Analysis and Presentation for Bank Marketing Data

By:

Vinay Kumar

MS by Research Scholar

IIT Kharagpur

Vinay2k2@gmail.com

+91-8348575432

Table of Contents

1. Introduction 2

2. Problem Statement and Data Set Description 2

3. Techniques Used 5

3.1. Data Science: A systematic approach for data analysis 5

3.2. Data Cleansing 7

3.3. Data engineering and exploratory analysis 9

3.4. Data transformation and Binning 12

3.5. More Plots 12

3.6. Training and Testing 14

3.7. Decision tree model 15

3.8. Random forest model 17

4. Conclusion and Managerial Implication 18

5. Future plan 19

1. Introduction

For the campaigning need organizations rely mainly on either mass campaign, which targets a larger population or direct campaign, which targets specific clients. Formal studies have shown mass campaign to be as less effective as 1% positive response. In contrast, the direct campaign focuses on specific potential clients and is often data driven, and more effective.

In the age of Big-Data, it is nearly impossible to scale without data-driven techniques and solutions. The bank in the question is considering how to optimize this campaign in future. We can make data-driven decision to suggest marketing manager about effective client selection, which would increase the conversion rate. Direct marketing is effective yet it has some drawbacks, such as causing negative attitude towards banks due to the intrusion of privacy. It would be interesting to investigate how we can decrease the outbound call rate and use inbound calls for cross-selling intelligently to increase the duration of the call. We will discuss later why the duration of a call is an important parameter here.

We will be building few classifiers to predict whether a particular client will subscribe to term deposit or not. If classifier has very high accuracy it can help the manager to filter clients and use available resources more efficiently to achieve the campaign goal. Besides, we would also try to find influential factors for decision so that we can establish efficient and precise campaigning strategy. Proper strategy would reduce cost and improve long term relations with the clients.

2. Problem Statement and Data Set Description

Bank Marketing Analysis

Text file containing some real data that relates to a marketing campaign run by a bank. (available at : https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) The aim of the marketing campaign was to get customers to subscribe to a bank term deposit product. Whether they did this or not is variable ‘y’ in the data set.

The bank in question is considering how to optimize this campaign in future. What would your recommendations to the marketing manager be?

The variables are as follows:

Input variables:

# bank client data:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.', 'unknown', 'unemployed', 'management', 'housemaid', 'entrepreneur', 'student', 'blue-collar','self-employed','retired','technician','services')

3 - marital : marital status (categorical: 'married', 'divorced', 'single'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'unknown', 'secondary', 'primary', 'tertiary')

5 - default: has credit in default? (binary: 'yes', 'no')

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: 'yes', 'no')

8 - loan: has personal loan? (binary: 'yes', 'no')

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: 'unknown', 'telephone', 'cellular')

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: 'unknown', 'other', 'failure', 'success')

Output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Initial impression about data

Let's discuss about initial impression that can be created using the dataset

- Total 45211 records
- 7 numeric attributes : age, balance, day, duration, campaign, pdays, previous
- 10 Factors:
    - 6 multi-valued categorical attributes : job, marital, education, contact, month, poutcome
    - 3 yes/no binary attributes: default, housing, loan
    - 1 target attribute y
- No missing values: Preprocessing should be easier

Initial findings

We used R language for this analysis with RStudio IDE.

```
> bank_marketing_data <- read.table("Bank_Marketing.txt", header=TRUE, sep="\t")
> summary(bank_marketing_data)
```

- Most of the clients have never been contacted since contact is unknown for 28.79%.
- 81.74% of the time outcome of previous marketing campaign is unknown.
- Duration seems to have lot more variation, it may be a good predictor
- Data is very imbalanced, only 11.69% yes in outcome.

Data Science applicability: Model selection

Since response variable is categorical Logistic regression can be applied. Data is very structured and for major columns data is categorical. Decision tree or Random forest can be more appropriate to exploit sub-feature space i.e. categories. Neural network and Support vector machine can also be used for classification. Here we will limit our discussion to Decision tree and Random forest. We wish to apply NN and SVM in future.

3. Techniques Used

3.1. Data Science: A systematic approach for data analysis

We will discuss bit of data science project phases and how to classify data into categories before we start applying data science techniques. Towards end we would reiterate upon importance of this discussion. As shown in Fig. 1. a data science project goes through typically four phases.

