $ poc\_success <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_failure <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_other <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...

Thus data preparation is done and we will now seperate both test n train data.

glimpse(all\_data)

## Observations: 45,211

## Variables: 35

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", ...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

## $ co\_cellular <dbl> 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ month\_1 <dbl> 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0,...

## $ month\_2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_3 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...

## $ poc\_success <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_failure <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_other <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...

table(all\_data$y)

##

## no yes

## 27927 3720

table(train$y)

##

## no yes

## 27927 3720

all\_data$y=as.numeric(all\_data$y=="yes")

table(all\_data$y)

##

## 0 1

## 27927 3720

glimpse(all\_data)

## Observations: 45,211

## Variables: 35

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...

## $ data <chr> "train", "train", "train", "train", "train", "trai...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

## $ co\_cellular <dbl> 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ month\_1 <dbl> 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0,...

## $ month\_2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_3 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...

## $ poc\_success <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_failure <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_other <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...

Separating test and train:

train=all\_data %>%

filter(data=='train') %>%

select(-data) *#31647,34*

test=all\_data %>%

filter(data=='test') %>%

select(-data,-y)

Lets view the structure of test n train datasets:

glimpse(train) *#31647,34*

## Observations: 31,647

## Variables: 34

## $ age <int> 45, 34, 40, 58, 59, 36, 34, 38, 52, 48, 28, 43, 37...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 0, 311, 5810, 169, 24, -868, -140, 98, 1203, 38...

## $ housing <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,...

## $ loan <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ day <int> 26, 10, 6, 12, 16, 11, 30, 25, 28, 6, 3, 19, 21, 2...

## $ duration <int> 105, 268, 738, 139, 181, 100, 198, 456, 103, 61, 1...

## $ campaign <int> 10, 1, 2, 1, 3, 5, 2, 3, 1, 5, 5, 1, 4, 1, 2, 4, 1...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, 245, -1, -1, 198, ...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0,...

## $ ID <int> 22944, 13870, 19301, 31334, 3849, 10192, 12411, 12...

## $ y <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...

## $ job\_1 <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ job\_2 <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...

## $ edu\_sec <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0,...

## $ co\_cellular <dbl> 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,...

## $ month\_1 <dbl> 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0,...

## $ month\_2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_3 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...

## $ poc\_success <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_failure <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_other <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...

glimpse(test) *#13564,33*

## Observations: 13,564

## Variables: 33

## $ age <int> 33, 47, 35, 28, 58, 32, 46, 36, 37, 58, 55, 54, 38...

## $ default <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ balance <int> 2, 1506, 231, 447, 71, 23, -246, 265, 0, -364, 0, ...

## $ housing <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1,...

## $ loan <dbl> 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0,...

## $ day <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,...

## $ duration <int> 76, 92, 139, 217, 71, 160, 255, 348, 137, 355, 160...

## $ campaign <int> 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...

## $ pdays <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1...

## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ ID <int> 3, 4, 6, 7, 14, 23, 29, 30, 40, 47, 49, 51, 56, 59...

## $ job\_1 <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0,...

## $ job\_2 <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,...

## $ job\_3 <dbl> 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ job\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ divorced <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...

## $ single <dbl> 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0,...

## $ edu\_primary <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,...

## $ edu\_sec <dbl> 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1,...

## $ edu\_tert <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0,...

## $ co\_cellular <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ co\_tel <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_3 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_4 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ month\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_success <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_failure <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

## $ poc\_other <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

now lets divide the train dataset in the ratio 75:25.

set.seed(5)

s=sample(1:nrow(train),0.75\*nrow(train))

train\_75=train[s,] *#23735,34*

test\_25=train[-s,]*#7912,34*

Step 3: Model Building

We will use train for logistic regression model building and use train\_25 to test the performance of the model thus built.

Lets build logistic regression model on train dataset.

**library**(car)

##

## Attaching package: 'car'

## The following object is masked from 'package:dplyr':

##

## recode

for\_vif=lm(y~.,data=train)

summary(for\_vif)

##

## Call:

## lm(formula = y ~ ., data = train)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.06429 -0.11770 -0.03423 0.03642 1.04121

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -1.505e-01 1.283e-02 -11.729 < 2e-16 \*\*\*

