Bank Term Deposit Scheme

Data Analysis and Approach to Model Building

Anushree Tomar

7-10-2021

The banks are moving with the pace of technology and incorporating different techniques to get the clients on-board. There are multiple marketing techniques in the market different banks are resorting to get people involved into different banking schemes. One such technique is phone calling the clients, getting their details and letting them know about the different schemes. It might require multiple calls to the same client to figure out if the client will be on-board or not. It is where Machine Learning can be incorporated and the result can be predicted based on the information received. This information will be valuable to pay more attention to the customers who might be willing to get on-board and be in their contact. The models can be trained on the data set and the banks can plan out a strategy which will be beneficial for them.

Dataset Source:

The data sets are provided with the details of the campaign which we used to build a model which can predict if the client will say 'yes' or 'no' for the scheme. The scheme in question is term deposit and is the same for all the clients. If the client gets on-board, it is denoted with 'yes' and if he does not, it is denoted with 'no'.

Data Description:

The data set consists of the 21 attributes along with their values. The term deposit is denoted with variable `y`. The data can be understood in the 4 parts:

1. Bank client data attributes
2. Related with the last contact of the current campaign attributes
3. Other Attributes
4. Social and Economic Context Attributes

1. Bank client data attributes

Attribute	Values
key	1, 2. 3, 4....
age	numeric
job	type of job (categorical:'admin.','blue-collar','entrepreneur','housemaid',

Attribute	Values
	‘management’,‘retired’,‘self-employed’,‘services’,‘student’,‘technician’,‘unemployed’,‘unknown’)
marital	marital status (categorical: ‘divorced’,‘married’,‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed)
education	categorical: ‘basic.4y’,‘basic.6y’,‘basic.9y’,‘high.school’,‘illiterate’, ‘professional.course’,‘university.degree’,‘unknown’
default	has credit in default? (categorical: ‘no’,‘yes’,‘unknown’)
housing	has housing loan? (categorical: ‘no’, ‘yes’, ‘unknown’)
loan	has personal loan? (categorical: ‘no’,‘yes’,‘unknown’)

2. Related with the last contact of the current campaign attributes

Attributes	Values
contact	contact communication type (categorical: ‘cellular’,‘telephone’)
month	last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’, . . . , ‘nov’, ‘dec’)
day_of_week	last contact day of the week (categorical: ‘mon’,‘tue’,‘wed’,‘thu’,‘fri’)
duration	last contact duration, in seconds (numeric)

Note: duration attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.

3. Other Attributes

Attributes	Values
campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client (numeric)

Attributes	Values
poutcome	outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

4. Social and Economic Context Attributes

Attributes	Values
emp.var.rate	employment variation rate - quarterly indicator (numeric)
cons.price.idx	consumer price index - monthly indicator (numeric)
cons.conf.idx	consumer confidence index - monthly indicator (numeric)
euribor3m	euribor 3 month rate - daily indicator (numeric)
nr.employed	number of employees - quarterly indicator (numeric)

Data Dictionary

Here's a brief version of what you'll find in the data description file.

Variable	Description
key	Unique Key
y	If the client would say yes or no for the deposit scheme

Data Insights

Import Libraries

First we import libraries required for Exploratory Data Analysis.

```
library(data.table)
library(DataExplorer)
library(ggplot2)
library(scales)
library(corrplot)
```

Import Files

Let's Analyse train and test data sets:-

Observing Data

Here we will observe the data:-

Train Dataset

```
##    key age         job marital           education default housing loan
## 1: 444  45  management married   university.degree      no     yes   no
## 2: 445  34      admin. married           basic.9y      no      no   no
## 3: 446  47 blue-collar married            unknown unknown      no   no
## 4: 447  42  technician married professional.course      no      no   no
## 5: 448  57  technician married           basic.4y unknown      no  yes
## 6: 449  57  technician married           basic.4y unknown      no   no
##      contact month day_of_week duration campaign pdays previous    poutcome
## 1: telephone   may         tue      140        1   999        0 nonexistent
## 2: telephone   may         tue      175        1   999        0 nonexistent
## 3: telephone   may         tue      136        1   999        0 nonexistent
## 4: telephone   may         tue     1623        1   999        0 nonexistent
## 5: telephone   may         tue       50        1   999        0 nonexistent
## 6: telephone   may         tue      101        1   999        0 nonexistent
##    emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   y
## 1:          1.1         93.994         -36.4     4.857        5191  no
## 2:          1.1         93.994         -36.4     4.857        5191  no
## 3:          1.1         93.994         -36.4     4.857        5191  no
## 4:          1.1         93.994         -36.4     4.857        5191 yes
## 5:          1.1         93.994         -36.4     4.857        5191  no
## 6:          1.1         93.994         -36.4     4.857        5191  no
```

Test Dataset

```
##    key age       job marital   education default housing loan   contact month
## 1:   1  56 housemaid married    basic.4y      no      no   no telephone   may
## 2:   2  57  services married high.school unknown      no   no telephone   may
## 3:   3  37  services married high.school      no     yes   no telephone   may
## 4:   4  40    admin. married    basic.6y      no      no   no telephone   may
## 5:   5  56  services married high.school      no      no  yes telephone   may
## 6:   6  45  services married    basic.9y unknown      no   no telephone   may
##    day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 1:         mon      261        1   999        0 nonexistent          1.1
## 2:         mon      149        1   999        0 nonexistent          1.1
## 3:         mon      226        1   999        0 nonexistent          1.1
## 4:         mon      151        1   999        0 nonexistent          1.1
## 5:         mon      307        1   999        0 nonexistent          1.1
## 6:         mon      198        1   999        0 nonexistent          1.1
##    cons.price.idx cons.conf.idx euribor3m nr.employed
## 1:         93.994         -36.4     4.857        5191
## 2:         93.994         -36.4     4.857        5191
## 3:         93.994         -36.4     4.857        5191
## 4:         93.994         -36.4     4.857        5191
## 5:         93.994         -36.4     4.857        5191
## 6:         93.994         -36.4     4.857        5191
```

Check for missing value

Train Dataset

```
##    rows columns discrete_columns continuous_columns all_missing_columns
## 1: 4170      22               11                 11                   0
##    total_missing_values complete_rows total_observations memory_usage
## 1:                    0          4170              91740       463032
```

Test Dataset

```
##     rows columns discrete_columns continuous_columns all_missing_columns
## 1: 37018      21               10                 11                   0
##    total_missing_values complete_rows total_observations memory_usage
## 1:                    0         37018             777378      3862416
```

we checked train and test dataset for missing values but fortunately the data is cleaned.