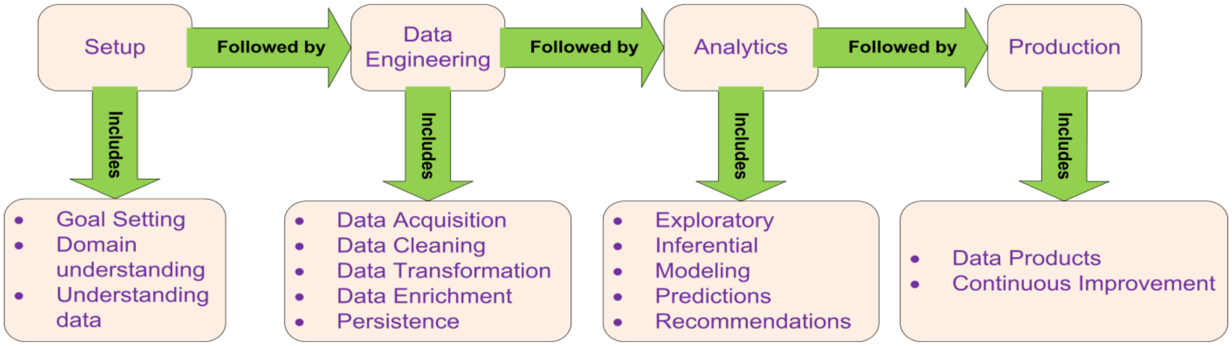


Fig.1. Data Science Project activities/phases

Dataset is a collection of observations where each observation is typically called a record. Each record has a set of attributes that point to characteristics, behavior or outcomes. We have a structured dataset here and we need to work on dataset to learn about entities and predict their future behavior/ outcomes i.e. whether they will subscribe to term deposit or not. Before we start analyzing the data we need to understand/observe typically eight pieces of information listed in Table 1. The challenge is how we should look at data to draw better managerial implication. We suggest following analogy listed in Table 1. , which will enables us to draw better managerial implications. We discuss more about this in the conclusion section.

Table 1: Concept and their relevance to our dataset

Concept	Relevance to our dataset
Entity •A thing that exists about which we research and predict in data science. •Entity has a business context. •Customer of a business	Clients
Characteristics •Every entity has a set of characteristics. These are unique properties •Properties too have a business context •Customer: Age, gender, education	age , job, marital, education, contact

Environment •Environment points to the eco-system in which the entity exists or functions. •Environment affects an entity's behavior •Customer: Country, City, Work Place	No details in thedataset
Event •A significant business activity in which an entity participates. •Events happen in a said environment. •Customer: Site visit, Store visit, Phone Call	Phone call is a event here
Behavior •What an entity does during an event. •Entities may have different behaviors in different environments •Customer: Click-stream, Purchase, Duration	Anything which happened during last transaction or event Duration, day, month,
Profile Data •Including characteristics what other business details stored previously •Data collected over time	Data collected over time: default, balance, housing, loan, contact, campaign, pdays, previous, poutcome
Outcome •The result of an activity deemed significant by the business. • Outcome values can be • Boolean (Yes/No, Pass/Fail) • Continuous (a numeric value) • Class (identification of type) •Customer: Sale (Boolean), sale value (continuous)	Y (subscribe or not)
Observation •A measurement of an event deemed significant by the business. •Captures information about •Entities involved •Characteristics of the entities •Behavior •Environment in which the behavior happens •Outcomes •An observation is also called a system of record • Customer: A phone call record, a buying transaction, an email offer	Records (Rows in the dataset)

We start by loading data into the environment. Let's look at the structure of the dataset. We would like to know what type of different attributes is present in different columns.

```
#Loading data into the environment
>bank_marketing_data <- read.table("Bank_Marketing.txt",
header=TRUE,sep="\t")
#Lets look at dataset and generate initial understanding about the column types
>str(bank_marketing_data)
```

```
'data.frame':    45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2  3 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 3 3 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 ...
 $ 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 ...
 $ 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 ...
 $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

For the above line of code let's look at the output which is present in blue colored box. How this output can help? By looking at the structure we can decide if a variable is categorical or numerical. If a variable is numerical we can use box-plot to visualize individual column and see if there is overlap. We will do box-plots after partitioning the dataset.

3.2. Data Cleansing

Missing Values or NA Values Check:

Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. TRUE value for following signifies dataset contain no NA values.

```
# If newdata has same number of observation that implies no NA value present
>newdata <- na.omit(bank_marketing_data)
>nrow(newdata)==nrow(bank_marketing_data)
```

```
[1] TRUE
```

Outlier detection and treatment

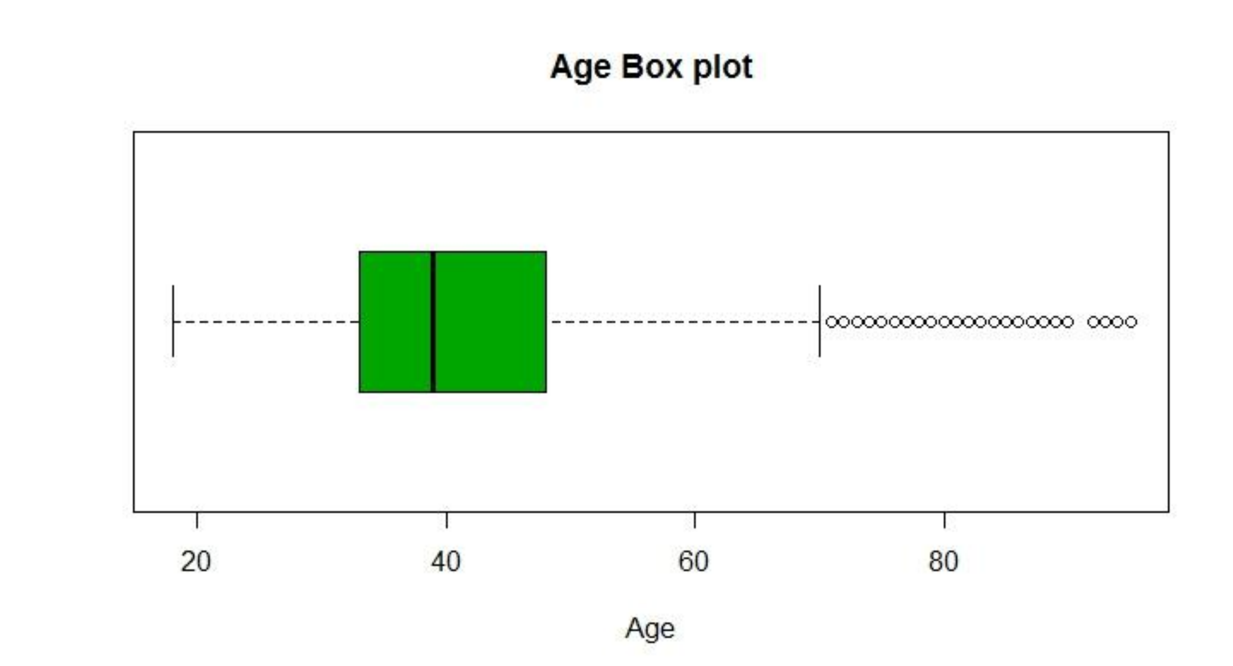
We will check in this section if outliers are present in the dataset. We may remove the records having outlier for better accuracy. First we see the summary and then we can create box-plot to see if there are outliers. Box plot might be confusing some time like one we have on page 9. We can confirm by looking at the histogram. Histogram on page 9 confirms that there is no outlier in age column.

