# Table of Contents


Bank Marketing Analysis ........ 1
Business Understanding ........ 2
Data Understanding ........ 2
Data Cleaning ........ 3
Exploratory Analysis ........ 3
Modelling ........ 12
Data Preparation ........ 13
Logistic Regression ........ 13
Model Evaluation ........ 25
Model Deployment and Recommendations ........ 30
Reducing Customer Acquision Cost ........ 32


# Bank Marketing Analysis

### Designing a Telemarketing Strategy To Reduce Acquisition Costs

A bank sells a product (called term deposit) to prospects mainly through telemarketing. If a prospect customer buys the product, we say that he has 'responded'.

The aim of this analysis is to **reduce the marketing cost by atleast 50%** and acquire a comparable number of customers (say 80-90%).

We'll use *telemarketing data* from a past campaign of the bank. The sales team had recorded customer data like age, salary, whether he has a loan, house, the month of call etc.

The idea is to use machine learning to predict the likelihood of a person 'buying the product 'responding'. We'll identify those who are most likely to respond and telemarket only to them, thereby reducing the total cost of acquisition per customer.

The standard process followed in analytics projects is:

1. Business Understanding

2. Data Understanding
3. Modelling

4. Model Evaluation

5. Model Deployment and Recommendations

## Business Understanding

The **overall goal** is to reduce telemarketing costs by about 50% and acquire atleast 80-90% of the customers.

The specific **objective of this analysis** is to build a **'response model'** to predict the likelihood of a prospect buying the product (or responding).

## Data Understanding

The datafile is named bank-marketing.csv. You can download it here: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

```
## 'data.frame':    45211 obs. of  19 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 ...
##  $ salary   : int  100000 60000 120000 20000 0 100000 100000 120000 55000
60000 ...
##  $ 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 ...
##  $ targeted : Factor w/ 2 levels "no","yes": 2 2 2 1 1 2 1 1 2 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
...
##  $ response : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

We have 45211 observations, i.e. we have data of 45211 potential customers. There are 20 variables (or 20 columns). We have two types of variables:

1. Two type of attributes (or variables):

- **Customer attributes** like age, job, salary, marital (status), education (primary / secondary / tertiary) etc.
- **Bank related attributes** like targeted (whether he/she was targeted before), loan (yes / no), contact (whether he/she has been contacted before etc.)

The last attribute 'response' tells us whether the person had responded to the bank's marketing campaign. It thus contains only two values - Yes or No. It is called the **target attribute** since that is what we want to predict using the other attributes.

## Data Cleaning

### Missing Values and Outliers

We always start with cleaning the data i.e. removing the missing values, any erroneous entries etc. Let's see if this data contains **missing values**.

```r
sum(is.na(bank_data))

## [1] 0
```

We have simply summed up all the missing values (denoted as 'NA'). There are none of them, so we move ahead.

Next, we should ideally look at **outliers** in the data. Outliers are extreme values for which treatment is done so that the data only represents the general trends and ignores extreme cases. For example, a person having an income of Rs 20 lacs per month will be an outlier in this dataset.

For now, we will not do outlier treatment since this data doesn't have many of them.

## Exploratory Analysis

In Exploratory Data Analysis, or EDA, we use plots and summaries of data to understand the patterns in it. Let's see some examples.

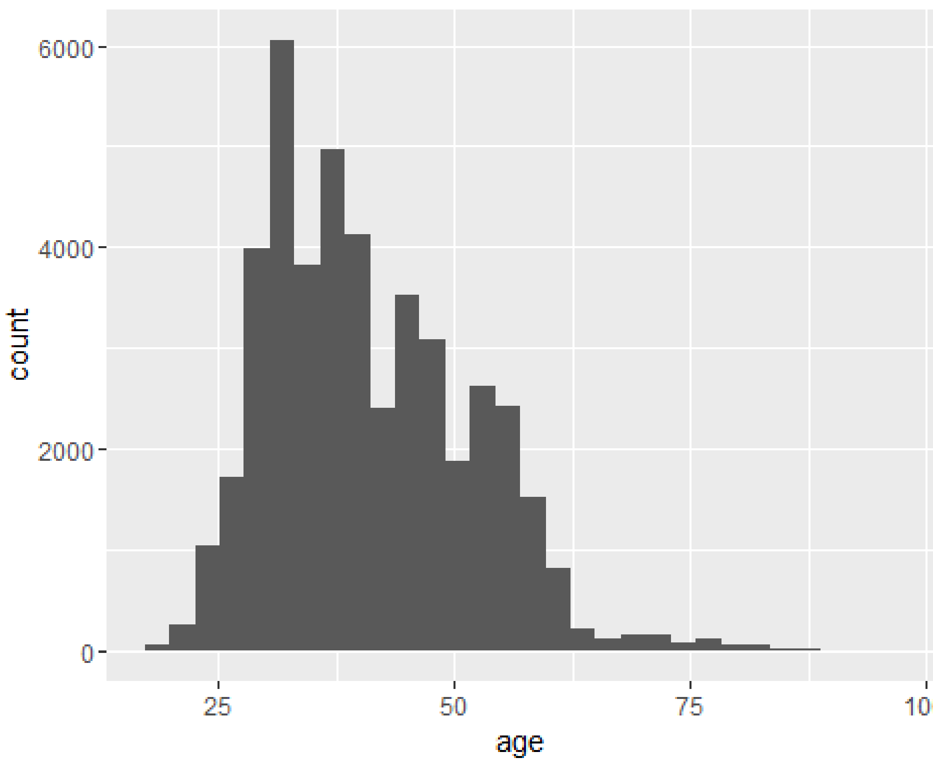
### Univariate Analysis

In univariate analysis, we analyse one variable at a time.

The following plot shows the distribution of peoples' ages in the data. The ggplot library is a great data visualisation tool in R.

Since age is a numeric variable, we plot a **histogram**.

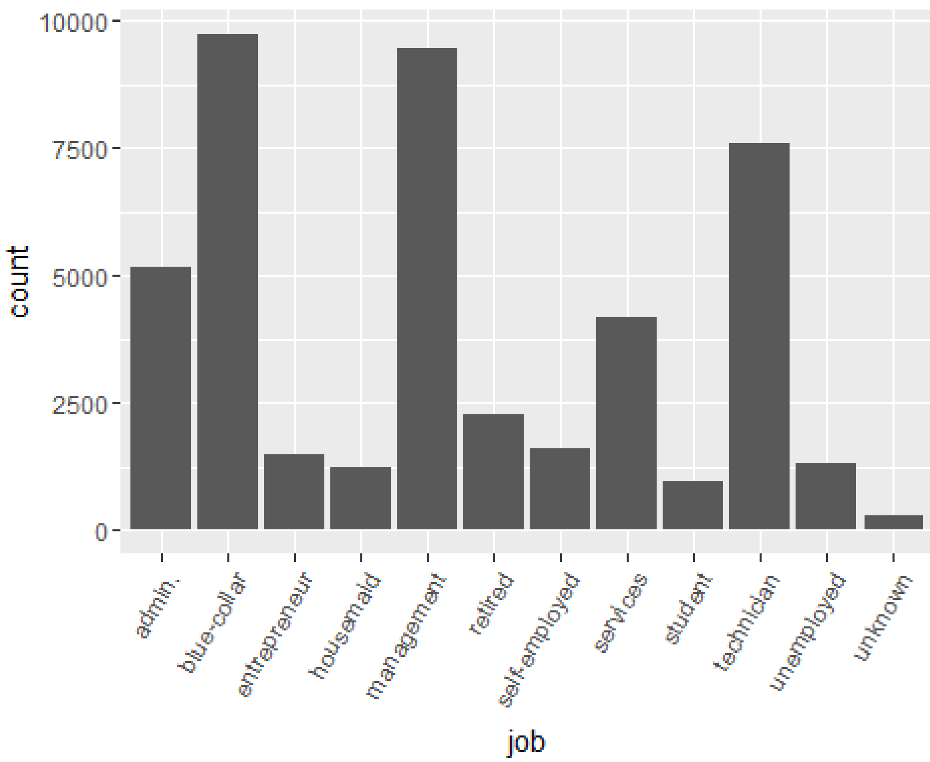
```r
library(ggplot2)
ggplot(bank_data, aes(age)) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




You can see that most people are between 25-50 years old. Very few people older than 60 years have been targeted.

Next, let's look at the types of jobs people have. Now *job* is not a numeric variable, it is a **categorical variable** and so we plot a **bar chart** for it.

```
ggplot(bank_data, aes(job)) + geom_bar() + theme(axis.text.x =
element_text(angle = 60, hjust = 1))
```




We have a large number of people with blue-collar jobs and in management, which are about 9000 each. The third highest category is technicians which are about 7500 people. Note that among 45,000 people, about 25,000 or 55% are either blue-collar workers, management employees technicians.

A very important variable for the bank would be **salary**. Let's have a look at the average and the median salary.

```
summary(bank_data$salary)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   20000   60000   57010   70000  120000
```

The average salary is about INR 57000 per month. Note that the maximum is only about INR 1.2 lacs per month.