## age 3.015e-04 1.837e-04 1.641 0.100800

## default 4.211e-03 1.143e-02 0.368 0.712674

## balance 1.054e-06 5.003e-07 2.106 0.035204 \*

## housing -3.438e-02 3.555e-03 -9.672 < 2e-16 \*\*\*

## loan -1.291e-02 4.153e-03 -3.110 0.001874 \*\*

## day 4.884e-04 1.926e-04 2.536 0.011233 \*

## duration 4.797e-04 5.887e-06 81.488 < 2e-16 \*\*\*

## campaign -2.866e-04 5.015e-04 -0.571 0.567727

## pdays -7.456e-05 3.220e-05 -2.316 0.020582 \*

## previous 4.162e-04 7.179e-04 0.580 0.562084

## ID 7.404e-06 2.186e-07 33.867 < 2e-16 \*\*\*

## job\_1 3.799e-03 4.439e-03 0.856 0.392043

## job\_2 -4.361e-03 4.689e-03 -0.930 0.352264

## job\_3 1.029e-02 5.398e-03 1.907 0.056592 .

## job\_4 5.558e-02 1.129e-02 4.924 8.52e-07 \*\*\*

## job\_5 2.802e-02 8.136e-03 3.444 0.000574 \*\*\*

## job\_6 -4.295e-03 9.344e-03 -0.460 0.645792

## divorced 1.394e-02 4.858e-03 2.870 0.004103 \*\*

## single 1.722e-02 3.834e-03 4.492 7.08e-06 \*\*\*

## edu\_primary -6.400e-03 8.419e-03 -0.760 0.447196

## edu\_sec 4.772e-03 7.748e-03 0.616 0.538009

## edu\_tert 1.072e-02 8.298e-03 1.292 0.196520

## co\_cellular -9.688e-02 5.808e-03 -16.682 < 2e-16 \*\*\*

## co\_tel -1.092e-01 8.066e-03 -13.535 < 2e-16 \*\*\*

## month\_1 2.368e-02 4.220e-03 5.610 2.04e-08 \*\*\*

## month\_2 1.007e-01 1.240e-02 8.115 5.03e-16 \*\*\*

## month\_3 3.088e-01 1.525e-02 20.248 < 2e-16 \*\*\*

## month\_4 1.630e-01 1.257e-02 12.967 < 2e-16 \*\*\*

## month\_5 3.982e-02 6.821e-03 5.837 5.35e-09 \*\*\*

## month\_6 3.573e-02 7.516e-03 4.754 2.00e-06 \*\*\*

## poc\_success 3.651e-01 1.063e-02 34.353 < 2e-16 \*\*\*

## poc\_failure -2.516e-02 9.567e-03 -2.630 0.008552 \*\*

## poc\_other -7.933e-03 1.110e-02 -0.715 0.474773

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2656 on 31613 degrees of freedom

## Multiple R-squared: 0.3205, Adjusted R-squared: 0.3198

## F-statistic: 451.9 on 33 and 31613 DF, p-value: < 2.2e-16

In order to take care of multi collinearity,we remove variables whose VIF>5,as follows:

t=vif(for\_vif)

sort(t,decreasing = T)[1:5]

## edu\_sec edu\_tert pdays edu\_primary poc\_failure

## 6.725263 6.376653 4.658624 4.095902 3.929616

Removing variable edu\_sec

for\_vif=lm(y~.-edu\_sec,data=train)

t=vif(for\_vif)

sort(t,decreasing = T)[1:5]

## pdays poc\_failure ID co\_cellular poc\_other

## 4.658565 3.929525 3.659703 3.453187 2.186067

summary(for\_vif)

##

## Call:

## lm(formula = y ~ . - edu\_sec, data = train)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.06369 -0.11762 -0.03423 0.03654 1.04150

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -1.458e-01 1.027e-02 -14.198 < 2e-16 \*\*\*