```
################Outlier detection and treatment################
# Lets find the range of individual variables
>summary(bank_marketing_data)
```

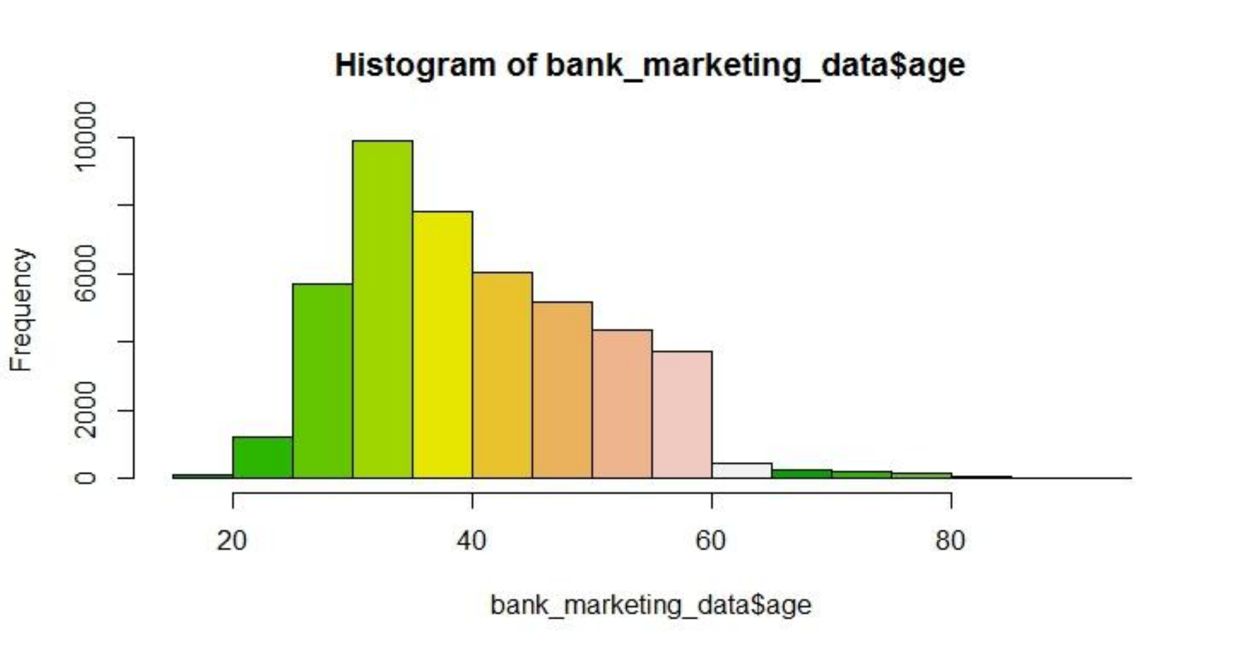
```
age               job                 marital           education          default
Min.   :18.00    blue-collar:9732    divorced: 5207    primary  : 6851    no :44396
1st Qu.:33.00    management :9458    married :27214    secondary:23202    yes:  815
Median :39.00    technician :7597    single  :12790    tertiary :13301
Mean   :40.94    admin.     :5171                      unknown  : 1857
3rd Qu.:48.00    services   :4154
Max.   :95.00    retired    :2264
                 (Other)    :6835

balance              housing      loan         contact            day
Min.   : -8019      no :20081    no :37967    cellular :29285    Min.   : 1.00
1st Qu.:    72      yes:25130    yes: 7244    telephone: 2906    1st Qu.: 8.00
Median :   448                                unknown  :13020    Median :16.00
Mean   :  1362                                                   Mean   :15.81
3rd Qu.:  1428                                                   3rd Qu.:21.00
Max.   :102127                                                   Max.   :31.00

month            duration
may     :13766   Min.   :   0.0
jul     : 6895   1st Qu.: 103.0
aug     : 6247   Median : 180.0
jun     : 5341   Mean   : 258.2
nov     : 3970   3rd Qu.: 319.0
apr     : 2932   Max.   :4918.0
(Other) : 6060

 campaign          pdays            previous              poutcome          y
 Min.   : 1.000    Min.   : -1.0    Min.   :  0.0000     failure: 4901     no :39922
 1st Qu.: 1.000    1st Qu.: -1.0    1st Qu.:  0.0000     other  : 1840     yes: 5289
 Median : 2.000    Median : -1.0    Median :  0.0000     success: 1511
 Mean   : 2.764    Mean   : 40.2    Mean   :  0.5803     unknown:36959
 3rd Qu.: 3.000    3rd Qu.: -1.0    3rd Qu.:  0.0000
 Max.   :63.000    Max.   :871.0    Max.   :275.0000
```

3.3. Data engineering and exploratory analysis