Let's now look at the summary of the target variable **response**.

```
summary(bank_data$response)
```

So out of about 45,000 people, only 5000 had responded. The exact **response rate** ca calculated as:

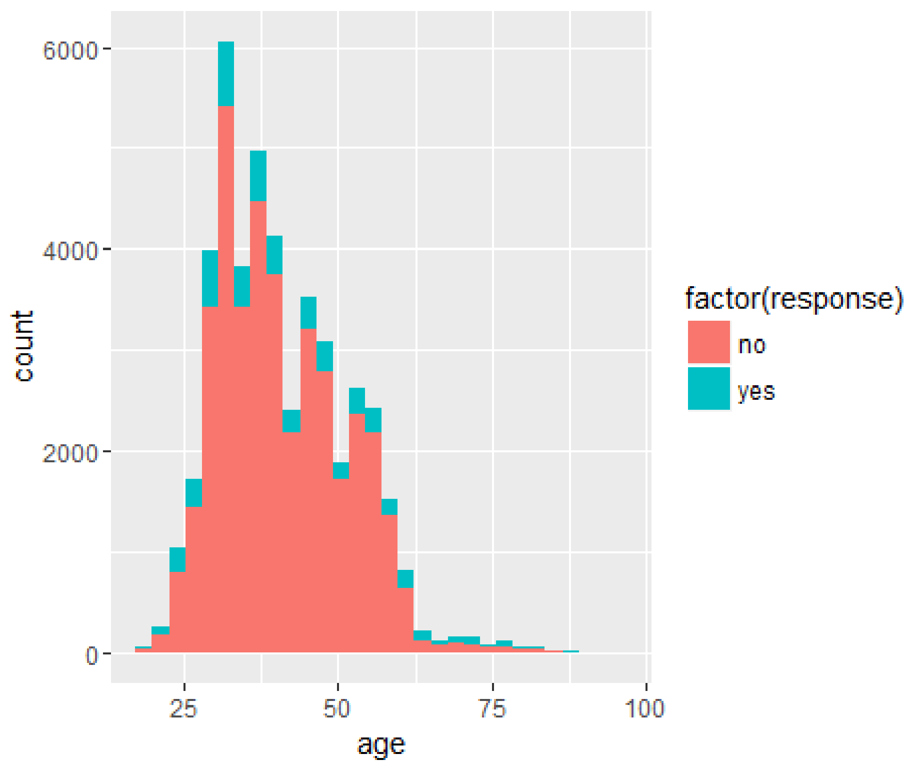
```
## [1] 0.1169848
```

An 11.7% response rate means that if the you make 100 calls to market the product (term deposit), about 11.7% will subscribe for a term-deposit. In marketing, 11.7% is a decently good rate.

### Multivariate Analysis

Now let's analyse two variables at a time, one of which should obviously be the target variable 'response'.

```
ggplot(bank_data, aes(age, fill = factor(response))) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




So now the plot shows information of two variables - the age on x-axis and response as a colour. This chart does not show any obvious trend of response rate with age.

The analysis will be easier if we could divide the age into **buckets**, e.g. 0-10 years, 10-20 years etc. This is called bucketing and is often done to divide **numeric variables** into smaller buckets.

```
bank_data$buckets.age <- cut(bank_data$age, breaks = c(10, 20, 30, 40, 50,
60, 70, 80, 90, 100))
str(bank_data)

## 'data.frame':    45211 obs. of  20 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 ...
##  $ salary      : int  100000 60000 120000 20000 0 100000 100000 120000
55000 60000 ...
##  $ 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 ...
##  $ targeted   : Factor w/ 2 levels "no","yes": 2 2 2 1 1 2 1 1 2 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 ...
##  $ response   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ buckets.age: Factor w/ 9 levels "(10,20]","(20,30]",..: 5 4 3 4 3 3 2 4
5 4 ...
```

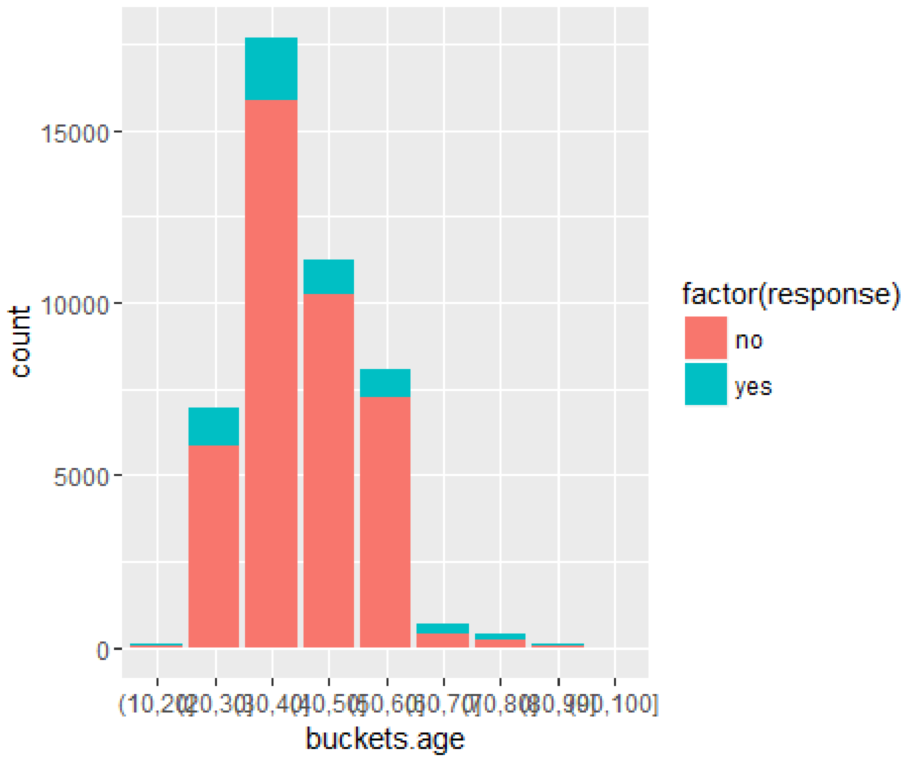
```r
sum(is.na(bank_data))
```

```
## [1] 0
```

Note that a new variable named *buckets.age* is now added to bank_data. Ages are now bucketed into this variable.

Let's use the buckets to see if age affects the response rate.

```r
ggplot(bank_data, aes(buckets.age, fill = factor(response))) + geom_bar()
```




It will be easier if we could see the response rate in numbers as well. We can convert the 'yes-no' values to '1-0' respectively and then calculate the response rate by summing up the 1s.

```
bank_data$response.numeric <- ifelse(bank_data$response == "yes", 1, 0)
str(bank_data)

## 'data.frame':    45211 obs. of  21 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 ...
##  $ salary            : int  100000 60000 120000 20000 0 100000 100000 120000
55000 60000 ...
##  $ 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 ...
##  $ targeted          : Factor w/ 2 levels "no","yes": 2 2 2 1 1 2 1 1 2 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 ...
##  $ response       : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1
...
##  $ buckets.age    : Factor w/ 9 levels "(10,20]","(20,30]",..: 5 4 3 4 3
3 2 4 5 4 ...
##  $ response.numeric: num  0 0 0 0 0 0 0 0 0 0 ...
```

We can now aggregate the 1s by each age bucket and see the response rates for each bucket.

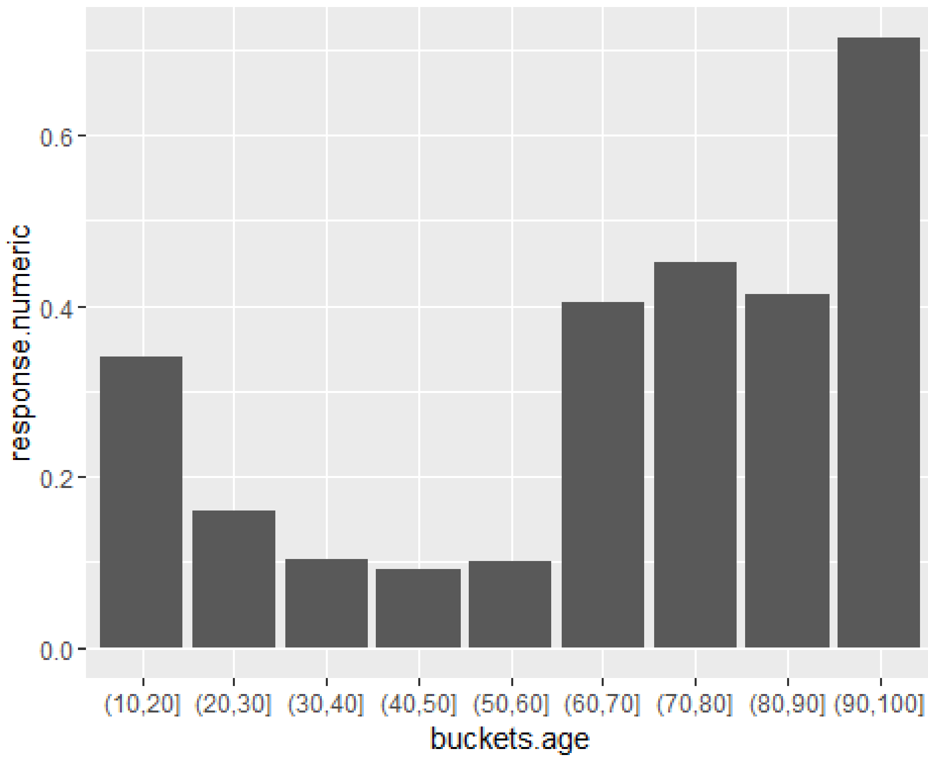
```
agg_age <- aggregate(response.numeric ~ buckets.age, data = bank_data, mean)
agg_age
```