## age 2.911e-04 1.829e-04 1.591 0.111552

## default 4.193e-03 1.143e-02 0.367 0.713794

## balance 1.054e-06 5.003e-07 2.107 0.035125 \*

## housing -3.431e-02 3.553e-03 -9.656 < 2e-16 \*\*\*

## loan -1.279e-02 4.148e-03 -3.083 0.002048 \*\*

## day 4.882e-04 1.926e-04 2.534 0.011271 \*

## duration 4.797e-04 5.887e-06 81.488 < 2e-16 \*\*\*

## campaign -2.863e-04 5.015e-04 -0.571 0.568116

## pdays -7.464e-05 3.220e-05 -2.318 0.020460 \*

## previous 4.169e-04 7.179e-04 0.581 0.561378

## ID 7.402e-06 2.186e-07 33.863 < 2e-16 \*\*\*

## job\_1 3.789e-03 4.439e-03 0.854 0.393337

## job\_2 -4.341e-03 4.688e-03 -0.926 0.354497

## job\_3 1.016e-02 5.393e-03 1.884 0.059612 .

## job\_4 5.490e-02 1.123e-02 4.887 1.03e-06 \*\*\*

## job\_5 2.817e-02 8.132e-03 3.464 0.000532 \*\*\*

## job\_6 -4.206e-03 9.343e-03 -0.450 0.652607

## divorced 1.404e-02 4.855e-03 2.891 0.003844 \*\*

## single 1.719e-02 3.834e-03 4.484 7.35e-06 \*\*\*

## edu\_primary -1.078e-02 4.507e-03 -2.391 0.016789 \*

## edu\_tert 6.368e-03 4.357e-03 1.462 0.143829

## co\_cellular -9.676e-02 5.804e-03 -16.671 < 2e-16 \*\*\*

## co\_tel -1.091e-01 8.066e-03 -13.528 < 2e-16 \*\*\*

## month\_1 2.367e-02 4.220e-03 5.609 2.05e-08 \*\*\*

## month\_2 1.006e-01 1.240e-02 8.109 5.31e-16 \*\*\*

## month\_3 3.088e-01 1.525e-02 20.246 < 2e-16 \*\*\*

## month\_4 1.630e-01 1.257e-02 12.964 < 2e-16 \*\*\*

## month\_5 3.982e-02 6.821e-03 5.837 5.36e-09 \*\*\*

## month\_6 3.572e-02 7.516e-03 4.752 2.02e-06 \*\*\*

## poc\_success 3.651e-01 1.063e-02 34.356 < 2e-16 \*\*\*

## poc\_failure -2.513e-02 9.567e-03 -2.627 0.008626 \*\*

## poc\_other -7.876e-03 1.110e-02 -0.710 0.477961

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2656 on 31614 degrees of freedom

## Multiple R-squared: 0.3205, Adjusted R-squared: 0.3198

## F-statistic: 466 on 32 and 31614 DF, p-value: < 2.2e-16

Now lets remove edu-sec from train dataset

colnames(train) *#34var*

## [1] "age" "default" "balance" "housing" "loan"

## [6] "day" "duration" "campaign" "pdays" "previous"

## [11] "ID" "y" "job\_1" "job\_2" "job\_3"

## [16] "job\_4" "job\_5" "job\_6" "divorced" "single"

## [21] "edu\_primary" "edu\_sec" "edu\_tert" "co\_cellular" "co\_tel"

## [26] "month\_1" "month\_2" "month\_3" "month\_4" "month\_5"

## [31] "month\_6" "poc\_success" "poc\_failure" "poc\_other"

fit\_train=train %>%

select(-edu\_sec)

*#1 omited*

colnames(fit\_train) *#33var including target(y)*

## [1] "age" "default" "balance" "housing" "loan"

## [6] "day" "duration" "campaign" "pdays" "previous"

## [11] "ID" "y" "job\_1" "job\_2" "job\_3"

## [16] "job\_4" "job\_5" "job\_6" "divorced" "single"

## [21] "edu\_primary" "edu\_tert" "co\_cellular" "co\_tel" "month\_1"

## [26] "month\_2" "month\_3" "month\_4" "month\_5" "month\_6"

## [31] "poc\_success" "poc\_failure" "poc\_other"

Lets build model on fit\_train dataset:

fit=glm(y~.,family = "binomial",data=fit\_train) *#32 predictor var*

summary(fit) *#we get aic:14348*

##

## Call:

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

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -5.7162 -0.3503 -0.2137 -0.1199 3.2305