```
# We look at difference between mean and median in summary if it's more there
might be outliers
> boxplot(bank_marketing_data$age, main="Age Box plot",yaxt="n", xlab="Age",
horizontal=TRUE, col=terrain.colors(2))
```



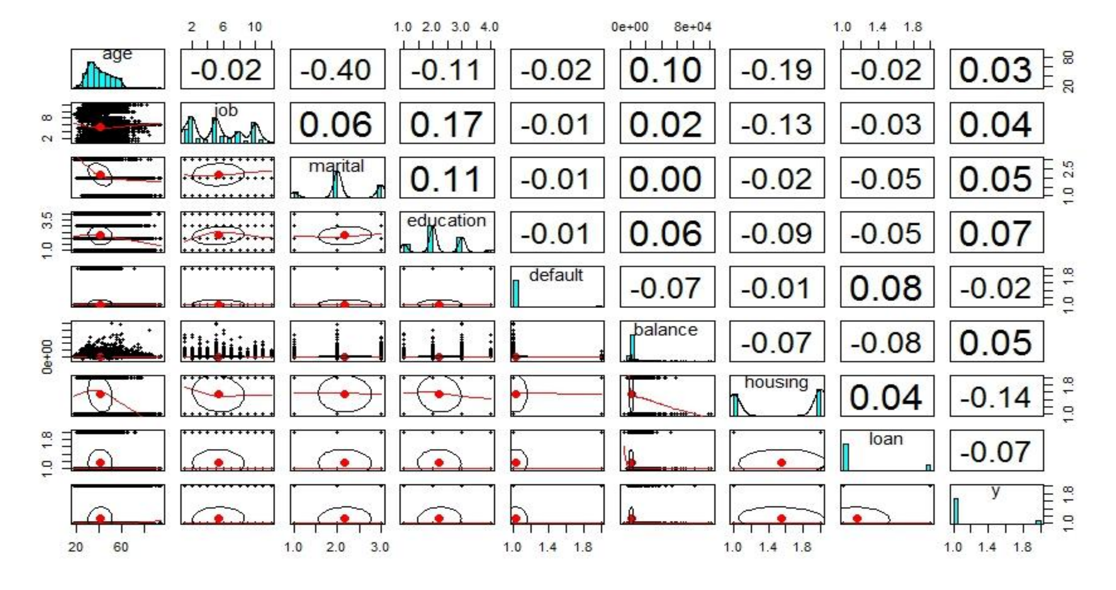
```
# By plotting histogram we can ensure if there are outliers or not
>hist(bank_marketing_data$age,col=terrain.colors(10))
```

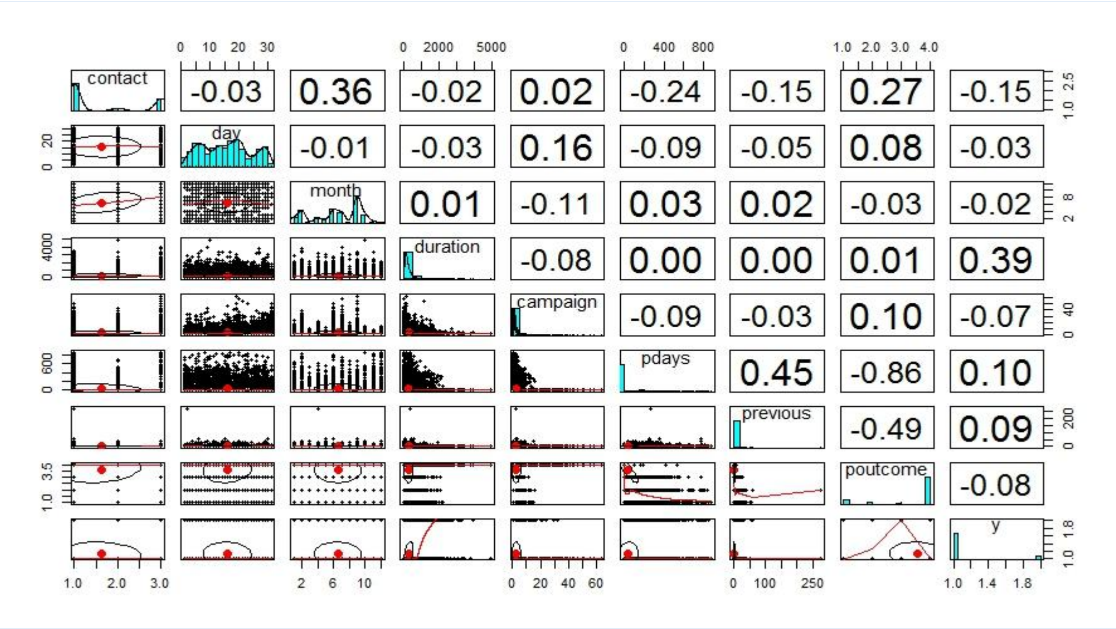


Correlation Analysis

What we saw in the box plot can be emphasized by correlation plot, It can tell if predictor is a good predictor or not a good predictor. This analysis can help us decide if we can drop some columns/predictors depending upon its correlation with the outcome variable.