```
##   buckets.age response.numeric
## 1     (10,20]       0.34020619
## 2     (20,30]       0.16039233
## 3     (30,40]       0.10244813
## 4     (40,50]       0.09066643
## 5     (50,60]       0.10053304
## 6     (60,70]       0.40513552
## 7     (70,80]       0.45103093
## 8     (80,90]       0.41304348
## 9    (90,100]       0.71428571
```

Note that the bucket 10-20 has about 34% response rate; 20-30 has 16% etc. The bucket 40-50 and 50-60 have low response rates (around 10%).

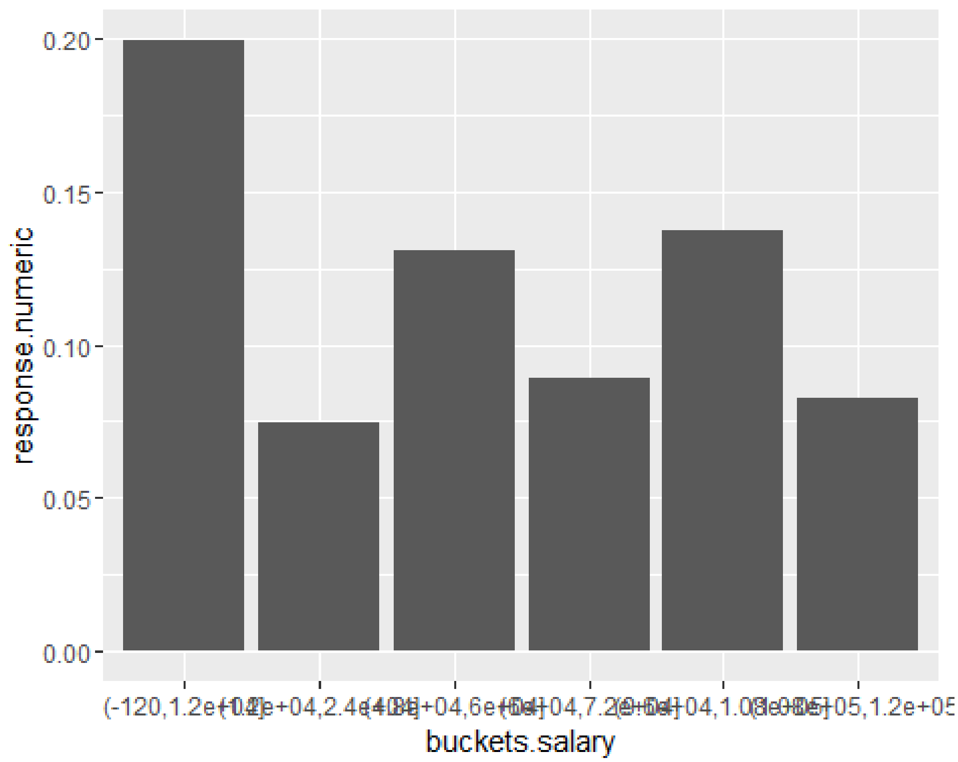
We can display the aggregate response rates in the plot as well.

```
ggplot(agg_age, aes(x = buckets.age, y = response.numeric)) + geom_bar(stat =
'identity')
```




Similarly, we can measure the response rate with salary and jobs.

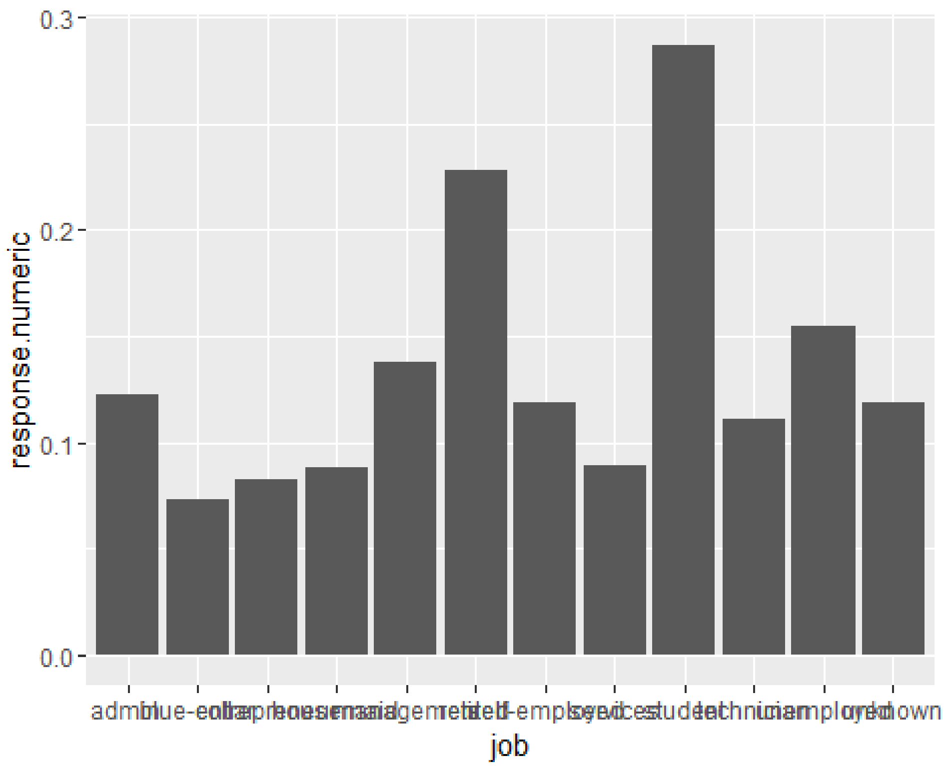
```
bank_data$buckets.salary <- cut(bank_data$salary, breaks = 10)
agg_salary <- aggregate(response.numeric ~ buckets.salary, data = bank_data,
mean)
agg_salary

##        buckets.salary response.numeric
## 1      (-120,1.2e+04]       0.19968367
## 2  (1.2e+04,2.4e+04]       0.07446227
## 3    (4.8e+04,6e+04]       0.13087713
## 4    (6e+04,7.2e+04]       0.08883004
## 5 (9.6e+04,1.08e+05]       0.13755551
## 6 (1.08e+05,1.2e+05]       0.08271688

ggplot(agg_salary, aes(x = buckets.salary, y = response.numeric)) +
geom_bar(stat = 'identity')
```




You can see that the response rate is highest for the lowest salary band. This might tell you something about the banking products you should be selling (which are used by people in this salary band.)

Let's also compare response rates across various jobs.

```
agg_job <- aggregate(response.numeric~job, data = bank_data, mean)
agg_job

##              job response.numeric
## 1         admin.       0.12202669
## 2    blue-collar       0.07274969
## 3   entrepreneur       0.08271688
## 4      housemaid       0.08790323
## 5     management       0.13755551
## 6        retired       0.22791519
## 7  self-employed       0.11842939
## 8       services       0.08883004
## 9        student       0.28678038
## 10    technician       0.11056996
## 11    unemployed       0.15502686
## 12       unknown       0.11805556

ggplot(agg_job, aes(job, response.numeric)) + geom_bar(stat = 'identity')
```




Interestingly, response rate is highest for students and second highest for retired people. It is quite low for blue-collar workers, housemaids and entrepreneurs.

Similarly, you can analyse response rates with other variables like education, marital status, loan etc.

But this way, we can only only analyse the effect of each variable separately. We saw that multiple attributes like age, salary etc. affect the reponse rate. How do we analyse the *combined effect* of the variables? Also, how can we know which variables affect response rate more than others?

**Machine Learning** helps us build **models** which extract the patterns in the data. We'll see that in the next section.

## Modelling

Let's now build some machine learning models to predict the type of potential customers who are more likely to respond.

To build machine learning models, we use only a part of the data to train the model. This is called **training data**.

Rest of the data is used to test or evaluate the model, which is called **test data**.

We'll use 70% data to train the model and the rest 30% to test it.

## Data Preparation

```
library(caret)

## Loading required package: lattice

library(caTools)
library(dummies)

## dummies-1.5.6 provided by Decision Patterns

bank_data <- bank_data[, -c(20, 21, 22)]


#creating dummy variables
bank_data$response <- as.integer(bank_data$response)
bank_data <- dummy.data.frame(bank_data)
bank_data$response <- as.factor(ifelse(bank_data$response == 1, "no", "yes"))


# splitting into train and test data
set.seed(1)
split_indices <- sample.split(bank_data$response, SplitRatio = 0.70)
train <- bank_data[split_indices, ]
test <- bank_data[!split_indices, ]
nrow(train)/nrow(bank_data)

## [1] 0.6999845

nrow(test)/nrow(bank_data)

## [1] 0.3000155
```

## Logistic Regression

Let's build the first model - **logistic regression**.

```
library(MASS)
library(car)
logistic_1 <- glm(response ~ ., family = "binomial", data = train)
summary(logistic_1)