##

## Coefficients:

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

## (Intercept) -6.010e+00 1.709e-01 -35.168 < 2e-16 \*\*\*

## age 4.180e-05 2.644e-03 0.016 0.987386

## default 8.625e-02 2.144e-01 0.402 0.687500

## balance 1.259e-05 6.196e-06 2.032 0.042187 \*

## housing -4.787e-01 5.322e-02 -8.996 < 2e-16 \*\*\*

## loan -2.255e-01 7.184e-02 -3.139 0.001694 \*\*

## day 5.060e-03 2.797e-03 1.809 0.070439 .

## duration 4.532e-03 8.160e-05 55.535 < 2e-16 \*\*\*

## campaign -5.043e-02 1.200e-02 -4.202 2.64e-05 \*\*\*

## pdays -2.180e-04 3.477e-04 -0.627 0.530631

## previous 3.051e-03 7.611e-03 0.401 0.688524

## ID 1.008e-04 3.054e-06 32.991 < 2e-16 \*\*\*

## job\_1 4.729e-02 6.733e-02 0.702 0.482426

## job\_2 -1.003e-01 7.675e-02 -1.307 0.191191

## job\_3 1.267e-01 7.742e-02 1.637 0.101638

## job\_4 2.036e-01 1.211e-01 1.681 0.092807 .

## job\_5 1.966e-01 1.100e-01 1.787 0.073887 .

## job\_6 -1.221e-01 1.309e-01 -0.933 0.350734

## divorced 2.563e-01 7.198e-02 3.561 0.000369 \*\*\*

## single 2.173e-01 5.616e-02 3.869 0.000109 \*\*\*

## edu\_primary -2.708e-01 7.517e-02 -3.603 0.000315 \*\*\*

## edu\_tert 7.046e-02 6.031e-02 1.168 0.242691

## co\_cellular -8.455e-01 8.821e-02 -9.586 < 2e-16 \*\*\*

## co\_tel -1.047e+00 1.227e-01 -8.530 < 2e-16 \*\*\*

## month\_1 4.834e-01 6.624e-02 7.298 2.92e-13 \*\*\*

## month\_2 5.866e-01 1.197e-01 4.902 9.49e-07 \*\*\*

## month\_3 2.151e+00 1.420e-01 15.148 < 2e-16 \*\*\*

## month\_4 1.014e+00 1.217e-01 8.334 < 2e-16 \*\*\*

## month\_5 6.524e-01 8.433e-02 7.736 1.02e-14 \*\*\*

## month\_6 6.471e-01 9.788e-02 6.611 3.81e-11 \*\*\*

## poc\_success 1.561e+00 1.019e-01 15.323 < 2e-16 \*\*\*

## poc\_failure -3.769e-01 1.087e-01 -3.468 0.000524 \*\*\*

## poc\_other -2.204e-01 1.256e-01 -1.755 0.079330 .

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 22913 on 31646 degrees of freedom

## Residual deviance: 14282 on 31614 degrees of freedom

## AIC: 14348

##

## Number of Fisher Scoring iterations: 6

Now lets remove all variables whose p value is >0.05 using step function.

fit=step(fit)