Find Duplicates in the data

Dimension of the data

```
## [1] 4170   22
```

Find Duplicates

```
## Empty data.table (0 rows and 22 cols): key,age,job,marital,education,default...
```

There is no duplicate records in the dataset.

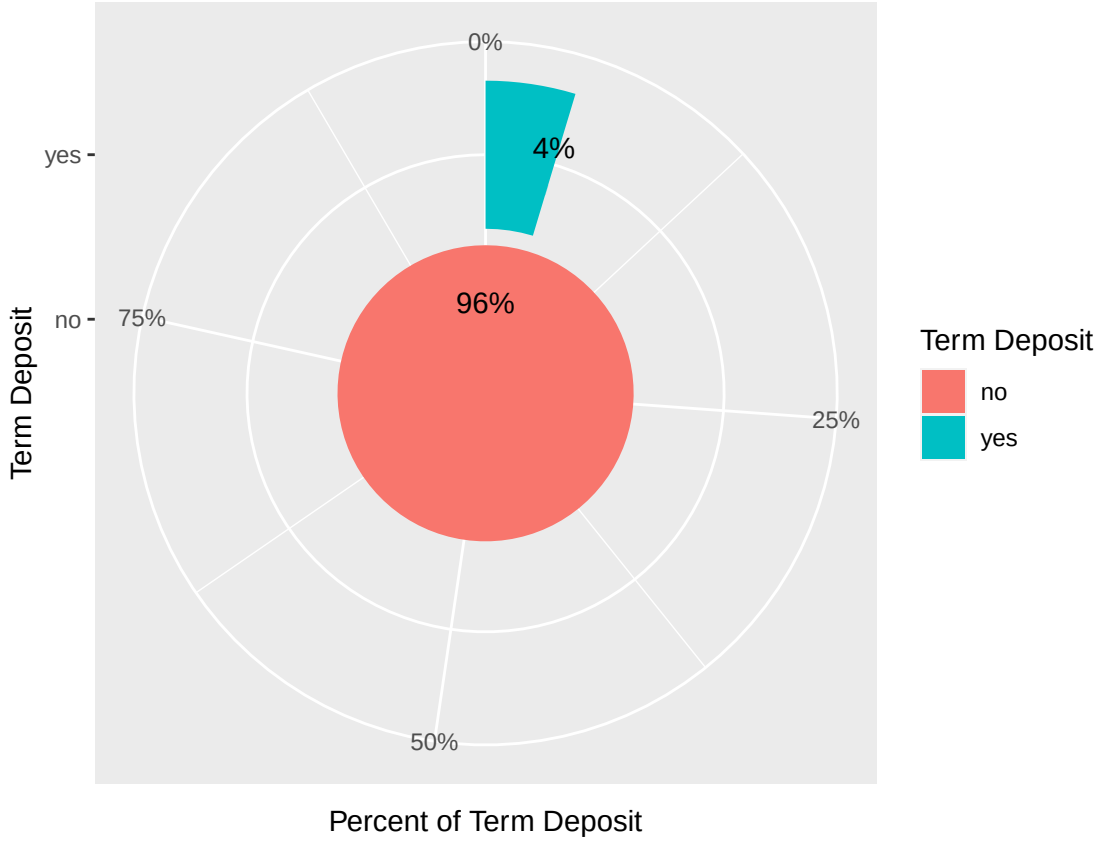
Data Visualization

Proportion of the target variable

```
## [1] "Proportion of target variable in the train data"
```

```
##
##        no       yes
## 95.563549  4.436451
```

Percent distribution of target variable



As you can see that 96% of the clients are not interested in the Term Deposit Scheme.But the Bank is interested in those clients who are interested in the scheme from the business point of view so we need to upsample the lower class in the dataset before model building.

Exploratory Data Analysis

Basic statistics

Summary Stats of train data

```
##       key             age                   job          marital
##  Min.   :  444   Min.   :20.00   blue-collar :1059   divorced: 465
##  1st Qu.: 4407   1st Qu.:33.00   admin.      : 962   married :2786
##  Median :10770   Median :39.00   technician  : 701   single  : 906
##  Mean   :11331   Mean   :40.59   services    : 411   unknown :  13
##  3rd Qu.:11812   3rd Qu.:47.00   management  : 328
##  Max.   :24452   Max.   :60.00   entrepreneur: 178
##                                  (Other)     : 531
##                education       default        housing          loan
##  basic.4y           : 489   no     :2943   no     :2067   no     :3470
##  basic.6y           : 247   unknown:1227   unknown: 134   unknown: 134
```

```
##  basic.9y           : 671                  yes    :1969   yes    : 566
##  high.school        : 930
##  professional.course: 531
##  university.degree  :1129
##  unknown            : 173
##       contact     month      day_of_week    duration         campaign
##  cellular : 885   aug: 693   fri: 947    Min.   :   4.0   Min.   : 1.000
##  telephone:3285   jun:2198   mon:1027    1st Qu.:  86.0   1st Qu.: 1.000
##                   may:1044   thu: 600    Median : 157.0   Median : 2.000
##                   nov: 235   tue: 835    Mean   : 241.8   Mean   : 3.172
##                              wed: 761    3rd Qu.: 294.0   3rd Qu.: 3.000
##                                          Max.   :3078.0   Max.   :56.000
##
##      pdays          previous              poutcome     emp.var.rate
##  Min.   :  3.0   Min.   :0.00000   failure    :  42   Min.   :-0.10
##  1st Qu.:999.0   1st Qu.:0.00000   nonexistent:4125   1st Qu.: 1.10
##  Median :999.0   Median :0.00000   success    :   3   Median : 1.40
##  Mean   :998.3   Mean   :0.01079                      Mean   : 1.24
##  3rd Qu.:999.0   3rd Qu.:0.00000                      3rd Qu.: 1.40
##  Max.   :999.0   Max.   :1.00000                      Max.   : 1.40
##
##  cons.price.idx  cons.conf.idx      euribor3m      nr.employed     y
##  Min.   :93.20   Min.   :-42.00   Min.   :4.191   Min.   :5191   no :3985
##  1st Qu.:93.99   1st Qu.:-41.80   1st Qu.:4.858   1st Qu.:5191   yes: 185
##  Median :94.47   Median :-41.80   Median :4.959   Median :5228
##  Mean   :94.11   Mean   :-39.51   Mean   :4.892   Mean   :5217
##  3rd Qu.:94.47   3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :94.47   Max.   :-36.10   Max.   :4.966   Max.   :5228
##
```