##
## Call:
## glm(formula = response ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.7547  -0.3728  -0.2496  -0.1458   3.4916
##
## Coefficients: (11 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)         -4.457e+00  3.990e-01 -11.170  < 2e-16 ***
## age                 -2.450e-03  2.641e-03  -0.928  0.35354
## jobadmin.            3.545e-01  2.679e-01   1.323  0.18579
## `jobblue-collar`     2.078e-02  2.673e-01   0.078  0.93802
## jobentrepreneur     -6.857e-02  2.951e-01  -0.232  0.81627
## jobhousemaid        -2.042e-01  2.988e-01  -0.684  0.49427
## jobmanagement        2.123e-01  2.659e-01   0.799  0.42448
## jobretired           5.813e-01  2.727e-01   2.132  0.03304 *
## `jobself-employed`   2.009e-01  2.833e-01   0.709  0.47819
## jobservices          6.568e-02  2.735e-01   0.240  0.81021
## jobstudent           7.528e-01  2.825e-01   2.665  0.00769 **
## jobtechnician        1.799e-01  2.659e-01   0.677  0.49866
## jobunemployed        1.390e-01  2.852e-01   0.487  0.62599
## jobunknown                  NA         NA      NA       NA
## salary                      NA         NA      NA       NA
## maritaldivorced     -4.294e-02  8.107e-02  -0.530  0.59636
## maritalmarried      -2.074e-01  6.652e-02  -3.118  0.00182 **
## maritalsingle               NA         NA      NA       NA
## educationprimary    -2.118e-01  1.511e-01  -1.402  0.16080
## educationsecondary -9.074e-02  1.418e-01  -0.640  0.52218
## educationtertiary    8.320e-02  1.224e-01   0.680  0.49654
## educationunknown            NA         NA      NA       NA
## targetedno           8.358e-03  9.161e-02   0.091  0.92731
## targetedyes                 NA         NA      NA       NA
## defaultno           -8.588e-03  1.928e-01  -0.045  0.96447
## defaultyes                  NA         NA      NA       NA
## balance              1.529e-05  6.019e-06   2.539  0.01111 *
## housingno            7.693e-01  5.306e-02  14.500  < 2e-16 ***
## housingyes                  NA         NA      NA       NA
## loanno               3.636e-01  7.175e-02   5.067 4.04e-07 ***
## loanyes                     NA         NA      NA       NA
## contactcellular      1.643e+00  8.798e-02  18.670  < 2e-16 ***
## contacttelephone     1.509e+00  1.210e-01  12.475  < 2e-16 ***
## contactunknown              NA         NA      NA       NA
## day                  7.950e-03  2.977e-03   2.671  0.00757 **
## monthapr            -8.189e-01  1.425e-01  -5.745 9.21e-09 ***
## monthaug            -1.530e+00  1.372e-01 -11.149  < 2e-16 ***
## monthdec            -1.471e-01  2.332e-01  -0.631  0.52825
## monthfeb            -9.456e-01  1.441e-01  -6.562 5.31e-11 ***
## monthjan            -2.124e+00  1.817e-01 -11.686  < 2e-16 ***
## monthjul            -1.659e+00  1.406e-01 -11.798  < 2e-16 ***
## monthjun            -3.797e-01  1.473e-01  -2.577  0.00996 **
## monthmar             7.240e-01  1.733e-01   4.177 2.95e-05 ***
## monthmay            -1.191e+00  1.366e-01  -8.725  < 2e-16 ***
## monthnov            -1.678e+00  1.456e-01 -11.524  < 2e-16 ***
## monthoct             1.125e-01  1.627e-01   0.692  0.48908
## monthsep                    NA         NA      NA       NA
## duration             4.272e-03  7.820e-05  54.631  < 2e-16 ***
## campaign            -9.677e-02  1.230e-02  -7.866 3.65e-15 ***
## pdays                5.737e-04  3.556e-04   1.613  0.10667
```

```
## previous              1.169e-02  7.290e-03   1.604  0.10872
## poutcomefailure      -6.733e-02  1.113e-01  -0.605  0.54529
## poutcomeother         4.195e-02  1.279e-01   0.328  0.74290
## poutcomesuccess       2.269e+00  1.007e-01  22.535  < 2e-16 ***
## poutcomeunknown              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15016  on 31603  degrees of freedom
## AIC: 15104
##
## Number of Fisher Scoring iterations: 6
```

```
#stepAIC(logistic_1, direction = "both")
# stepAIC has removed some variables and only the following ones remain
logistic_2 <- glm(formula = response ~ jobadmin. + jobhousemaid +
jobmanagement +
    jobretired + jobstudent + jobtechnician + maritalmarried +
    educationprimary + educationsecondary + balance + housingno +
    loanno + contactcellular + contacttelephone + day + monthapr +
    monthaug + monthfeb + monthjan + monthjul + monthjun + monthmar +
    monthmay + monthnov + duration + campaign + pdays + previous +
    poutcomesuccess, family = "binomial", data = train)

# checking vif for logistic_2
vif(logistic_2)
```

```
##          jobadmin.       jobhousemaid      jobmanagement
##           1.279574           1.075275           1.848829
##         jobretired         jobstudent      jobtechnician
##           1.269010           1.186903           1.356935
##     maritalmarried   educationprimary educationsecondary
##           1.094159           1.481352           1.639156
##            balance          housingno             loanno
##           1.033559           1.410136           1.057855
##    contactcellular   contacttelephone                day
##           2.472082           1.951957           1.315402
##           monthapr           monthaug           monthfeb
##           2.146210           2.555384           1.980877
##           monthjan           monthjul           monthjun
##           1.398465           2.495161           2.511823
##           monthmar           monthmay           monthnov
##           1.337560           3.180751           1.937072
##           duration           campaign              pdays
##           1.131015           1.102743           1.357429
##           previous    poutcomesuccess
##           1.161571           1.133248
```

```
summary(logistic_2)

##
## Call:
## glm(formula = response ~ jobadmin. + jobhousemaid + jobmanagement +
##     jobretired + jobstudent + jobtechnician + maritalmarried +
##     educationprimary + educationsecondary + balance + housingno +
##     loanno + contactcellular + contacttelephone + day + monthapr +
##     monthaug + monthfeb + monthjan + monthjul + monthjun + monthmar +
##     monthmay + monthnov + duration + campaign + pdays + previous +
##     poutcomesuccess, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.7557  -0.3727  -0.2504  -0.1459   3.4618
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.406e+00  1.614e-01 -27.295  < 2e-16 ***
## jobadmin.           2.964e-01  7.371e-02   4.021 5.79e-05 ***
## jobhousemaid       -2.749e-01  1.541e-01  -1.784 0.074388 .
## jobmanagement       1.560e-01  6.858e-02   2.274 0.022951 *
## jobretired          4.643e-01  8.931e-02   5.198 2.01e-07 ***
## jobstudent          7.148e-01  1.171e-01   6.101 1.05e-09 ***
## jobtechnician       1.234e-01  6.775e-02   1.821 0.068580 .
## maritalmarried     -2.172e-01  4.532e-02  -4.792 1.65e-06 ***
## educationprimary   -3.120e-01  8.208e-02  -3.800 0.000144 ***
## educationsecondary -1.738e-01  5.534e-02  -3.141 0.001685 **
## balance             1.519e-05  5.983e-06   2.539 0.011128 *
## housingno           7.677e-01  5.242e-02  14.646  < 2e-16 ***
## loanno              3.658e-01  7.136e-02   5.126 2.95e-07 ***
## contactcellular     1.651e+00  8.739e-02  18.894  < 2e-16 ***
## contacttelephone    1.503e+00  1.199e-01  12.534  < 2e-16 ***
## day                 8.415e-03  2.956e-03   2.847 0.004420 **
## monthapr           -8.538e-01  1.037e-01  -8.232  < 2e-16 ***
## monthaug           -1.560e+00  9.726e-02 -16.039  < 2e-16 ***
## monthfeb           -9.689e-01  1.087e-01  -8.910  < 2e-16 ***
## monthjan           -2.153e+00  1.516e-01 -14.204  < 2e-16 ***
## monthjul           -1.692e+00  1.003e-01 -16.878  < 2e-16 ***
## monthjun           -4.069e-01  1.128e-01  -3.608 0.000309 ***
## monthmar            6.990e-01  1.440e-01   4.854 1.21e-06 ***
## monthmay           -1.222e+00  9.672e-02 -12.638  < 2e-16 ***
## monthnov           -1.717e+00  1.077e-01 -15.931  < 2e-16 ***
## duration            4.272e-03  7.812e-05  54.690  < 2e-16 ***
## campaign           -9.724e-02  1.229e-02  -7.915 2.48e-15 ***
## pdays               4.559e-04  2.182e-04   2.089 0.036703 *
## previous            1.183e-02  7.001e-03   1.690 0.091013 .
## poutcomesuccess     2.290e+00  8.089e-02  28.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15025  on 31617  degrees of freedom
## AIC: 15085
##
## Number of Fisher Scoring iterations: 6
```

```r
# removing monthmay since vif is high
logistic_3 <- glm(formula = response ~ jobadmin. + jobhousemaid +
jobmanagement +
    jobretired + jobstudent + jobtechnician + maritalmarried +
    educationprimary + educationsecondary + balance + housingno +
    loanno + contactcellular + contacttelephone + day + monthapr +
    monthaug + monthfeb + monthjan + monthjul + monthjun + monthmar +
     + monthnov + duration + campaign + pdays + previous +
    poutcomesuccess, family = "binomial", data = train)

summary(logistic_3)
```

```
##
## Call:
## glm(formula = response ~ jobadmin. + jobhousemaid + jobmanagement +
##     jobretired + jobstudent + jobtechnician + maritalmarried +
##     educationprimary + educationsecondary + balance + housingno +
##     loanno + contactcellular + contacttelephone + day + monthapr +
##     monthaug + monthfeb + monthjan + monthjul + monthjun + monthmar +
##     +monthnov + duration + campaign + pdays + previous + poutcomesuccess,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6947  -0.3816  -0.2524  -0.1457   3.4782
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.564e+00  1.343e-01 -41.441  < 2e-16 ***
## jobadmin.           3.234e-01  7.321e-02   4.418 9.98e-06 ***
## jobhousemaid       -2.133e-01  1.530e-01  -1.394 0.163357
## jobmanagement       1.885e-01  6.815e-02   2.766 0.005680 **
## jobretired          5.705e-01  8.814e-02   6.473 9.61e-11 ***
## jobstudent          7.570e-01  1.161e-01   6.523 6.88e-11 ***
## jobtechnician       1.505e-01  6.744e-02   2.232 0.025612 *
## maritalmarried     -1.959e-01  4.503e-02  -4.351 1.35e-05 ***
## educationprimary   -3.293e-01  8.158e-02  -4.037 5.41e-05 ***
## educationsecondary -1.842e-01  5.502e-02  -3.347 0.000817 ***
## balance             1.712e-05  5.917e-06   2.894 0.003800 **
## housingno           9.394e-01  5.098e-02  18.425  < 2e-16 ***
## loanno              3.815e-01  7.118e-02   5.359 8.36e-08 ***
```