## Start: AIC=14348.09

## y ~ age + default + balance + housing + loan + day + duration +

## campaign + pdays + previous + ID + job\_1 + job\_2 + job\_3 +

## job\_4 + job\_5 + job\_6 + divorced + single + edu\_primary +

## edu\_tert + co\_cellular + co\_tel + month\_1 + month\_2 + month\_3 +

## month\_4 + month\_5 + month\_6 + poc\_success + poc\_failure +

## poc\_other

##

## Df Deviance AIC

## - age 1 14282 14346

## - previous 1 14282 14346

## - default 1 14282 14346

## - pdays 1 14282 14346

## - job\_1 1 14283 14347

## - job\_6 1 14283 14347

## - edu\_tert 1 14284 14348

## - job\_2 1 14284 14348

## <none> 14282 14348

## - job\_3 1 14285 14349

## - job\_4 1 14285 14349

## - poc\_other 1 14285 14349

## - job\_5 1 14285 14349

## - day 1 14285 14349

## - balance 1 14286 14350

## - loan 1 14292 14356

## - poc\_failure 1 14294 14358

## - divorced 1 14294 14358

## - edu\_primary 1 14295 14359

## - single 1 14297 14361

## - campaign 1 14302 14366

## - month\_2 1 14306 14370

## - month\_6 1 14325 14389

## - month\_1 1 14336 14400

## - month\_5 1 14340 14404

## - month\_4 1 14350 14414

## - co\_tel 1 14357 14421

## - housing 1 14364 14428

## - co\_cellular 1 14370 14434

## - month\_3 1 14500 14564

## - poc\_success 1 14524 14588

## - ID 1 15382 15446

## - duration 1 18490 18554

##

## Step: AIC=14346.09

## y ~ default + balance + housing + loan + day + duration + campaign +

## pdays + previous + ID + job\_1 + job\_2 + job\_3 + job\_4 + job\_5 +

## job\_6 + divorced + single + edu\_primary + edu\_tert + co\_cellular +

## co\_tel + month\_1 + month\_2 + month\_3 + month\_4 + month\_5 +

## month\_6 + poc\_success + poc\_failure + poc\_other

##

## Df Deviance AIC

## - previous 1 14282 14344

## - default 1 14282 14344

## - pdays 1 14282 14344

## - job\_1 1 14283 14345

## - job\_6 1 14283 14345

## - edu\_tert 1 14284 14346

## - job\_2 1 14284 14346

## <none> 14282 14346

## - job\_3 1 14285 14347

## - job\_4 1 14285 14347

## - poc\_other 1 14285 14347

## - day 1 14285 14347

## - balance 1 14286 14348

## - job\_5 1 14286 14348

## - loan 1 14292 14354

## - poc\_failure 1 14294 14356

## - divorced 1 14295 14357

## - edu\_primary 1 14296 14358

## - single 1 14300 14362

## - campaign 1 14302 14364

## - month\_2 1 14306 14368

## - month\_6 1 14325 14387

## - month\_1 1 14336 14398

## - month\_5 1 14340 14402

## - month\_4 1 14350 14412

## - co\_tel 1 14358 14420

## - housing 1 14365 14427

## - co\_cellular 1 14370 14432

## - month\_3 1 14501 14563

## - poc\_success 1 14524 14586

## - ID 1 15383 15445

## - duration 1 18491 18553

##

## Step: AIC=14344.23

## y ~ default + balance + housing + loan + day + duration + campaign +

## pdays + ID + job\_1 + job\_2 + job\_3 + job\_4 + job\_5 + job\_6 +

## divorced + single + edu\_primary + edu\_tert + co\_cellular +

## co\_tel + month\_1 + month\_2 + month\_3 + month\_4 + month\_5 +

## month\_6 + poc\_success + poc\_failure + poc\_other

##

## Df Deviance AIC

## - default 1 14282 14342

## - pdays 1 14283 14343

## - job\_1 1 14283 14343

## - job\_6 1 14283 14343

## - edu\_tert 1 14284 14344

## - job\_2 1 14284 14344

## <none> 14282 14344

## - job\_3 1 14285 14345

## - job\_4 1 14285 14345

## - poc\_other 1 14285 14345

## - day 1 14286 14346

## - balance 1 14286 14346

## - job\_5 1 14286 14346

## - loan 1 14292 14352

## - poc\_failure 1 14294 14354

## - divorced 1 14295 14355

## - edu\_primary 1 14296 14356

## - single 1 14300 14360

## - campaign 1 14302 14362

## - month\_2 1 14306 14366

## - month\_6 1 14325 14385

## - month\_1 1 14337 14397

## - month\_5 1 14341 14401

## - month\_4 1 14350 14410

## - co\_tel 1 14358 14418

## - housing 1 14365 14425

## - co\_cellular 1 14370 14430

## - month\_3 1 14501 14561

## - poc\_success 1 14542 14602

## - ID 1 15384 15444

## - duration 1 18491 18551

##

## Step: AIC=14342.39

## y ~ balance + housing + loan + day + duration + campaign + pdays +

## ID + job\_1 + job\_2 + job\_3 + job\_4 + job\_5 + job\_6 + divorced +

## single + edu\_primary + edu\_tert + co\_cellular + co\_tel +

## month\_1 + month\_2 + month\_3 + month\_4 + month\_5 + month\_6 +

## poc\_success + poc\_failure + poc\_other

##

## Df Deviance AIC

## - pdays 1 14283 14341