Summary Stats of test data

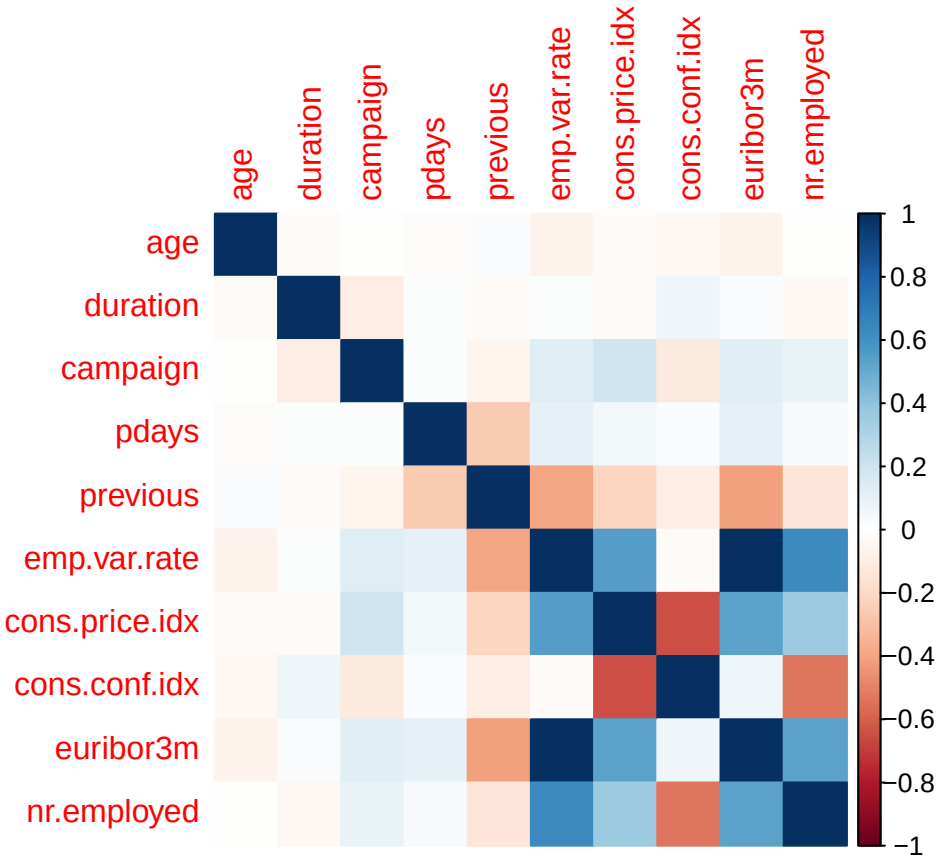
```
##       key             age                 job           marital
##  Min.   :    1   Min.   :17.00   admin.     :9460   divorced: 4147
##  1st Qu.:12497   1st Qu.:32.00   blue-collar:8195   married :22142
##  Median :22445   Median :38.00   technician :6042   single  :10662
##  Mean   :21638   Mean   :39.96   services   :3558   unknown :   67
##  3rd Qu.:31934   3rd Qu.:47.00   management :2596
##  Max.   :41188   Max.   :98.00   retired    :1609
##                                  (Other)    :5558
##                education        default         housing           loan
##  university.degree  :11039   no     :29645   no     :16555   no     :30480
##  high.school        : 8585   unknown: 7370   unknown:  856   unknown:  856
##  basic.9y           : 5374   yes    :    3   yes    :19607   yes    : 5682
##  professional.course: 4712
##  basic.4y           : 3687
##  basic.6y           : 2045
##  (Other)            : 1576
##       contact          month       day_of_week    duration
##  cellular :25259   may    :12725   fri:6880    Min.   :   0.0
##  telephone:11759   jul    : 7174   mon:7487    1st Qu.: 104.0
##                    aug    : 5485   thu:8023    Median : 182.0
##                    nov    : 3866   tue:7255    Mean   : 260.1
```

```
##                   jun    : 3120   wed:7373    3rd Qu.: 322.0
##                   apr    : 2632               Max.   :4918.0
##                   (Other): 2016
##     campaign          pdays          previous             poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   failure    : 4210
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:31438
##  Median : 2.000   Median :999.0   Median :0.0000   success    : 1370
##  Mean   : 2.499   Mean   :958.4   Mean   :0.1912
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :43.000   Max.   :999.0   Max.   :7.0000
##
##   emp.var.rate      cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.327
##  Median : 1.10000   Median :93.44   Median :-42.00   Median :4.856
##  Mean   :-0.04861   Mean   :93.52   Mean   :-40.61   Mean   :3.478
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045
##
##   nr.employed
##  Min.   :4964
##  1st Qu.:5099
##  Median :5191
##  Mean   :5161
##  3rd Qu.:5228
##  Max.   :5228
##
```

Attributes `education`,`default` and `month` have new levels in the test data.

Correlation of Numeric/Integer attributes

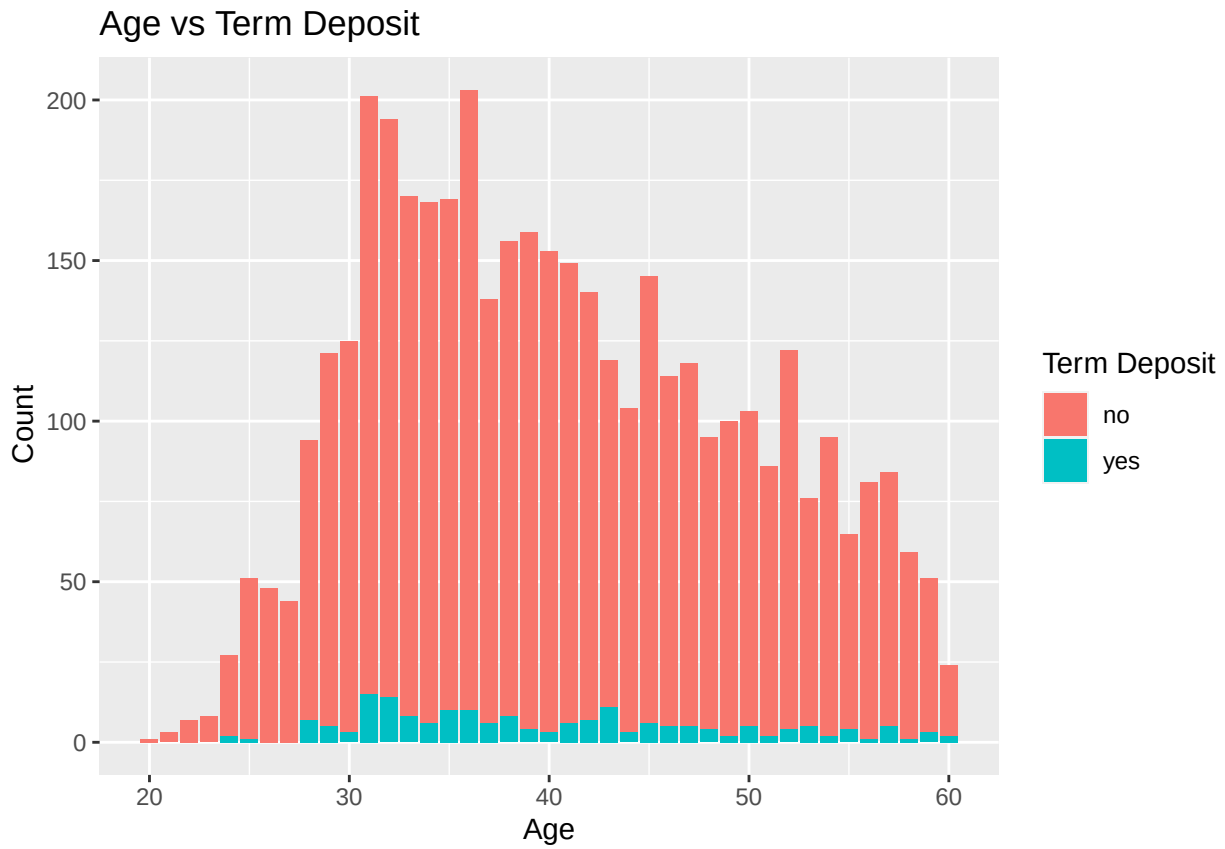
Find the correlation among continuous variable