```
## contactcellular     1.834e+00  8.457e-02  21.681  < 2e-16 ***
## contacttelephone    1.716e+00  1.173e-01  14.633  < 2e-16 ***
## day                 9.709e-03  2.951e-03   3.289 0.001004 **
## monthapr           -4.432e-03  8.090e-02  -0.055 0.956316
## monthaug           -7.991e-01  7.815e-02 -10.225  < 2e-16 ***
## monthfeb           -1.644e-01  9.005e-02  -1.826 0.067853 .
## monthjan           -1.392e+00  1.404e-01  -9.915  < 2e-16 ***
## monthjul           -8.762e-01  7.845e-02 -11.168  < 2e-16 ***
## monthjun            4.922e-01  8.870e-02   5.549 2.87e-08 ***
## monthmar            1.467e+00  1.328e-01  11.045  < 2e-16 ***
## monthnov           -9.104e-01  8.835e-02 -10.305  < 2e-16 ***
## duration            4.246e-03  7.769e-05  54.648  < 2e-16 ***
## campaign           -1.018e-01  1.235e-02  -8.240  < 2e-16 ***
## pdays               4.146e-04  2.194e-04   1.890 0.058793 .
## previous            1.414e-02  7.667e-03   1.844 0.065154 .
## poutcomesuccess     2.354e+00  8.025e-02  29.327  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15183  on 31618  degrees of freedom
## AIC: 15241
##
## Number of Fisher Scoring iterations: 6
```

```
vif(logistic_3)
```

```
##          jobadmin.       jobhousemaid      jobmanagement
##           1.278345           1.073900           1.852500
##         jobretired         jobstudent      jobtechnician
##           1.255725           1.186584           1.355827
##     maritalmarried   educationprimary educationsecondary
##           1.093686           1.478612           1.640305
##            balance          housingno             loanno
##           1.034041           1.347969           1.057721
##    contactcellular   contacttelephone                day
##           2.331771           1.885092           1.314812
##           monthapr           monthaug           monthfeb
##           1.295233           1.649712           1.344307
##           monthjan           monthjul           monthjun
##           1.191206           1.524433           1.517659
##           monthmar           monthnov           duration
##           1.119970           1.296945           1.127920
##           campaign              pdays           previous
##           1.100228           1.390133           1.197572
##    poutcomesuccess
##           1.136761
```

```r
# all vifs below 3 now, so removing variables based on significance level
# removing jobhousemaid

logistic_4 <- glm(formula = response ~ jobadmin. + jobmanagement +
    jobretired + jobstudent + jobtechnician + maritalmarried +
    educationprimary + educationsecondary + balance + housingno +
    loanno + contactcellular + contacttelephone + day + monthapr +
    monthaug + monthfeb + monthjan + monthjul + monthjun + monthmar +
     + monthnov + duration + campaign + pdays + previous +
    poutcomesuccess, family = "binomial", data = train)

summary(logistic_4)

##
## Call:
## glm(formula = response ~ jobadmin. + jobmanagement + jobretired +
##     jobstudent + jobtechnician + maritalmarried + educationprimary +
##     educationsecondary + balance + housingno + loanno + contactcellular +
##     contacttelephone + day + monthapr + monthaug + monthfeb +
##     monthjan + monthjul + monthjun + monthmar + +monthnov + duration +
##     campaign + pdays + previous + poutcomesuccess, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6895  -0.3818  -0.2529  -0.1457   3.4820
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.569e+00  1.342e-01 -41.485  < 2e-16 ***
## jobadmin.           3.346e-01  7.282e-02   4.594 4.34e-06 ***
## jobmanagement       2.019e-01  6.753e-02   2.990 0.002790 **
## jobretired          5.886e-01  8.724e-02   6.747 1.51e-11 ***
## jobstudent          7.701e-01  1.157e-01   6.656 2.82e-11 ***
## jobtechnician       1.628e-01  6.691e-02   2.433 0.014984 *
## maritalmarried     -1.967e-01  4.503e-02  -4.367 1.26e-05 ***
## educationprimary   -3.408e-01  8.122e-02  -4.196 2.71e-05 ***
## educationsecondary -1.823e-01  5.500e-02  -3.314 0.000919 ***
## balance             1.714e-05  5.919e-06   2.896 0.003784 **
## housingno           9.350e-01  5.088e-02  18.376  < 2e-16 ***
## loanno              3.804e-01  7.119e-02   5.342 9.17e-08 ***
## contactcellular     1.833e+00  8.459e-02  21.671  < 2e-16 ***
## contacttelephone    1.713e+00  1.173e-01  14.606  < 2e-16 ***
## day                 9.628e-03  2.951e-03   3.263 0.001104 **
## monthapr           -4.825e-03  8.091e-02  -0.060 0.952441
## monthaug           -8.037e-01  7.807e-02 -10.296  < 2e-16 ***
## monthfeb           -1.650e-01  9.003e-02  -1.833 0.066842 .
## monthjan           -1.388e+00  1.403e-01  -9.896  < 2e-16 ***
## monthjul           -8.788e-01  7.845e-02 -11.202  < 2e-16 ***
## monthjun            4.877e-01  8.865e-02   5.501 3.78e-08 ***
```

```
## monthmar            1.464e+00  1.328e-01  11.025  < 2e-16 ***
## monthnov           -9.111e-01  8.834e-02 -10.314  < 2e-16 ***
## duration            4.243e-03  7.766e-05  54.641  < 2e-16 ***
## campaign           -1.015e-01  1.234e-02  -8.222  < 2e-16 ***
## pdays               4.166e-04  2.193e-04   1.900 0.057494 .
## previous            1.408e-02  7.651e-03   1.840 0.065727 .
## poutcomesuccess     2.354e+00  8.023e-02  29.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15185  on 31619  degrees of freedom
## AIC: 15241
##
## Number of Fisher Scoring iterations: 6
```

```
vif(logistic_4) # vifs are all below 3
```

```
##          jobadmin.      jobmanagement         jobretired
##           1.265142           1.818987           1.230380
##         jobstudent      jobtechnician     maritalmarried
##           1.179704           1.335063           1.093653
##   educationprimary educationsecondary            balance
##           1.464271           1.639060           1.034000
##          housingno             loanno    contactcellular
##           1.342635           1.057621           2.333056
##   contacttelephone                day           monthapr
##           1.884936           1.314397           1.295365
##           monthaug           monthfeb           monthjan
##           1.647107           1.344399           1.190980
##           monthjul           monthjun           monthmar
##           1.523644           1.516960           1.119701
##           monthnov           duration           campaign
##           1.297082           1.126930           1.099934
##              pdays           previous    poutcomesuccess
##           1.389528           1.196844           1.136652
```

```
# removing monthapr, monthfeb, pdays, previous
logistic_5 <- glm(formula = response ~ jobadmin. + jobmanagement +
    jobretired + jobstudent + jobtechnician + maritalmarried +
    educationprimary + educationsecondary + balance + housingno +
    loanno + contactcellular + contacttelephone + day +
    monthaug + monthjan + monthjul + monthjun + monthmar +
     + monthnov + duration + campaign  +
    poutcomesuccess, family = "binomial", data = train)

summary(logistic_5)
```

```
##
## Call:
## glm(formula = response ~ jobadmin. + jobmanagement + jobretired +
##     jobstudent + jobtechnician + maritalmarried + educationprimary +
##     educationsecondary + balance + housingno + loanno + contactcellular +
##     contacttelephone + day + monthaug + monthjan + monthjul +
##     monthjun + monthmar + +monthnov + duration + campaign +
poutcomesuccess,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6847  -0.3811  -0.2531  -0.1468   3.4937
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.587e+00  1.326e-01 -42.146  < 2e-16 ***
## jobadmin.           3.410e-01  7.276e-02   4.686 2.78e-06 ***
## jobmanagement       2.061e-01  6.746e-02   3.055  0.00225 **
## jobretired          5.915e-01  8.716e-02   6.787 1.15e-11 ***
## jobstudent          7.819e-01  1.156e-01   6.763 1.35e-11 ***
## jobtechnician       1.630e-01  6.690e-02   2.437  0.01482 *
## maritalmarried     -1.928e-01  4.497e-02  -4.288 1.80e-05 ***
## educationprimary   -3.370e-01  8.117e-02  -4.152 3.29e-05 ***
## educationsecondary -1.787e-01  5.494e-02  -3.253  0.00114 **
## balance             1.724e-05  5.909e-06   2.917  0.00353 **
## housingno           8.965e-01  4.937e-02  18.159  < 2e-16 ***
## loanno              3.806e-01  7.115e-02   5.349 8.84e-08 ***
## contactcellular     1.867e+00  8.052e-02  23.183  < 2e-16 ***
## contacttelephone    1.745e+00  1.147e-01  15.213  < 2e-16 ***
## day                 1.090e-02  2.785e-03   3.915 9.04e-05 ***
## monthaug           -7.961e-01  7.053e-02 -11.287  < 2e-16 ***
## monthjan           -1.382e+00  1.372e-01 -10.078  < 2e-16 ***
## monthjul           -8.852e-01  7.136e-02 -12.405  < 2e-16 ***
## monthjun            5.168e-01  8.586e-02   6.019 1.75e-09 ***
## monthmar            1.493e+00  1.292e-01  11.561  < 2e-16 ***
## monthnov           -9.040e-01  8.293e-02 -10.900  < 2e-16 ***
## duration            4.241e-03  7.748e-05  54.728  < 2e-16 ***
## campaign           -1.030e-01  1.229e-02  -8.383  < 2e-16 ***
## poutcomesuccess     2.433e+00  7.682e-02  31.668  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15200  on 31623  degrees of freedom
## AIC: 15248
##
## Number of Fisher Scoring iterations: 6
```