```
################Correlation Analysis################
>library(psych)
>pairs.panels(bank_marketing_data[, c(1:8,17)])
>pairs.panels(bank_marketing_data[, c(9:17)])
```



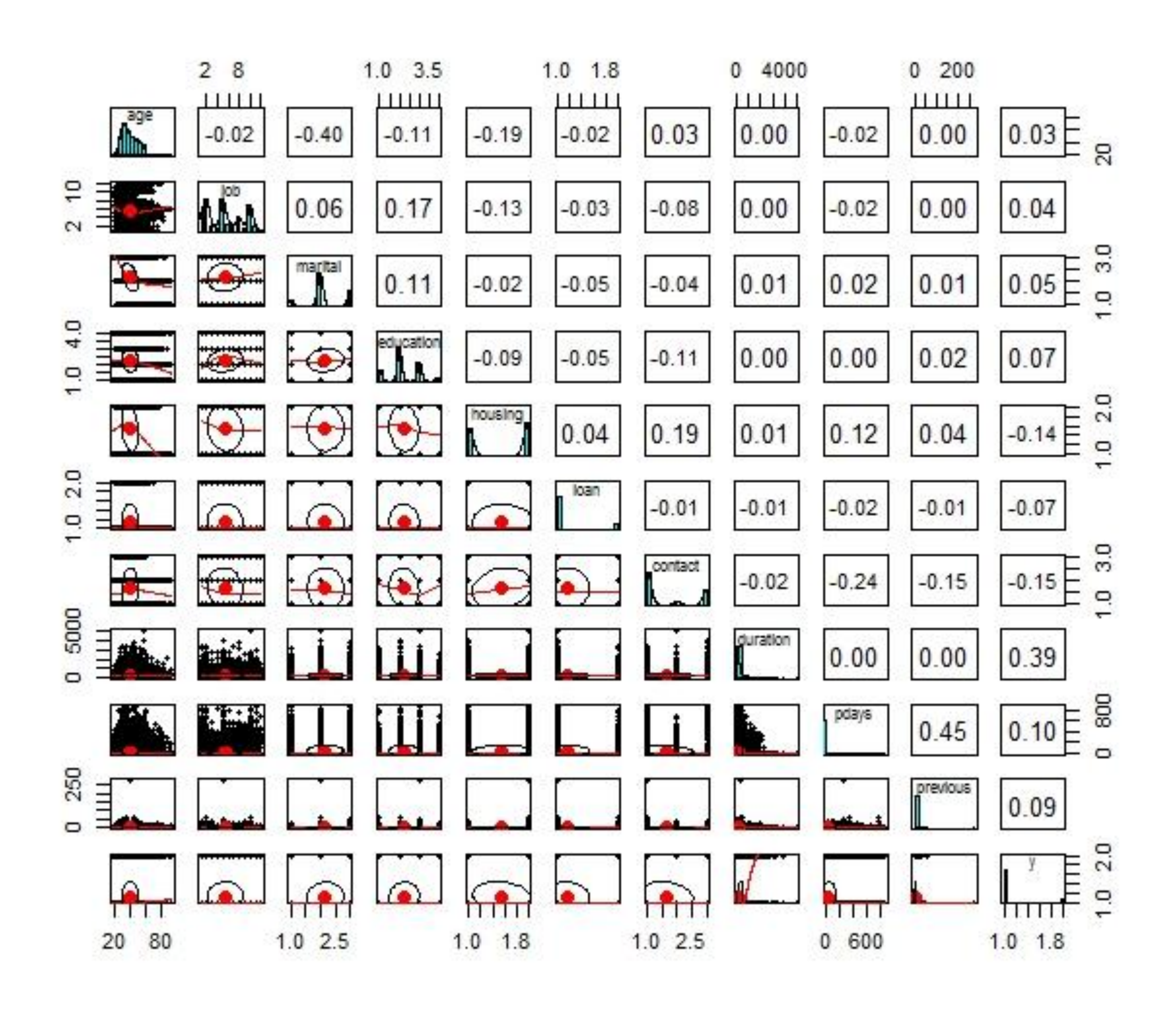


Subset Selection/ Feature-space reduction:

Features-space can be reduced by selecting subsets based upon correlation values obtained in page 10.

```
#################Subset Selection#################
>bank_marketing_data_sub <-bank_marketing_data[, c(1:4,7:9,12,14,15,17)]
>str(bank_marketing_data_sub)
>pairs.panels(bank_marketing_data_sub)
```

```
'data.frame':     45211 obs. of  11 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 3 2 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2  ...
 $ 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...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```



3.4. Data transformation and Binning

We do data transformation and binning for better modeling. We convert categorical variable into numerical using binning.