we found no strong Positive and negative relationship between the variables in the Correlation plot except `euribor3m`(euribor 3 month rate) and `emp.var.rate`(employment variation rate) which are Positively correlated to each other.while `cons.price.idx`(consumer price index) and `nr.employed` are moderately(+) correlated to `emp.var.rate`.`cons.conf.idx` and `cons.price.idx` are moderately(-) correlated.

Distribution of the Term Deposit by age

Lets analyse which age group is accepting or rejecting the Term Deposit scheme:-



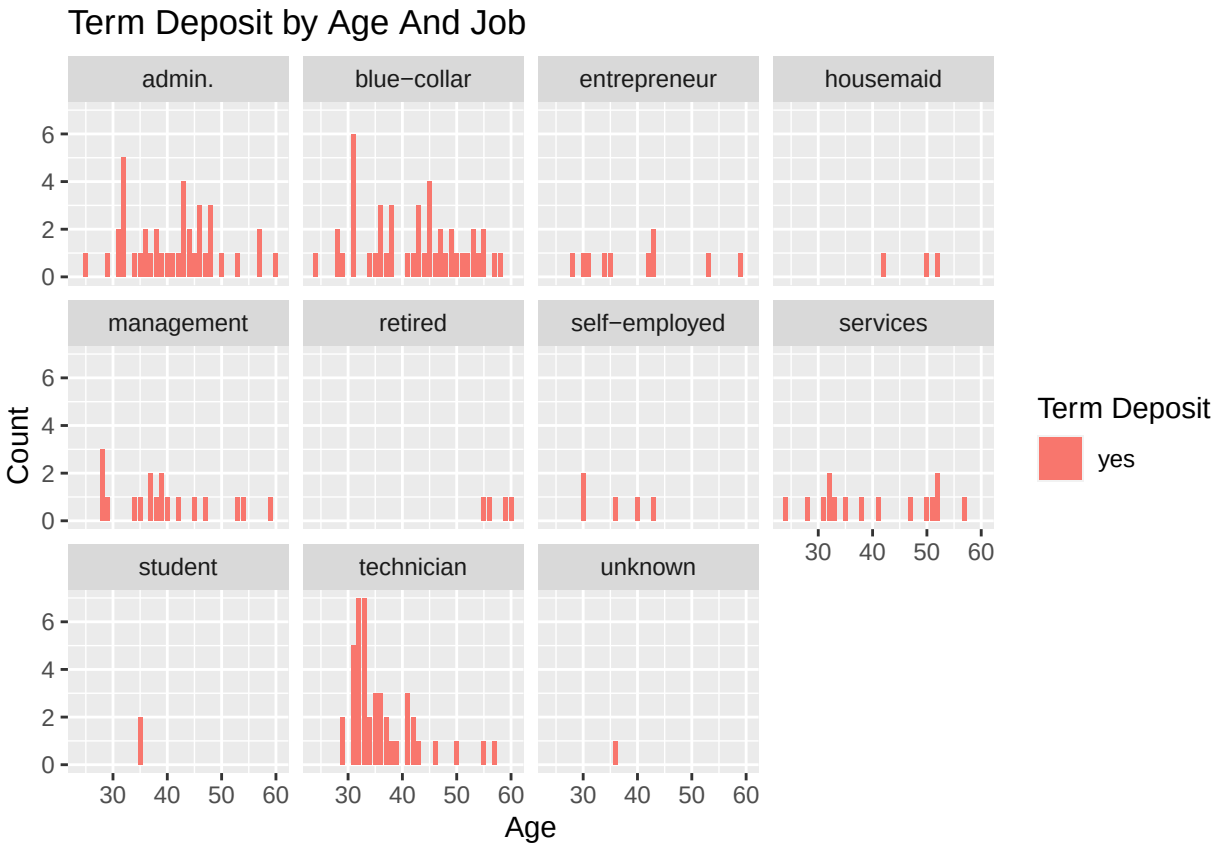
As you can see, the minimum age of the clients contacted by the bank is 20 years and the maximum is 60 years. From the above graph you can visualize that most of the clients who accepted the Term Deposit scheme is lie mostly between 30 years to 50 years.