```r
vif(logistic_5)

##          jobadmin.      jobmanagement         jobretired 
##           1.264792           1.817724           1.230185 
##         jobstudent      jobtechnician     maritalmarried 
##           1.177893           1.334885           1.091911 
##   educationprimary educationsecondary            balance 
##           1.463544           1.636925           1.033673 
##          housingno             loanno    contactcellular 
##           1.265137           1.056964           2.115809 
##   contacttelephone                day           monthaug 
##           1.806276           1.178723           1.346639 
##           monthjan           monthjul           monthjun 
##           1.137174           1.264032           1.423663 
##           monthmar           monthnov           duration 
##           1.062195           1.144795           1.123197 
##           campaign    poutcomesuccess 
##           1.092087           1.041965

#removing jobtechnician
logistic_6 <- glm(formula = response ~ jobadmin. + jobmanagement + 
    jobretired + jobstudent + maritalmarried + 
    educationprimary + educationsecondary + balance + housingno + 
    loanno + contactcellular + contacttelephone + day + 
    monthaug + monthjan + monthjul + monthjun + monthmar + 
     + monthnov + duration + campaign  + 
    poutcomesuccess, family = "binomial", data = train)

summary(logistic_6)

## 
## Call:
## glm(formula = response ~ jobadmin. + jobmanagement + jobretired + 
##     jobstudent + maritalmarried + educationprimary + educationsecondary + 
##     balance + housingno + loanno + contactcellular + contacttelephone + 
##     day + monthaug + monthjan + monthjul + monthjun + monthmar + 
##     +monthnov + duration + campaign + poutcomesuccess, family = 
"binomial", 
##     data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.6982  -0.3810  -0.2534  -0.1468   3.4930  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -5.531e+00  1.304e-01 -42.419  < 2e-16 ***
## jobadmin.             2.856e-01  6.893e-02   4.144 3.42e-05 ***
## jobmanagement         1.448e-01  6.241e-02   2.321 0.020292 *  
## jobretired            5.463e-01  8.499e-02   6.427 1.30e-10 ***
```

```
## jobstudent          7.216e-01  1.128e-01   6.396 1.60e-10 ***
## maritalmarried     -1.989e-01  4.491e-02  -4.429 9.48e-06 ***
## educationprimary   -3.765e-01  7.942e-02  -4.741 2.13e-06 ***
## educationsecondary -1.838e-01  5.492e-02  -3.346 0.000818 ***
## balance             1.727e-05  5.895e-06   2.929 0.003395 **
## housingno           8.976e-01  4.938e-02  18.178  < 2e-16 ***
## loanno              3.794e-01  7.113e-02   5.333 9.64e-08 ***
## contactcellular     1.871e+00  8.055e-02  23.234  < 2e-16 ***
## contacttelephone    1.746e+00  1.147e-01  15.226  < 2e-16 ***
## day                 1.111e-02  2.784e-03   3.990 6.59e-05 ***
## monthaug           -7.763e-01  7.008e-02 -11.078  < 2e-16 ***
## monthjan           -1.384e+00  1.371e-01 -10.095  < 2e-16 ***
## monthjul           -8.868e-01  7.136e-02 -12.426  < 2e-16 ***
## monthjun            5.181e-01  8.590e-02   6.031 1.63e-09 ***
## monthmar            1.499e+00  1.292e-01  11.609  < 2e-16 ***
## monthnov           -9.047e-01  8.290e-02 -10.912  < 2e-16 ***
## duration            4.237e-03  7.744e-05  54.717  < 2e-16 ***
## campaign           -1.033e-01  1.228e-02  -8.405  < 2e-16 ***
## poutcomesuccess     2.436e+00  7.680e-02  31.719  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15206  on 31624  degrees of freedom
## AIC: 15252
##
## Number of Fisher Scoring iterations: 6
```

```r
# removing jobmanagement
logistic_7 <- glm(formula = response ~ jobadmin. +
    jobretired + jobstudent + maritalmarried +
    educationprimary + educationsecondary + balance + housingno +
    loanno + contactcellular + contacttelephone + day +
    monthaug + monthjan + monthjul + monthjun + monthmar +
     + monthnov + duration + campaign  +
    poutcomesuccess, family = "binomial", data = train)

summary(logistic_7)
```

```
##
## Call:
## glm(formula = response ~ jobadmin. + jobretired + jobstudent +
##     maritalmarried + educationprimary + educationsecondary +
##     balance + housingno + loanno + contactcellular + contacttelephone +
##     day + monthaug + monthjan + monthjul + monthjun + monthmar +
##     +monthnov + duration + campaign + poutcomesuccess, family =
"binomial",
##     data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6989  -0.3814  -0.2541  -0.1467   3.4880
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.461e+00  1.266e-01 -43.123  < 2e-16 ***
## jobadmin.           2.570e-01  6.773e-02   3.794 0.000148 ***
## jobretired          5.138e-01  8.376e-02   6.134 8.57e-10 ***
## jobstudent          6.756e-01  1.110e-01   6.085 1.17e-09 ***
## maritalmarried     -1.970e-01  4.489e-02  -4.388 1.15e-05 ***
## educationprimary   -4.411e-01  7.419e-02  -5.945 2.76e-09 ***
## educationsecondary -2.457e-01  4.777e-02  -5.143 2.70e-07 ***
## balance             1.757e-05  5.890e-06   2.983 0.002854 **
## housingno           8.985e-01  4.938e-02  18.196  < 2e-16 ***
## loanno              3.806e-01  7.113e-02   5.351 8.72e-08 ***
## contactcellular     1.878e+00  8.052e-02  23.328  < 2e-16 ***
## contacttelephone    1.750e+00  1.147e-01  15.258  < 2e-16 ***
## day                 1.113e-02  2.784e-03   3.997 6.42e-05 ***
## monthaug           -7.735e-01  7.006e-02 -11.040  < 2e-16 ***
## monthjan           -1.388e+00  1.371e-01 -10.122  < 2e-16 ***
## monthjul           -8.876e-01  7.135e-02 -12.440  < 2e-16 ***
## monthjun            5.188e-01  8.592e-02   6.038 1.56e-09 ***
## monthmar            1.506e+00  1.291e-01  11.664  < 2e-16 ***
## monthnov           -9.020e-01  8.292e-02 -10.878  < 2e-16 ***
## duration            4.234e-03  7.739e-05  54.718  < 2e-16 ***
## campaign           -1.029e-01  1.227e-02  -8.389  < 2e-16 ***
## poutcomesuccess     2.436e+00  7.676e-02  31.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22840  on 31646  degrees of freedom
## Residual deviance: 15211  on 31625  degrees of freedom
## AIC: 15255
##
## Number of Fisher Scoring iterations: 6
```

```
vif(logistic_7)
```

```
##          jobadmin.         jobretired         jobstudent
##           1.095196           1.136371           1.085759
##     maritalmarried   educationprimary educationsecondary
##           1.088615           1.226031           1.238695
##            balance          housingno             loanno
##           1.032898           1.266552           1.057005
##    contactcellular   contacttelephone                day
##           2.116738           1.808209           1.177530
```

```
##          monthaug          monthjan          monthjul 
##          1.327510          1.137262          1.264405 
##          monthjun          monthmar          monthnov 
##          1.426528          1.061397          1.144361 
##          duration          campaign   poutcomesuccess 
##          1.121823          1.091999          1.041863
```

```
logistic_8 <- glm(formula = response ~ jobadmin. + 
    jobretired + jobstudent + maritalmarried + 
    educationprimary + educationsecondary + balance + housingno + 
    loanno + day + 
    duration + campaign  + 
    poutcomesuccess, family = "binomial", data = train)

logistic_final <- logistic_8
```

We now have a logistic model named logistic_final. Next, we'll use the model to predict the response in the test data.

```
predictions_logit <- predict(logistic_final, newdata = test[, -55], type = 
"response")
summary(predictions_logit)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 0.0000624 0.0275900 0.0522100 0.1165000 0.1111000 1.0000000
```

So now we have predicted the 'probabilities of responding' (for the test data). Note that the average probability (as shown above in summary(predictions_logit) is 11.6%, which is the average response rate.

Next comes the interesting part. We need to convert the probabilities to an actual prediction, i.e. **yes or no**. Can we just say that anything *above 50% probability of response is yes and no otherwise*? Yes, we could, but we can do better.

We can rather experiment with other **probability thresholds** like 30%, 40% etc. We will go with whatever gives us the highest (loosely speaking) **accuracy**. In fact, apart from accuracy, there are other metrics to **evaluate the model** like sensitivity, specificity etc.

## Model Evaluation

In model evaluation, we use the test data to evaluate how good the model is (note that it was trained on 'train data' and hasn't seen the test data, so we are not cheating).

Let us first look at how **accurate** the predictions are. For now, let's use a probability cutoff of 50% and then we'll iterate.