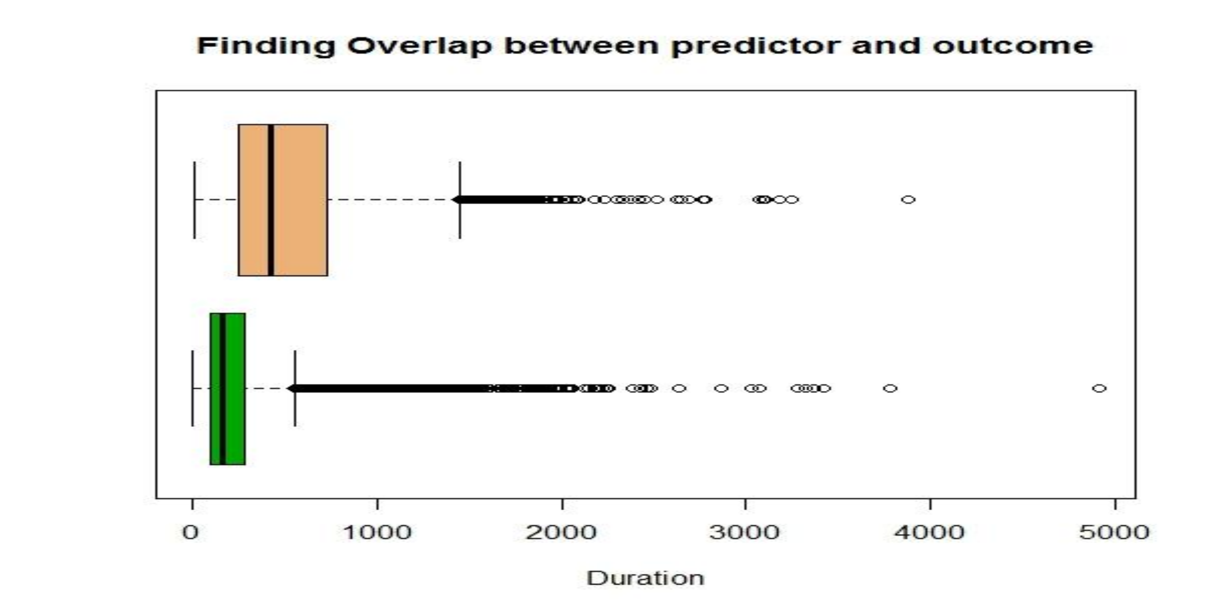
```
################Binning and Data Transformation################
>bank_marketing_data_sub$age <- cut(bank_marketing_data_sub$age,
c(1,20,40,60,100))
>bank_marketing_data_sub$is_divorced <- ifelse(
bank_marketing_data_sub$marital == "divorced", 1, 0)
>bank_marketing_data_sub$is_single <- ifelse( bank_marketing_data_sub$marital
== "single", 1, 0)
>bank_marketing_data_sub$is_married <- ifelse(
bank_marketing_data_sub$marital == "married", 1, 0)
>bank_marketing_data_sub$marital <- NULL
>str(bank_marketing_data_sub)
```

```
'data.frame':     45211 obs. of  13 variables:
 $ age         : Factor w/ 4 levels "(1,20]","(20,40]",..: 3 3 2 3  ...
 $ job         : Factor w/ 12 levels "admin.","blue-collar",..: 5 10  ...
 $ education   : Factor w/ 4 levels "primary","secondary",..: 3 2 2  ...
 $ 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 ...
 $ duration    : int  261 151 76 92 198 139 217 380 50 55 ...
 $ pdays       : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ y           : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1  ...
 $ is_divorced: num  0 0 0 0 0 0 0 1 0 0 ...
 $ is_single  : num  0 1 0 0 1 0 1 0 0 1 ...
 $ is_married : num  1 0 1 1 0 1 0 0 1 0 ...
```