Distribution of Age and Job by accepted Term Deposit scheme

Let's go more deeper into it where we will find the clients who accepted the term deposit scheme belong to which kind of `job`.

```
##     key age         job marital           education default housing loan
## 1:  447  42  technician married professional.course      no      no   no
## 2:  470  42  management married   university.degree      no      no   no
## 3: 1810  43      admin. married   university.degree unknown     yes   no
## 4: 1820  44 blue-collar  single            basic.6y unknown     yes   no
## 5: 2003  59  management married   university.degree      no      no   no
## 6: 2040  24    services  single         high.school      no      no   no
##      contact month day_of_week duration campaign pdays previous    poutcome
## 1: telephone   may         tue     1623        1   999        0 nonexistent
## 2: telephone   may         tue     1677        1   999        0 nonexistent
## 3: telephone   may         fri     2016        2   999        0 nonexistent
## 4: telephone   may         fri      665        2   999        0 nonexistent
## 5: telephone   may         mon      460        5   999        0 nonexistent
## 6: telephone   may         mon      757        2   999        0 nonexistent
##    emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   y
## 1:          1.1         93.994         -36.4     4.857        5191 yes
```

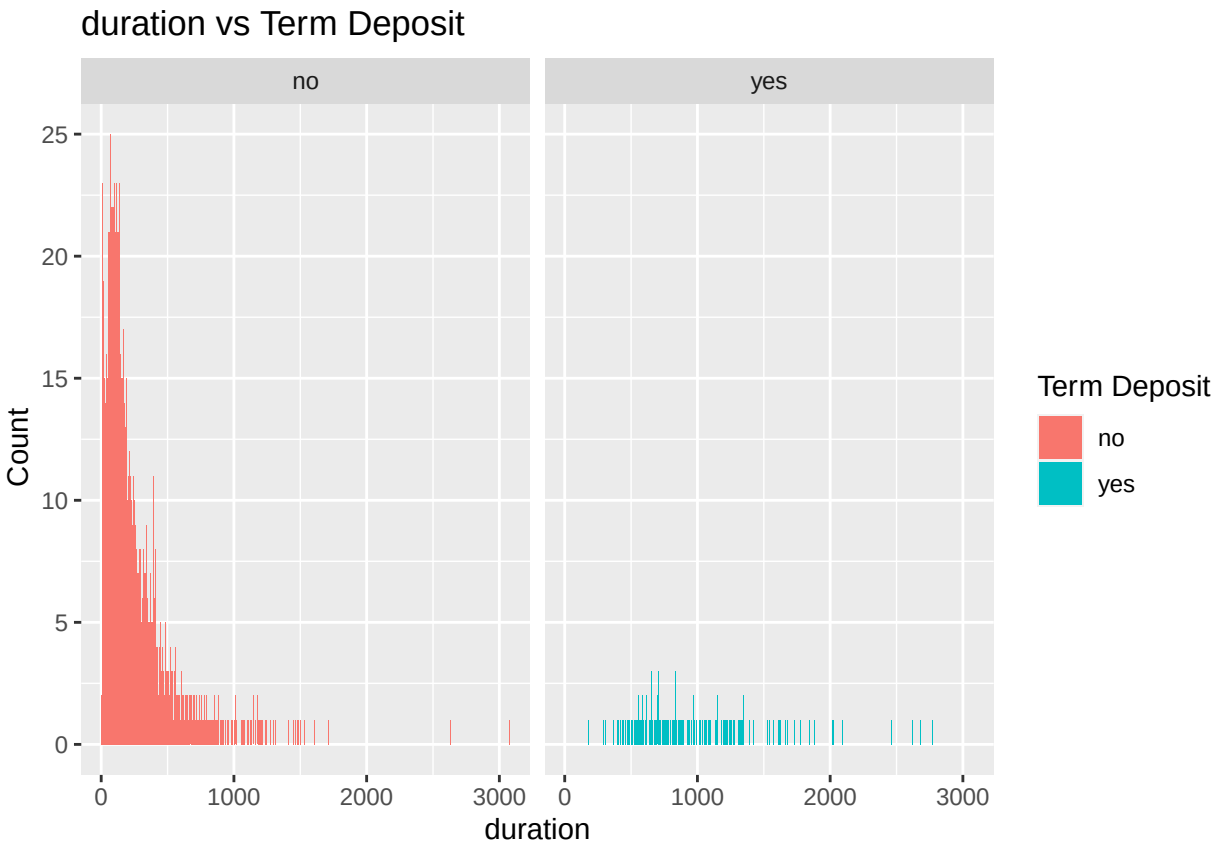
```
## 2:          1.1         93.994         -36.4     4.857        5191 yes
## 3:          1.1         93.994         -36.4     4.855        5191 yes
## 4:          1.1         93.994         -36.4     4.855        5191 yes
## 5:          1.1         93.994         -36.4     4.857        5191 yes
## 6:          1.1         93.994         -36.4     4.857        5191 yes
```



From the above graph we can assume that our future Potential clients may belong to `blue-collar`,`technician`,`admin`, and `management`(decreasing order) type of job .

Distribution of the Term Deposit by duration

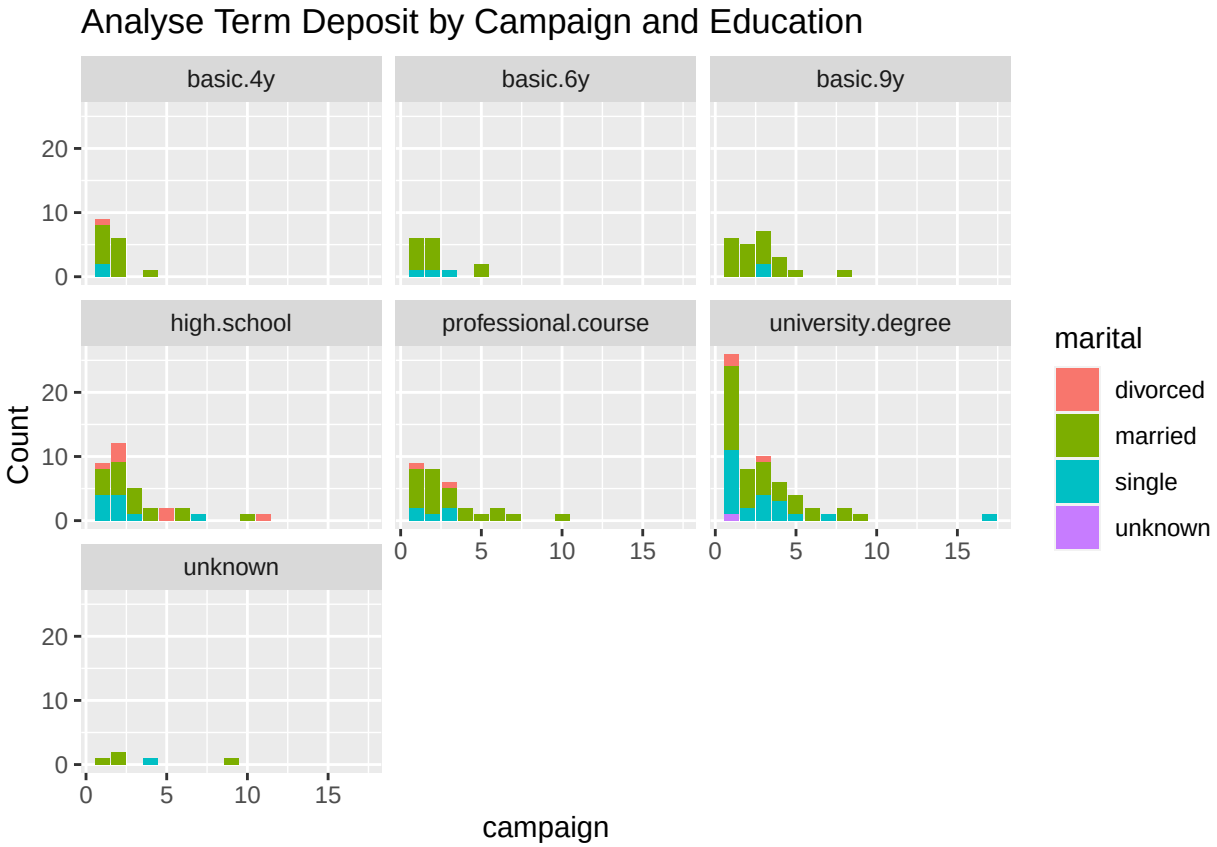
Now analyse the affect of duration on Term Deposit scheme.