```
predicted_response <- factor(ifelse(predictions_logit >= 0.50, "yes", "no"))
conf <- confusionMatrix(predicted_response, test$response, positive = "yes")
conf
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    no   yes
##        no  11705  1083
##        yes   272   504
##                                           
##                Accuracy : 0.9001          
##                  95% CI : (0.8949, 0.9051)
##     No Information Rate : 0.883           
##     P-Value [Acc > NIR] : 1.306e-10       
##                                           
##                   Kappa : 0.3788          
##  Mcnemar's Test P-Value : < 2.2e-16       
##                                           
##             Sensitivity : 0.31758         
##             Specificity : 0.97729         
##          Pos Pred Value : 0.64948         
##          Neg Pred Value : 0.91531         
##              Prevalence : 0.11700         
##          Detection Rate : 0.03716         
##    Detection Prevalence : 0.05721         
##       Balanced Accuracy : 0.64744         
##                                           
##        'Positive' Class : yes             
## 
```

Firstly, note that the **accuracy** is approx. 90% which means that the model has made about 90% predictions correct (whether yes or no).

There are two other important metrics - **sensitivity** and **specificity**.

**Sensitivity** is the fraction of correctly identified responses, i.e. out of those who will actually respond, how many has the model identified.

**Specificity** is the fraction of **incorrectly identified responses**, i.e. out of those who will actually NOT respond, how many has the model identified.

These two metrics can be calculated using the table of predictions as follows:

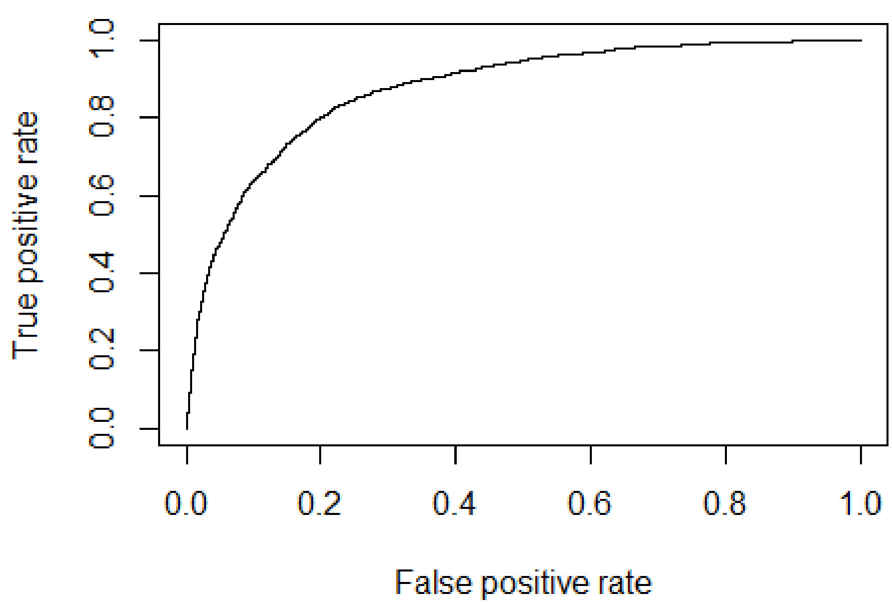
```
sensitivity <- conf$byClass[1]
specificity <- conf$byClass[2]
sensitivity

## Sensitivity 
##   0.3175803

specificity

## Specificity 
##   0.9772898
```

The values of sensitivity and specificity are about 31.75% and 97.72% respectively. This means that the model predicts 97.72% of those who will NOT buy correctly while only 31.75% of those who'll buy.

Since the number of "yes" responders are few (only 11% respond), it is hard to predict them. So if you market the product to about 10,000 people, you know that about 1100 will respond. The model will identify about 31% or 350 of them correctly.

But these predictions are based on an arbitrary cut-off of 0.50 probability. Now that we know what accuracy, specificity and sensitivity mean, we can find a cutoff which optimises the most important metric. In our case, it is sensitivity.

```
library(ROCR)

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

predictions_object <- prediction(predictions_logit, test$response)
perf_object <- performance(predictions_object, "tpr", "fpr")
plot(perf_object)
```

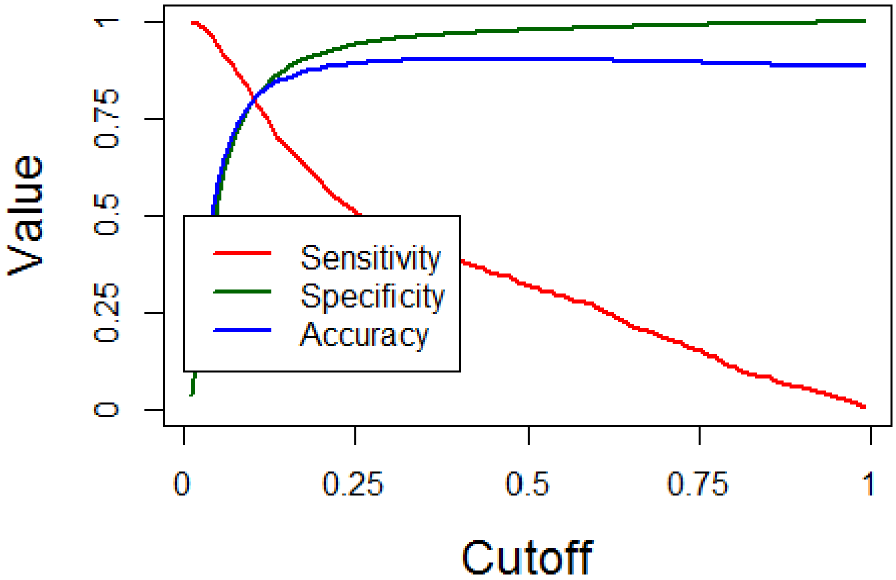


The plot shown above is called the **ROC** curve. It has False Positive Rate (FPR) and True Positive Rate (TPR, or sensitivity) on the x and y axes respectively.

The objective is to **maximise the TPR** and **minimise the FPR** which means that we want the curve to be aligned towards the **top-left**.

Now, let's find oput the optimal probabilty cutoff, i.e. the value abpve which we'll predict "yes" and "no" otherwise. We can plot the three metrics against cutoff values ranging from 0% to 100% and choose the one which gives high accuracy, sensitivity and specificity.

```
perform_fn <- function(cutoff)
  {
  predicted_response <- factor(ifelse(predictions_logit >= cutoff, "yes",
"no"))
  conf <- confusionMatrix(predicted_response, test$response, positive =
"yes")
  acc <- conf$overall[1]
  sens <- conf$byClass[1]
  spec <- conf$byClass[2]
  out <- t(as.matrix(c(sens, spec, acc)))
  colnames(out) <- c("sensitivity", "specificity", "accuracy")
  return(out)
}

# creating cutoff values from 0.01 to 0.99 for plotting and initialising a
matrix of size 1000x4
s = seq(.01,.99,length=100)
OUT = matrix(0,100,3)

# calculate the sens, spec and acc for different cutoff values
for(i in 1:100)
  {
  OUT[i,] = perform_fn(s[i])
}


# plotting cutoffs
plot(s,
OUT[,1],xlab="Cutoff",ylab="Value",cex.lab=1.5,cex.axis=1.5,ylim=c(0,1),type=
"l",lwd=2,axes=FALSE,col=2)
axis(1,seq(0,1,length=5),seq(0,1,length=5),cex.lab=1.5)
axis(2,seq(0,1,length=5),seq(0,1,length=5),cex.lab=1.5)
lines(s,OUT[,2],col="darkgreen",lwd=2)
lines(s,OUT[,3],col=4,lwd=2)
box()
legend(0,.50,col=c(2,"darkgreen",4,"darkred"),lwd=c(2,2,2,2),c("Sensitivity",
"Specificity","Accuracy"))
```




The plot above shows the sensitivity, specificity and accuracy for cutoff probabilities ranging from 0 to 100. It is clear that a cutoff around 12-13% will optimise the three metrics.

```
cutoffs <- s[which(abs(OUT[, 1] - OUT[, 2]) < 0.01)]
```

Let's choose a cutoff value of 12% for the final model.

```
predicted_response <- factor(ifelse(predictions_logit >= 0.12, "yes", "no"))
conf_final <- confusionMatrix(predicted_response, test$response, positive =
"yes")
acc <- conf_final$overall[1]
sens <- conf_final$byClass[1]
spec <- conf_final$byClass[2]
acc

##  Accuracy
## 0.8262312

sens

## Sensitivity
##   0.7529931

spec

## Specificity
##   0.8359355
```

We have accuracy = 82.62%, sensitivity = 75.29% and specificity = 83.59%. This is a remarkable improvement over cutoff = 0.50, where the sensitivity was around 31% only.

Now, if you market the product to 10,000 people (out of which around 1100 usually respond), the model will be able to identify 75% of 1100 or approx. 825 people correctly.

## Model Deployment and Recommendations

Now that we have a model which predicts the probability of response, we can arrive at some interesting recommendations.

Our objective is to reduce the marketing cost and get almost the same number of customers as before.

The usual response rate is 11%, which means that if we telemarket to 10,000 people, 1100 will buy the product.

We can rather telemarket to only thoso whose **probability of purchase is high**. Let's look at the probabilities of purchase. Note that we will use only test data for this analysis.