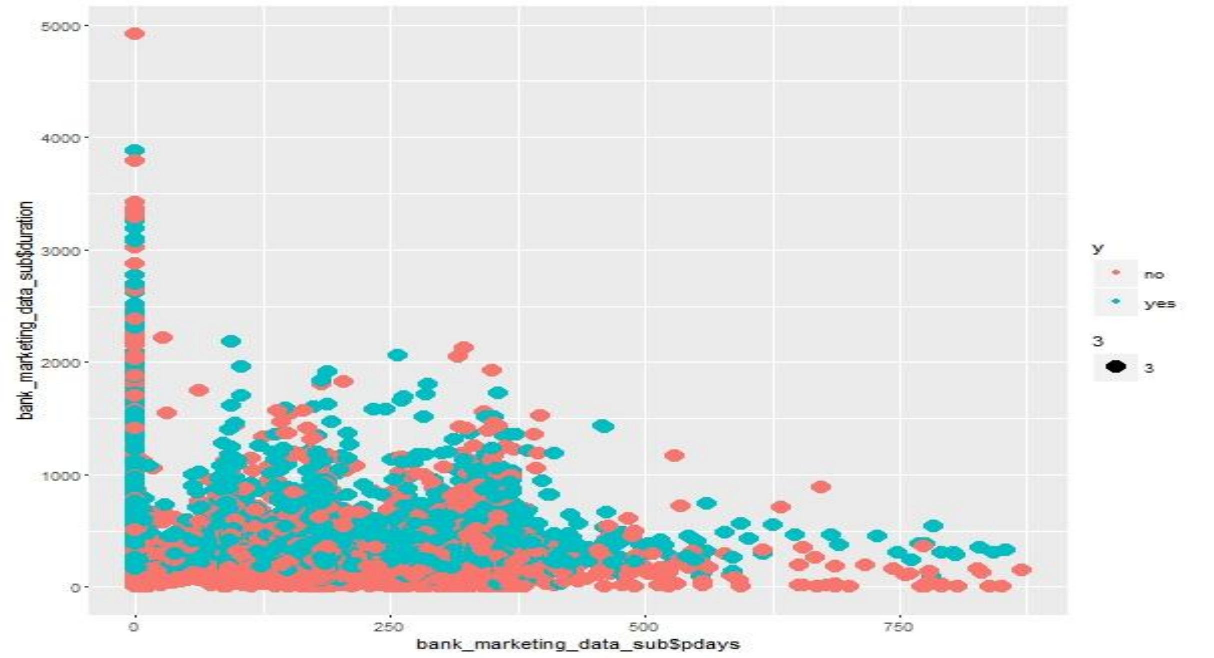
3.5. More Plots

We can do more plotting to generate more insight. We can try to observe overlap between individual numerical variable and the outcome, If there is no overlap and we have clear boundary then that particular predictor can play a major role. For example the box-plot in page 13 shows that there is no clear boundary between yes (green) and no (orange). If we had no overlap then just by looking at duration we could say if it is yes or no. Unfortunately for the entire numerical variable in the supplied dataset we have overlaps and hence none of them can become major predictor. Similar plot can be drawn between two predictors using qplot from ggplot2 . This emphasizes the correlation between variables we found earlier and can help investigate individual relationship visually. General visualization section in the supplied code (Bank_Maketing.R) shows more ways to visualize data.

```
#Finding overlap between predictor and outcome/target variable
>boxplot(duration~y,data=bank_marketing_data_sub, main="Finding Overlap
between predictor and outcome",
        yaxt="n", xlab="Duration", horizontal=TRUE,
        col=terrain.colors(3))
```



```
##Similarly boundry can be researched between two predictors also
>library(ggplot2)
>qplot(bank_marketing_data_sub$pdays,bank_marketing_data_sub$duration,data=ba
nk_marketing_data_sub,colour=y,size=3)
```



3.6. Training and Testing

We have imbalanced data as discussed already in the earlier section. How do we split imbalanced data consistently? We can use CreateDataPartition method present in caret package to split in such a way that training and testing data will have same ratio of target variable.

```
#################Training and testing split#################
#CreateDataPartition present in caret packagesplit in such a way that
#training and testing data will have same ratio for target variable
>library(caret)
#Rows selection for training data set
>inTrain <- createDataPartition(y=bank_marketing_data_sub$y
,p=0.7,list=FALSE)
>training <- bank_marketing_data_sub[inTrain,]
>testing <- bank_marketing_data_sub[-inTrain,]
# As we said we have imbalanced data so how can we do  sampling
#Caret will take care of that, So createDataPartition does the magic
>dim(training);dim(testing),p=0.7,list=FALSE)
```

```
[1] 31649    13
[1] 13562    13
```

```
# We can see imbalancing has been taken care of or not
>table(training$y); table(testing$y)
```

```
no      yes
27946   3703

no     yes
11976   1586
```

So as we can see in above output imbalance has been taken care of by CreateDataPartition method.

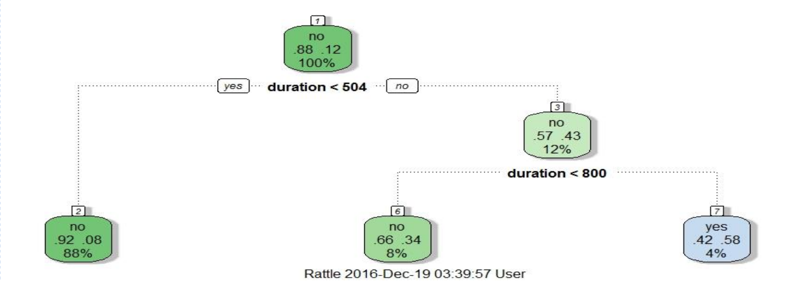
3.7. Decision tree model

```
> #################Decision Tree#################
> library(rpart)
> library(rpart.plot)
> library(rattle)
> library(caret)
> dt_model<- rpart(y ~ ., data = training)
> summary(dt_model)
```

```
n= 31649

          CP nsplit rel error    xerror       xstd
1 0.02835539      0 1.0000000 1.0000000 0.01544198
2 0.01000000      2 0.9432892 0.9489603 0.01509352
Variable importance
duration
     100
Node number 1: 31649 observations,    complexity param=0.02835539
  predicted class=no   expected loss=0.1170021  P(node) =1
    class counts: 27946  3703
   probabilities: 0.883 0.117
  left son=2 (27892 obs) right son=3 (3757 obs)
  Primary splits:
      duration < 503.5 to the left,  improve=820.4202, (0 missing)
      pdays    < 8.5   to the left,  improve=175.6106, (0 missing)
      previous < 0.5   to the left,  improve=173.1874, (0 missing)
      age      splits as  RLLR,      improve=167.5416, (0 missing)
      contact  splits as  RRL,       improve=142.7370, (0 missing)

Node number 2: 27892 observations
  predicted class=no   expected loss=0.0752187  P(node) =0.8812917
    class counts: 25794  2098
   probabilities: 0.925 0.075

Node number 3: 3757 observations,    complexity param=0.02835539
  predicted class=no   expected loss=0.4272026  P(node) =0.1187083
    class counts:  2152  1605
   probabilities: 0.573 0.427
  left son=6 (2375 obs) right son=7 (1382 obs)
  Primary splits:
      duration   < 800.5 to the left,  improve=96.77683, (0 missing)
      contact    splits as  RRL,       improve=37.62359, (0 missing)
      is_married < 0.5   to the right, improve=22.39650, (0 missing)
      housing    splits as  RL,        improve=18.55489, (0 missing)
      is_single  < 0.5   to the left,  improve=15.30389, (0 missing)

Node number 6: 2375 observations
  predicted class=no   expected loss=0.3406316  P(node) =0.07504187
   class counts:  1566   809
   probabilities: 0.659 0.341

Node number 7: 1382 observations
  predicted class=yes  expected loss=0.4240232  P(node) =0.04366647
    class counts:   586   796
    probabilities: 0.424 0.576
```

```
> #################Testing Decision Tree################
> predictions <- predict(dt_model, testing, type = "class")
> #What is predicted
> table(predictions)
> # Lets look at the confusion matrix
> confusion.matrix <- prop.table(table(predictions, testing$y))
> confusion.matrix
> confusionMatrix(predictions,testing$y)
> fancyRpartPlot(dt model)
```

```
predictions
   no   yes
13015   547

predictions          no         yes
        no  0.86535909 0.09430762
        yes 0.01769650 0.02263678
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  11736  1279
       yes   240    30
       Accuracy : 0.888
```