## - job\_1 1 14283 14341

## - job\_6 1 14283 14341

## - edu\_tert 1 14284 14342

## - job\_2 1 14284 14342

## <none> 14282 14342

## - job\_3 1 14285 14343

## - job\_4 1 14285 14343

## - poc\_other 1 14285 14343

## - day 1 14286 14344

## - balance 1 14286 14344

## - job\_5 1 14287 14345

## - loan 1 14292 14350

## - poc\_failure 1 14295 14353

## - divorced 1 14295 14353

## - edu\_primary 1 14296 14354

## - single 1 14300 14358

## - campaign 1 14302 14360

## - month\_2 1 14306 14364

## - month\_6 1 14325 14383

## - month\_1 1 14337 14395

## - month\_5 1 14341 14399

## - month\_4 1 14350 14408

## - co\_tel 1 14358 14416

## - housing 1 14365 14423

## - co\_cellular 1 14370 14428

## - month\_3 1 14501 14559

## - poc\_success 1 14542 14600

## - ID 1 15385 15443

## - duration 1 18491 18549

##

## Step: AIC=14340.79

## y ~ balance + housing + loan + day + duration + campaign + ID +

## job\_1 + job\_2 + job\_3 + job\_4 + job\_5 + job\_6 + divorced +

## single + edu\_primary + edu\_tert + co\_cellular + co\_tel +

## month\_1 + month\_2 + month\_3 + month\_4 + month\_5 + month\_6 +

## poc\_success + poc\_failure + poc\_other

##

## Df Deviance AIC

## - job\_1 1 14283 14339

## - job\_6 1 14284 14340

## - edu\_tert 1 14284 14340

## - job\_2 1 14284 14340

## <none> 14283 14341

## - job\_3 1 14286 14342

## - job\_4 1 14286 14342

## - day 1 14286 14342

## - balance 1 14287 14343

## - job\_5 1 14287 14343

## - poc\_other 1 14290 14346

## - loan 1 14293 14349

## - divorced 1 14295 14351

## - edu\_primary 1 14296 14352

## - single 1 14301 14357

## - campaign 1 14302 14358

## - month\_2 1 14306 14362

## - poc\_failure 1 14321 14377

## - month\_6 1 14326 14382

## - month\_1 1 14338 14394

## - month\_5 1 14341 14397

## - month\_4 1 14352 14408

## - co\_tel 1 14358 14414

## - housing 1 14367 14423

## - co\_cellular 1 14370 14426

## - month\_3 1 14502 14558

## - poc\_success 1 14645 14701

## - ID 1 15386 15442

## - duration 1 18492 18548

##

## Step: AIC=14339.32

## y ~ balance + housing + loan + day + duration + campaign + ID +

## job\_2 + job\_3 + job\_4 + job\_5 + job\_6 + divorced + single +

## edu\_primary + edu\_tert + co\_cellular + co\_tel + month\_1 +

## month\_2 + month\_3 + month\_4 + month\_5 + month\_6 + poc\_success +

## poc\_failure + poc\_other

##

## Df Deviance AIC

## - job\_6 1 14285 14339

## - edu\_tert 1 14285 14339

## <none> 14283 14339

## - job\_3 1 14286 14340

## - job\_4 1 14286 14340

## - job\_2 1 14286 14340

## - day 1 14287 14341

## - job\_5 1 14287 14341

## - balance 1 14287 14341

## - poc\_other 1 14291 14345

## - loan 1 14294 14348

## - divorced 1 14296 14350

## - edu\_primary 1 14298 14352

## - single 1 14301 14355

## - campaign 1 14303 14357

## - month\_2 1 14307 14361

## - poc\_failure 1 14321 14375

## - month\_6 1 14327 14381

## - month\_1 1 14340 14394

## - month\_5 1 14342 14396

## - month\_4 1 14352 14406

## - co\_tel 1 14358 14412

## - housing 1 14369 14423

## - co\_cellular 1 14370 14424

## - month\_3 1 14503 14557

## - poc\_success 1 14646 14700

## - ID 1 15386 15440

## - duration 1 18494 18548

##

## Step: AIC=14338.63

## y ~ balance + housing + loan + day + duration + campaign + ID +

## job\_2 + job\_3 + job\_4 + job\_5 + divorced + single + edu\_primary +

## edu\_tert + co\_cellular + co\_tel + month\_1 + month\_2 + month\_3 +

## month\_4 + month\_5 + month\_6 + poc\_success + poc\_failure +

## poc\_other

##

## Df Deviance AIC

## - edu\_tert 1 14286 14338

## <none> 14285 14339

## - job\_2 1 14287 14339

## - job\_3 1 14287 14339

## - job\_4 1 14288 14340

## - day 1 14288 14340

## - balance 1 14289 14341

## - job\_5 1 14289 14341

## - poc\_other 1 14292 14344

## - loan 1 14295 14347

## - divorced 1 14297 14349

## - edu\_primary 1 14300 14352

## - single 1 14302 14354

## - campaign 1 14304 14356

## - month\_2 1 14308 14360

## - poc\_failure 1 14322 14374

## - month\_6 1 14327 14379

## - month\_1 1 14340 14392

## - month\_5 1 14343 14395

## - month\_4 1 14354 14406

## - co\_tel 1 14360 14412

## - housing 1 14369 14421

## - co\_cellular 1 14371 14423

## - month\_3 1 14504 14556

## - poc\_success 1 14647 14699

## - ID 1 15386 15438

## - duration 1 18494 18546

##

## Step: AIC=14338.38

## y ~ balance + housing + loan + day + duration + campaign + ID +

## job\_2 + job\_3 + job\_4 + job\_5 + divorced + single + edu\_primary +

## co\_cellular + co\_tel + month\_1 + month\_2 + month\_3 + month\_4 +

## month\_5 + month\_6 + poc\_success + poc\_failure + poc\_other

##

## Df Deviance AIC

## <none> 14286 14338

## - job\_2 1 14289 14339

## - job\_4 1 14289 14339

## - day 1 14290 14340

## - balance 1 14291 14341

## - job\_5 1 14291 14341

## - poc\_other 1 14294 14344

## - job\_3 1 14294 14344

## - loan 1 14296 14346

## - divorced 1 14299 14349

## - edu\_primary 1 14304 14354

## - campaign 1 14305 14355

## - single 1 14306 14356

## - month\_2 1 14310 14360

## - poc\_failure 1 14324 14374

## - month\_6 1 14329 14379

## - month\_1 1 14343 14393

## - month\_5 1 14345 14395

## - month\_4 1 14356 14406

## - co\_tel 1 14362 14412

## - housing 1 14372 14422

## - co\_cellular 1 14373 14423

## - month\_3 1 14507 14557

## - poc\_success 1 14650 14700

## - ID 1 15390 15440

## - duration 1 18495 18545

lets check the remaining significant variables

names(fit$coefficients) *#25 significant var*

## [1] "(Intercept)" "balance" "housing" "loan" "day"

## [6] "duration" "campaign" "ID" "job\_2" "job\_3"

## [11] "job\_4" "job\_5" "divorced" "single" "edu\_primary"

## [16] "co\_cellular" "co\_tel" "month\_1" "month\_2" "month\_3"

## [21] "month\_4" "month\_5" "month\_6" "poc\_success" "poc\_failure"

## [26] "poc\_other"

lets build final logistic model on significant variables on dataset fit\_train

fit\_final=glm(y~balance + housing + loan + duration + campaign + ID +

job\_3 + job\_5 + divorced + single + edu\_primary +

co\_cellular + co\_tel + month\_1 + month\_2 + month\_3 + month\_4 +

month\_5 + month\_6 + poc\_success + poc\_failure + poc\_other ,data=fit\_train,family="binomial")

summary(fit\_final)

##

## Call:

## glm(formula = y ~ balance + housing + loan + duration + campaign +

## ID + job\_3 + job\_5 + divorced + single + edu\_primary + co\_cellular +

## co\_tel + month\_1 + month\_2 + month\_3 + month\_4 + month\_5 +

## month\_6 + poc\_success + poc\_failure + poc\_other, family = "binomial",

## data = fit\_train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -5.7230 -0.3521 -0.2141 -0.1192 3.2428

##

## Coefficients:

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

## (Intercept) -5.931e+00 1.206e-01 -49.168 < 2e-16 \*\*\*

## balance 1.298e-05 6.148e-06 2.112 0.034726 \*

## housing -4.967e-01 5.208e-02 -9.536 < 2e-16 \*\*\*

## loan -2.323e-01 7.149e-02 -3.249 0.001158 \*\*

## duration 4.522e-03 8.147e-05 55.508 < 2e-16 \*\*\*

## campaign -4.798e-02 1.194e-02 -4.017 5.89e-05 \*\*\*

## ID 1.005e-04 3.012e-06 33.366 < 2e-16 \*\*\*

## job\_3 1.661e-01 5.339e-02 3.112 0.001860 \*\*

## job\_5 2.068e-01 8.908e-02 2.322 0.020243 \*

## divorced 2.567e-01 7.169e-02 3.581 0.000343 \*\*\*

## single 2.502e-01 4.948e-02 5.057 4.26e-07 \*\*\*

## edu\_primary -3.026e-01 7.263e-02 -4.167 3.09e-05 \*\*\*

## co\_cellular -8.257e-01 8.748e-02 -9.439 < 2e-16 \*\*\*

## co\_tel -1.033e+00 1.215e-01 -8.503 < 2e-16 \*\*\*

## month\_1 4.945e-01 6.585e-02 7.510 5.90e-14 \*\*\*

## month\_2 5.873e-01 1.195e-01 4.915 8.89e-07 \*\*\*

## month\_3 2.161e+00 1.415e-01 15.270 < 2e-16 \*\*\*

## month\_4 1.044e+00 1.208e-01 8.638 < 2e-16 \*\*\*

## month\_5 6.734e-01 8.355e-02 8.061 7.59e-16 \*\*\*

## month\_6 6.057e-01 9.448e-02 6.410 1.45e-10 \*\*\*

## poc\_success 1.540e+00 8.219e-02 18.739 < 2e-16 \*\*\*

## poc\_failure -4.203e-01 6.926e-02 -6.069 1.29e-09 \*\*\*

## poc\_other -2.482e-01 9.551e-02 -2.599 0.009359 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 22913 on 31646 degrees of freedom

## Residual deviance: 14295 on 31624 degrees of freedom

## AIC: 14341

##

## Number of Fisher Scoring iterations: 6

names(fit\_final$coefficients)

## [1] "(Intercept)" "balance" "housing" "loan" "duration"

## [6] "campaign" "ID" "job\_3" "job\_5" "divorced"

## [11] "single" "edu\_primary" "co\_cellular" "co\_tel" "month\_1"

## [16] "month\_2" "month\_3" "month\_4" "month\_5" "month\_6"

## [21] "poc\_success" "poc\_failure" "poc\_other"

*it shows: aic:14341 and 22 significant var in final model*

Thus logistic regression model is successfully built.