```
test$predicted_probs <- predictions_logit
test$predicted_response <- predicted_response
str(test)
```

```
## 'data.frame':    13564 obs. of  57 variables:
##  $ age               : int  47 35 28 57 45 57 33 28 46 51 ...
##  $ jobadmin.         : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ jobblue-collar    : int  1 0 0 0 0 1 0 1 0 0 ...
##  $ jobentrepreneur   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobhousemaid      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobmanagement     : int  0 1 1 0 0 0 0 0 1 1 ...
##  $ jobretired        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobself-employed  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobservices       : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ jobstudent        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobtechnician     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobunemployed     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jobunknown        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ salary            : int  20000 100000 100000 70000 50000 20000 70000
20000 100000 100000 ...
##  $ maritaldivorced   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ maritalmarried    : int  1 1 0 1 0 1 1 1 0 1 ...
##  $ maritalsingle     : int  0 0 1 0 1 0 0 0 1 0 ...
##  $ educationprimary  : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ educationsecondary: int  0 0 0 1 0 0 1 1 1 0 ...
##  $ educationtertiary : int  0 1 1 0 0 0 0 0 0 1 ...
##  $ educationunknown  : int  1 0 0 0 1 0 0 0 0 0 ...
##  $ targetedno        : int  1 0 1 0 1 0 0 0 0 0 ...
##  $ targetedyes       : int  0 1 0 1 0 1 1 1 1 1 ...
##  $ defaultno         : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ defaultyes        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ balance           : int  1506 231 447 162 13 52 0 723 -246 10635 ...
##  $ housingno         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ housingyes        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ loanno            : int  1 1 0 1 1 1 1 0 1 1 ...
##  $ loanyes           : int  0 0 1 0 0 0 0 1 0 0 ...
##  $ contactcellular   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contacttelephone  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contactunknown    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day               : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ monthapr          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthaug          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthdec          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthfeb          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthjan          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthjul          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthjun          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthmar          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthmay          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ monthnov          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthoct          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthsep          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ duration          : int  92 139 217 174 98 38 54 262 255 336 ...
##  $ campaign          : int  1 1 1 1 1 1 1 1 2 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 ...
##  $ poutcomefailure   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcomeother     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcomesuccess   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcomeunknown   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ response          : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1
...
##  $ predicted_probs   : num  0.0317 0.037 0.0391 0.031 0.0552 ...
##  $ predicted_response: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1
...
##  - attr(*, "dummies")=List of 10
##   ..$ job      : int  2 3 4 5 6 7 8 9 10 11 ...
##   ..$ marital  : int  15 16 17
##   ..$ education: int  18 19 20 21
##   ..$ targeted : int  22 23
##   ..$ default  : int  24 25
##   ..$ housing  : int  27 28
##   ..$ loan     : int  29 30
##   ..$ contact  : int  31 32 33
##   ..$ month    : int  35 36 37 38 39 40 41 42 43 44 ...
##   ..$ poutcome : int  51 52 53 54

test_predictions <- test[, c("response", "predicted_probs",
"predicted_response")]
head(test_predictions)
```

```
##    response predicted_probs predicted_response
## 4        no      0.03167298                 no
## 6        no      0.03701600                 no
## 7        no      0.03908693                 no
## 15       no      0.03095186                 no
## 17       no      0.05521066                 no
## 18       no      0.01445412                 no
```

```r
write.csv(test_predictions, file = "response_predictions.csv")
```

We have 13,564 observations in test data. Since we now have the probabilities of response, we can sort them and market only to those with high probabilities.

### Reducing Customer Acquision Cost

Let's assume that telemarketing to each person costs INR 1. In the test data, we have 13,564 observations, so the total cost is INR 13564.
Among these, about 11.7% respond, so we get 1587 customers for INR 13564, or Rs 8.54 per customer.

```r
summary(test_predictions$response)

##    no   yes
## 11977  1587
```

Let's sort the observations in decreasing order of probability.

```r
test_predictions <- test_predictions[order(test_predictions$predicted_probs,
decreasing = T), ]
head(test_predictions)

##       response predicted_probs predicted_response
## 24149       no       0.9999999                yes
## 24096       no       0.9999634                yes
## 2387        no       0.9998842                yes
## 24055       no       0.9996612                yes
## 10727      yes       0.9996206                yes
## 31501       no       0.9994970                yes
```

Now if we market to, say, only 50% population (approx. 6800 people), then about 1500 will respond (see below). The response rate is improved to 22.6%, almost double of what you'll get by randomly marketing. The acquision cost comes down to Rs 4.5 per customer.

```r
summary(test_predictions$response[1:6800])

##   no  yes
## 5321 1479

1479/6800

## [1] 0.2175
```

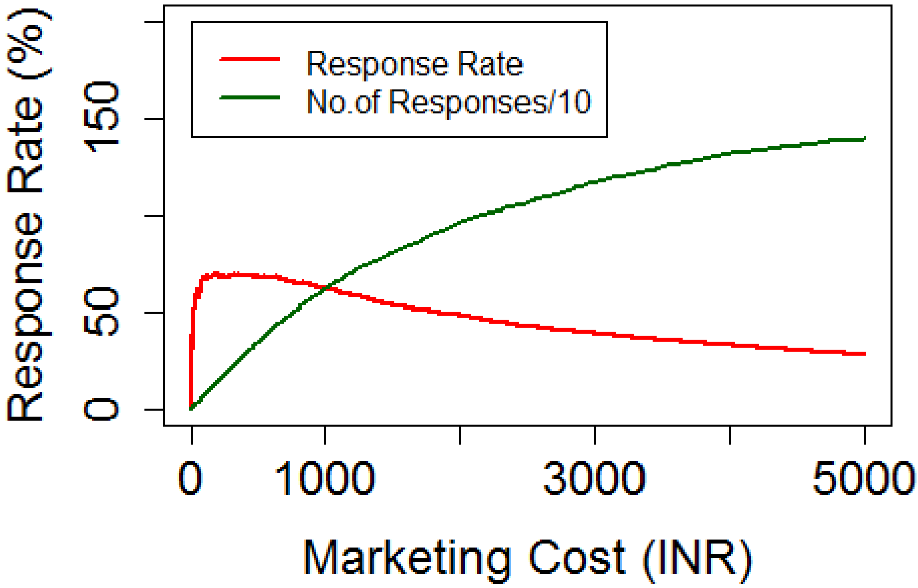
```
6800/1479

## [1] 4.597701
```

We can also visualise how the response rate varies with the marketing cost.

```
seq_prospects <- seq(1, 5000, by = 1)
cost_matrix <- matrix(0, length(seq_prospects), 3)
for (i in seq_prospects)
{
  cost_matrix[i, 1] = i
  response <- length(which(test_predictions$response[1:i] == "yes"))
  cost_matrix[i, 2] = response/i
  cost_matrix[i, 3] = response
}
colnames(cost_matrix) <- c("number of prospects targeted (marketing cost)",
"response rate", "number of responses")
head(cost_matrix)

##      number of prospects targeted (marketing cost) response rate
## [1,]                                             1     0.0000000
## [2,]                                             2     0.0000000
## [3,]                                             3     0.0000000
## [4,]                                             4     0.0000000
## [5,]                                             5     0.2000000
## [6,]                                             6     0.1666667
##      number of responses
## [1,]                   0
## [2,]                   0
## [3,]                   0
## [4,]                   0
## [5,]                   1
## [6,]                   1
```

The cost_matrix stores the number of prospects targeted, the response rates and the number of responses. The marketing cost is same as number of people targted since we've assumed Re 1 per call.

```
plot(cost_matrix[, 1], cost_matrix[,2]*100,xlab="Marketing Cost
(INR)",ylab="Response Rate (%)",cex.lab=1.5,cex.axis=1.5,
ylim=c(0,200),type="l",lwd=2,axes=TRUE,col=2)

lines(seq_prospects, cost_matrix[, 3]/10, col="darkgreen",lwd=2)
box()
legend(0, 200,col=c(2,"darkgreen"),lwd=c(2,2),c("Response Rate","No.of
Responses/10"))
```




The plot shows how the number of responses and the response rate varies with marketing cost (no. of prospects targeted).

You can see that for INR 3000, almost 1272 prospects are expected to respond. Earlier, about 1587 would respond at a cost of Rs 13500.

```
cost_matrix[3000:3010, ]

##       number of prospects targeted (marketing cost) response rate
##  [1,]                                          3000     0.3890000
##  [2,]                                          3001     0.3888704
##  [3,]                                          3002     0.3887408
##  [4,]                                          3003     0.3889444
##  [5,]                                          3004     0.3888149
##  [6,]                                          3005     0.3886855
##  [7,]                                          3006     0.3885562
##  [8,]                                          3007     0.3884270
##  [9,]                                          3008     0.3886303
## [10,]                                          3009     0.3885012
## [11,]                                          3010     0.3883721
##       number of responses
##  [1,]                1167
##  [2,]                1167
##  [3,]                1167
##  [4,]                1168
##  [5,]                1168
##  [6,]                1168
```

```
## [7,]                 1168
## [8,]                 1168
## [9,]                 1169
## [10,]                1169
## [11,]                1169
```

```r
1167/1587
```

```
## [1] 0.7353497
```

```r
3000/13500
```

```
## [1] 0.2222222
```

Thus, we can acquire **about 73% of the customers for only about 22% of the marketing cost.**