3.8. Random forest model

```
> library(randomForest)
> model <- randomForest(y ~ ., data=training)
> model
```

```
Call:
 randomForest(formula = y ~ ., data = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 10.34%
Confusion matrix:
       no  yes class.error
no  27105  841  0.03009375
yes  2433 1270  0.65703484
```

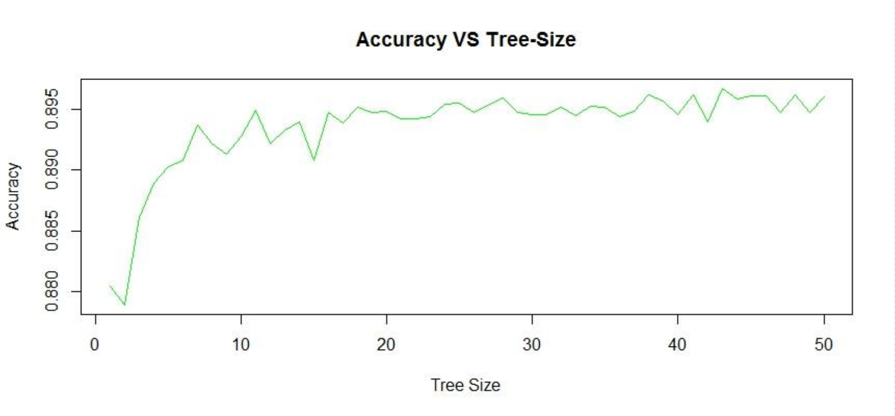
```
> #importance of each predictor
> importance(model)
```

```
MeanDecreaseGini
age                  191.48265
job                  372.70030
education            144.67461
housing              189.23100
loan                  66.17372
contact              133.93896
duration            1792.53250
pdays                509.05163
previous             224.80533
is_divorced           38.92399
is_single             48.52818
is_married            52.14683
```

```
> ############ Testing Random forest ############
> library(caret)
> predicted <- predict(model, testing)
> table(predicted)
> confusionMatrix(predicted, testing$y)
```

```
   predicted
   no    yes
12696   866
Confusion Matrix and Statistics
          Reference
Prediction    no   yes
       no  11815   806
       yes   161   780
Accuracy : 0.9287
```

```
> #Effect of increasing tree count
> accuracy=c()
> for (i in seq(1,50, by=1)) {
+ modFit <- randomForest(y ~ ., data=training, ntree=i)
+ accuracy <- c(accuracy, confusionMatrix(predict(modFit, testing,
type="class"), testing$y)$overall[1])
```



4. Conclusion and Managerial Implication

We suggested in the Section 3.1 that looking at the data in certain manner will enables us to draw better managerial implications. We can observe that we have eight pieces of information in Table 1. And if look at the important predictors in their order of importance, as per Random forest, they are: duration, pdays, job, previous, age, housing, education, contact, loan, marital. We know characteristics of an entity are not in control of bank or any company manager as they are individual characteristics. Removing them from important predictors leaves us with duration, pdays, previous, housing, contact, and loan as option. Among these duration can be enhanced and in this case it's the most important predictor as well. Based on data analyzed and domain following are few managerial suggestions for improving duration.

- As discussed in Section 1. out-bound call creates negative attitude towards bank due to the intrusion of privacy. Bank should decrease the outbound call rate and use inbound calls for cross-selling intelligently to increase the duration of the call. Agents may pitch about profits of term deposit for a particular client during inbound calls.
- Bank should have a clear promotion and they should put the value proposition up front – waiving fees, offering something free, or promoting a bundled service at a discounted

price increases response rates versus just informing customers about a product. The value proposition should be right up front, with a clear call to action.

- “Duration” has positive effect on people saying “yes”. This is because the longer the conversations on the phone, the higher interest the customer will show to the term deposit. The Bank ought to focus on the potential clients who have significant call duration and moreover who have reacted emphatically amid the past campaign.
- They should customize or personalize at whatever point possible – the more an agent utilize the client's name and background the more effective the response will be. Agents ought to demonstrate that they know their name, their business/family unit and something about them (size of business, for illustration) and response rates will increase. During inbound call we should have a customized report prepared for individual profiles and they should be used by agents without intersecting the frightening lines. This fosters the philosophy of keeping customer at the centre.
- Agents may target clients of job category of housemaid, services, technician etc as these set of people are averse to taking risks and look for safe deposit of their savings with fixed returns.
- To improve their lead generation banks may hire more people or develop analytics solution, as an alternative, like we discussed here for client selection. This would improve the quality of conversation as agents would be spending more time with selective clients only. Less hiring also implies reduction in cost for the company.

5. Future plan

Logistic regression, NN and SVM may be appropriate for dataset hence can be applied to check for better accuracy. As outcome has only two values Logistic regression may seem appropriate choice but as there are more categorical variables Logistic regression may not result in better accuracy, still it can be tried along with NN and SVM.