Now lets make predict scores

train$score=predict(fit\_final,newdata = train,type="response")

*#score means Pi*

lets see how the score (Pi ) behaves.

**library**(ggplot2)

ggplot(train,aes(y=y,x=score,color=factor(y)))+

geom\_point()+geom\_jitter()

Step 4. Finding Cutoff value and Perfomance measurements of the model.

lets find cutoff based on these probability scores.

cutoff\_data=data.frame(cutoff=0,TP=0,FP=0,FN=0,TN=0)

cutoffs=seq(0,1,length=100)

**for** (i **in** cutoffs){

predicted=as.numeric(train$score>i)

TP=sum(predicted==1 & train$y==1)

FP=sum(predicted==1 & train$y==0)

FN=sum(predicted==0 & train$y==1)

TN=sum(predicted==0 & train$y==0)

cutoff\_data=rbind(cutoff\_data,c(i,TP,FP,FN,TN))

}

*## lets remove the dummy data cotaining top row in data frame cutoff\_data*

cutoff\_data=cutoff\_data[-1,]

*#we now have 100 obs in df cutoff\_data*

lets calculate the performance measures:sensitivity,specificity,accuracy, KS and precision.

cutoff\_data=cutoff\_data %>%

mutate(P=FN+TP,N=TN+FP, *#total positives and negatives*

Sn=TP/P, *#sensitivity*

Sp=TN/N, *#specificity*

KS=abs((TP/P)-(FP/N)),

Accuracy=(TP+TN)/(P+N),

Lift=(TP/P)/((TP+FP)/(P+N)),

Precision=TP/(TP+FP),

Recall=TP/P

) %>%

select(-P,-N)

lets view cutoff dataset:

*#View(cutoff\_data)*

Lets find cutoff value based on ks MAXIMUM.

KS\_cutoff=cutoff\_data$cutoff[which.max(cutoff\_data$KS)]

KS\_cutoff

## [1] 0.1111111

hence 0.1111111 is the cutoff value by ks max method.

Step 5.Predict the final output on test dataset.(whether the client subscribe or no to term deposit)

lets predict test scores

test$score=predict(fit\_final,newdata =test,type = "response")*#on final test dataset.*

Predicting whether the client has subscribed or no in final test dataset.

test$left=as.numeric(test$score>KS\_cutoff)*#if score is greater dan cutoff then true(1) else false(0)*

table(test$left)

##

## 0 1

## 10168 3396

Thus final prediction is as follows:

test$leftfinal=factor(test$left,levels = c(0,1),labels=c("no","yes"))

table(test$leftfinal)

##

## no yes

## 10168 3396

writing into csv file final output test$leftfinal

write.csv(test$leftfinal,"P5\_sub\_1.csv")

**Thus 3396 customers out of 13564 subscribe to term deposit according to the model.**

Step 6:Creating confusion matrix and find how good our model is (by predicting on test\_25 dataset)

test\_25$score=predict(fit\_final,newdata =test\_25,type = "response")

table(test\_25$y,as.numeric(test\_25$score>KS\_cutoff))

##

## 0 1

## 0 5888 1145

## 1 109 770

table(test\_25$y)

##

## 0 1

## 7033 879

here TP=770,TN=5888,FP=,FN=

Accuracy=(TP+TN)/(P+N):

a=(770+5888)/7912

a

## [1] 0.8415066

Hence error will be:

1-a

## [1] 0.1584934

\*\*Error is 15.85%.(according to ks method)

Lets plot the ROC curve:

**library**(pROC)

## Type 'citation("pROC")' for a citation.

##

## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':

##

## cov, smooth, var

roccurve=roc(test\_25$y,test\_25$score) *#real outcome and predicted score is plotted*

plot(roccurve)

Thus area under the ROC curve is:

auc(roccurve) *#0.9218*

## Area under the curve: 0.9218

Conclusion:

*Thus the target no. of customers to be focused upon for term deposits by the bank are predicted successfully using logistic regression model with an accuracy of 84.15% using KS method. The KS score examined came out to be: 0.72/1.00[our model and predictions wer very good]*