From the graph it is clear that clients who contacted for a very short duration belong to `No` category. As the duration can be obtained after the phone call is performed it is obvious that duration is highly correlated to the target variable. So that at the time of Feature engineering we can drop this variable.

Distribution of campaign and education by Accepted Term Deposit

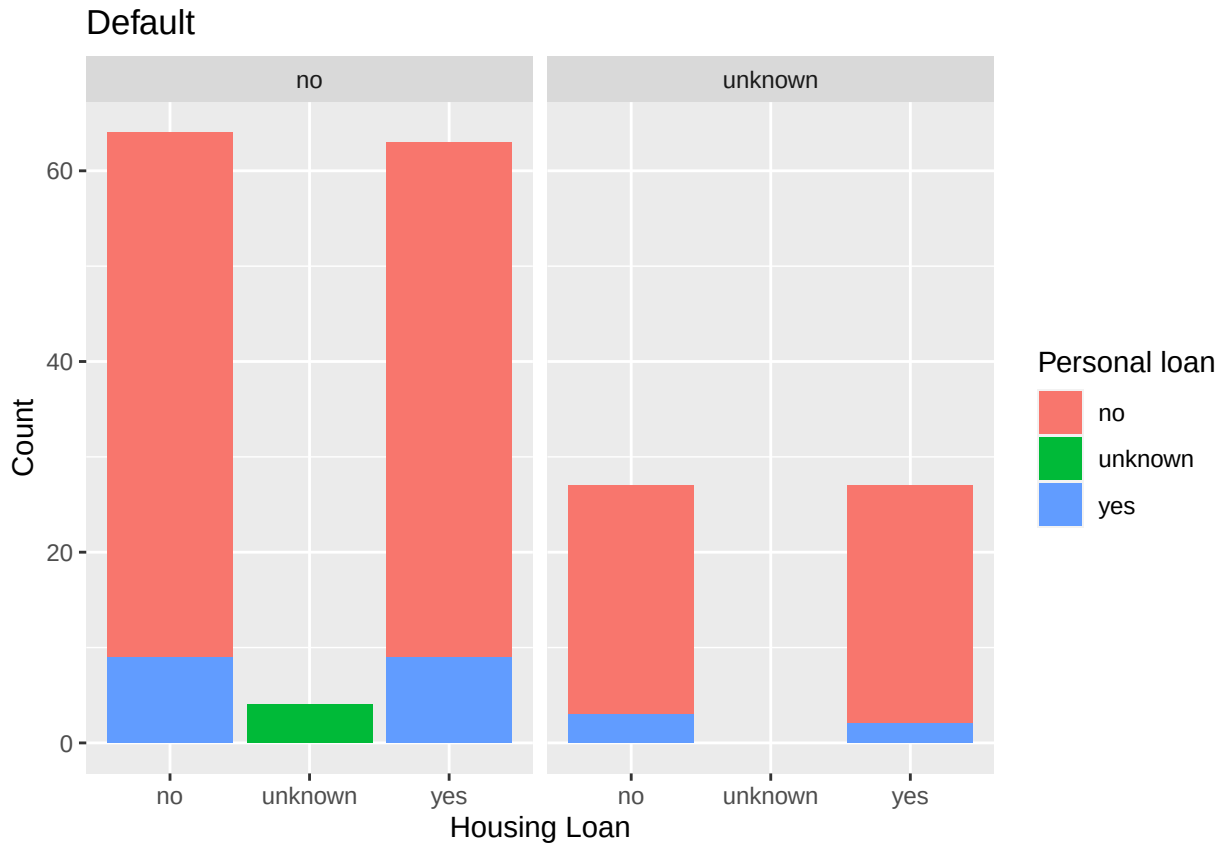
Let's analyse how many contacts performed during this campaign.



Graphical analysis saying that number of contacts performed on clients during campaign are having mostly `university.degree` followed by `high.school` and `professional.course` degree. and most of the clients are married as usual.

Relationship of Loan with accepted Target scheme

Now analyse the status of loan and credit in default of the clients who accepted the scheme-

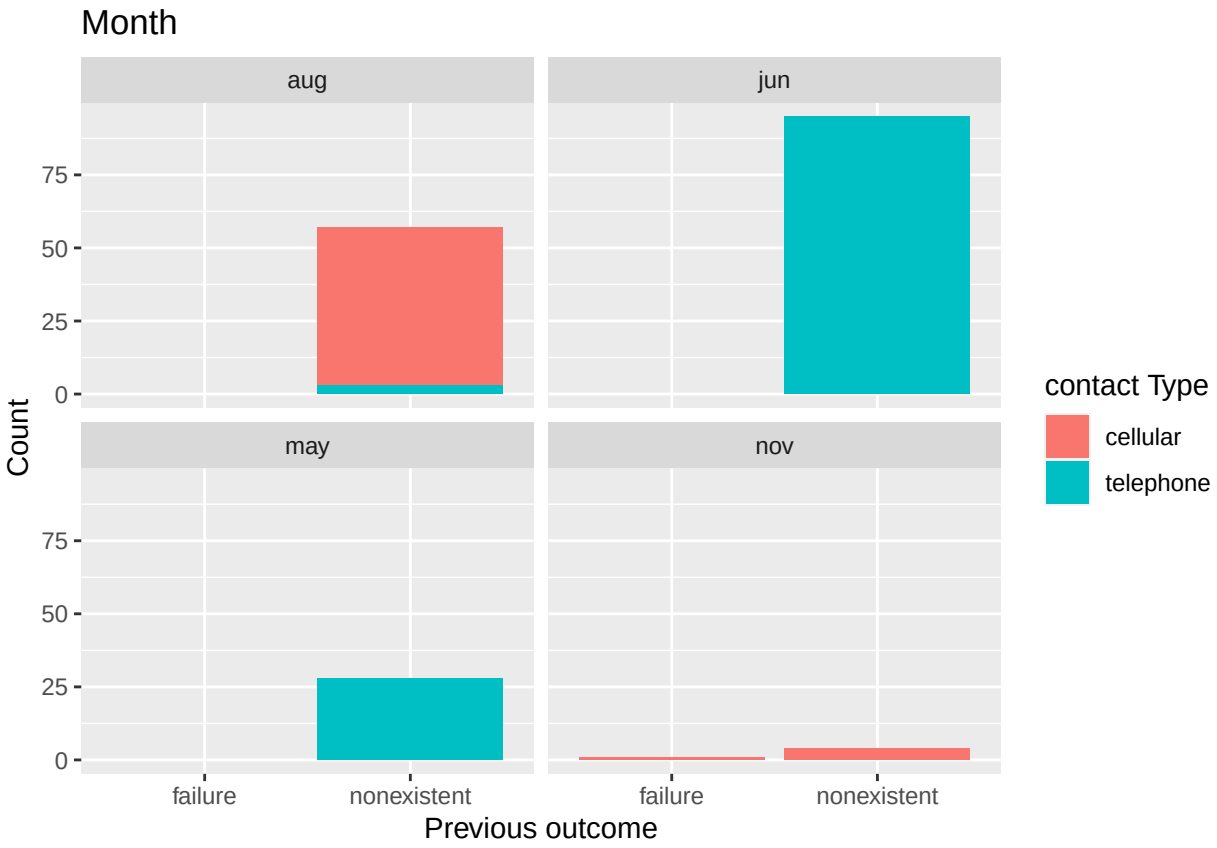


From the graph it is clear that most of the clients who accepted the term deposit scheme have `no` credit in default.

secondly fewer number of clients have personal loans.

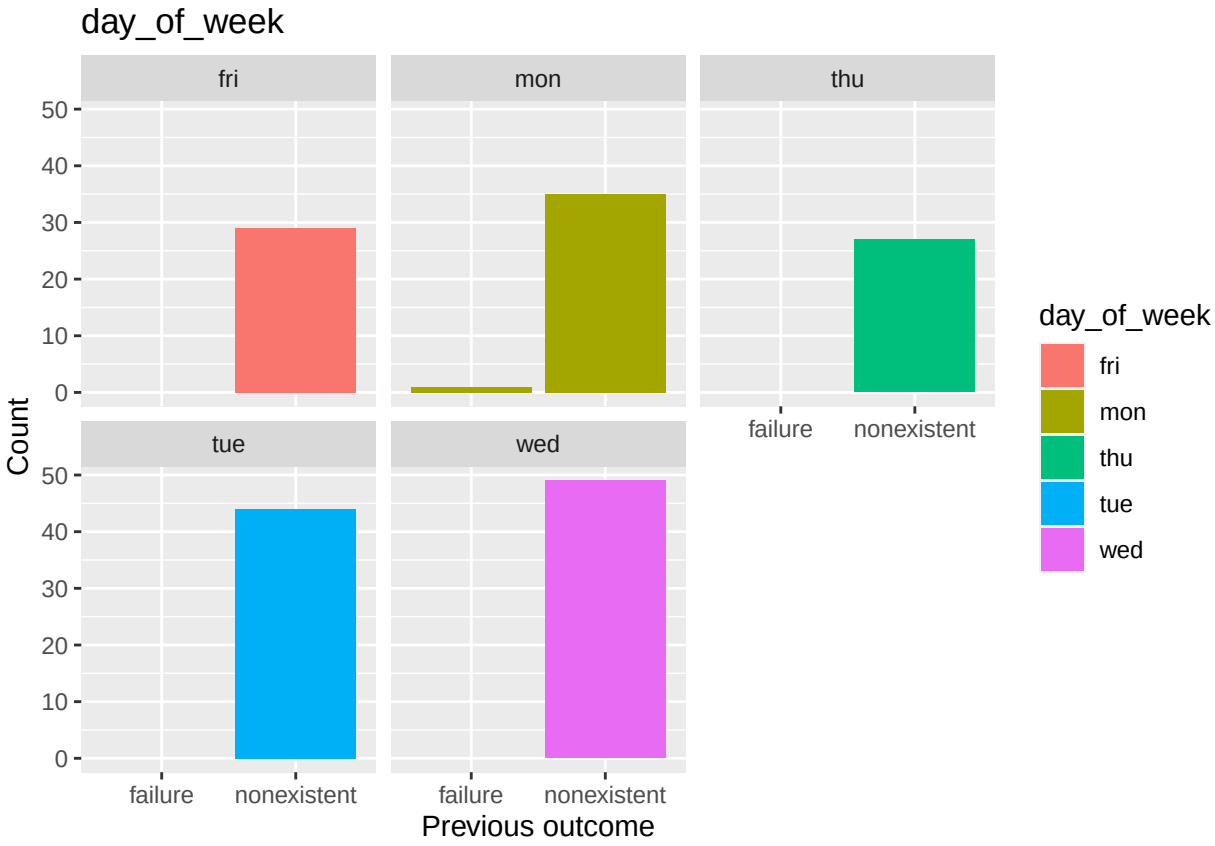
Thirdly there are an equal number of clients with or without housing loans.

Relationship of previous outcome with accepted Target scheme



Most of the clients belong to `nonexistent` categories who accepted the Term deposit scheme and contacted by telephone. secondly they were contacted in the month of `jun` followed by `aug` and `may`.

previous outcome with in day_of_week and accepted Target scheme



The result of the previous outcome suggests that clients mostly belong to `nonexistent` categories and their count is more on 'wed' followed by `tue` and `mon`.

Approach To create Predictive Model

Feature Engineering

Our next step is to perform Feature Engineering on the data so that we can get a best performing model.

Steps involved in Feature Engineering:-

- Grouping of `age` attribute
- Drop `duration` column
- Standardize numeric data
- Balancing the data by smote technique
- Re-leveling the train Data

Grouping(Binning) of age attribute

In this step we will create groups of the age attribute and label them as 0-4, 5-9, 10-14 and so on.

```
##    key   age         job marital           education default housing loan
## 1: 444 45-49  management married   university.degree      no     yes   no
## 2: 445 30-34      admin. married            basic.9y      no      no   no
## 3: 446 45-49 blue-collar married             unknown unknown      no   no
## 4: 447 40-44  technician married professional.course      no      no   no
## 5: 448 55-59  technician married            basic.4y unknown      no  yes
## 6: 449 55-59  technician married            basic.4y unknown      no   no
##      contact month day_of_week campaign pdays previous    poutcome emp.var.rate
## 1: telephone   may         tue        1   999        0 nonexistent          1.1
## 2: telephone   may         tue        1   999        0 nonexistent          1.1
## 3: telephone   may         tue        1   999        0 nonexistent          1.1
## 4: telephone   may         tue        1   999        0 nonexistent          1.1
## 5: telephone   may         tue        1   999        0 nonexistent          1.1
## 6: telephone   may         tue        1   999        0 nonexistent          1.1
##    cons.price.idx cons.conf.idx euribor3m nr.employed   y
## 1:         93.994         -36.4     4.857        5191  no
## 2:         93.994         -36.4     4.857        5191  no
## 3:         93.994         -36.4     4.857        5191  no
## 4:         93.994         -36.4     4.857        5191 yes
## 5:         93.994         -36.4     4.857        5191  no
## 6:         93.994         -36.4     4.857        5191  no
```

Standardize the data

In this step we Standardize numerical variables to the same range where mean and standard deviation of all numerical variables will be 0 and 1 respectively.

```
##    key   age         job marital           education default housing loan
## 1: 444 45-49  management married   university.degree      no     yes   no
## 2: 445 30-34      admin. married            basic.9y      no      no   no
## 3: 446 45-49 blue-collar married             unknown unknown      no   no
## 4: 447 40-44  technician married professional.course      no      no   no
## 5: 448 55-59  technician married            basic.4y unknown      no  yes
## 6: 449 55-59  technician married            basic.4y unknown      no   no
##      contact month day_of_week  campaign      pdays   previous    poutcome
## 1: telephone   may         tue -0.546413 0.02682852 -0.1044341 nonexistent
## 2: telephone   may         tue -0.546413 0.02682852 -0.1044341 nonexistent
## 3: telephone   may         tue -0.546413 0.02682852 -0.1044341 nonexistent
## 4: telephone   may         tue -0.546413 0.02682852 -0.1044341 nonexistent
## 5: telephone   may         tue -0.546413 0.02682852 -0.1044341 nonexistent
## 6: telephone   may         tue -0.546413 0.02682852 -0.1044341 nonexistent
##    emp.var.rate cons.price.idx cons.conf.idx  euribor3m nr.employed   y
## 1:   -0.3987949     -0.2570681      1.138698 -0.1965098   -1.553122  no
## 2:   -0.3987949     -0.2570681      1.138698 -0.1965098   -1.553122  no
## 3:   -0.3987949     -0.2570681      1.138698 -0.1965098   -1.553122  no
## 4:   -0.3987949     -0.2570681      1.138698 -0.1965098   -1.553122 yes
## 5:   -0.3987949     -0.2570681      1.138698 -0.1965098   -1.553122  no
## 6:   -0.3987949     -0.2570681      1.138698 -0.1965098   -1.553122  no
```

```
##      pdays            emp.var.rate
##  Min.   :-37.28979   Min.   :-3.8083
##  1st Qu.:  0.02683   1st Qu.:-0.3988
##  Median :  0.02683   Median : 0.4536
##  Mean   :  0.00000   Mean   : 0.0000
##  3rd Qu.:  0.02683   3rd Qu.: 0.4536
##  Max.   :  0.02683   Max.   : 0.4536
```

```
## [1] "standard deviation of pdays = 1"
```

Creating Balanced data

```
##      age         job marital           education default housing loan   contact
## 1: 45-49  management married   university.degree      no     yes   no telephone
## 2: 30-34      admin. married            basic.9y      no      no   no telephone
## 3: 45-49 blue-collar married             unknown unknown      no   no telephone
## 4: 40-44  technician married professional.course      no      no   no telephone
## 5: 55-59  technician married            basic.4y unknown      no  yes telephone
## 6: 55-59  technician married            basic.4y unknown      no   no telephone
##    month day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 1:   may         tue      140        1   999        0 nonexistent          1.1
## 2:   may         tue      175        1   999        0 nonexistent          1.1
## 3:   may         tue      136        1   999        0 nonexistent          1.1
## 4:   may         tue     1623        1   999        0 nonexistent          1.1
## 5:   may         tue       50        1   999        0 nonexistent          1.1
## 6:   may         tue      101        1   999        0 nonexistent          1.1
##    cons.price.idx cons.conf.idx euribor3m nr.employed   y
## 1:         93.994         -36.4     4.857        5191  no
## 2:         93.994         -36.4     4.857        5191  no
## 3:         93.994         -36.4     4.857        5191  no
## 4:         93.994         -36.4     4.857        5191 yes
## 5:         93.994         -36.4     4.857        5191  no
## 6:         93.994         -36.4     4.857        5191  no
```

```
##
##   no  yes
## 3985 3985
```

Re-Leveling of Attributes

We found new levels in the test data so that to solve our problem we releveled the train data as a test dataset.

Check levels after Re-leveling

```
## [1] "education levels in balanced_data and test dataset"
```

```
## [1] "basic.4y"            "basic.6y"            "basic.9y"
## [4] "high.school"         "illiterate"          "professional.course"
## [7] "university.degree"   "unknown"
```

```
## [1] "basic.4y"            "basic.6y"            "basic.9y"
## [4] "high.school"         "illiterate"          "professional.course"
## [7] "university.degree"   "unknown"
```

Model Building

After Feature Engineering our next step is to partition the data into train data and test data in the ratio of 70/30.

Machine learning algorithms

Classification algorithms we applied in the development of Bank term deposit Scheme with Feature Engineering steps mentioned above includes:-

- *Random Forest(Auto ML-h2o)*
- *Decision tree*
- *XGboost*
- *CatBoost*

Final Model Selection

After comparing the results of all models we find the `Decision tree` model and the 'CatBoost' model showing similar scores on Leaderboard.

Our Final Model is Performing very well with `Binning of Age attributes`,`Re-leveling` and with `Balanced data`.

Droping of `duration` column and `scaling` of the numerical columns is not improving the score of the model.

Evaluation Metric

Final Model selection done by Accuracy Score.

Submission

Files included in the submission are:-

1 *Source code*

- Data Analysis and Approach
- Predictive Model

2 *Model saved*

- data_tree_relevel4.rds
- model(Catboost Model)

- XGboost_new2_mod.rds

3 *Output file*

- relevel_4.csv

Thank You,

Anushree Tomar