FINAL PROJECT

Irem TANRIVERDI

OUTLINE

- Problem Description
- Data Description
- Explanatory Data analysis
- Model Building
- Model Selection

1. Problem Description

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

2. Data understanding

First and last 10 variables in the data

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
...	...	NA	NA	NA	NA	...	NA	NA	NA	...	NA	...	...	...	...	NA	NA
45202	53	management	married	tertiary	no	583	no	no	cellular	17	nov	226	1	184	4	success	yes
45203	34	admin.	single	secondary	no	557	no	no	cellular	17	nov	224	1	-1	0	unknown	yes
45204	23	student	single	tertiary	no	113	no	no	cellular	17	nov	266	1	-1	0	unknown	yes
45205	73	retired	married	secondary	no	2850	no	no	cellular	17	nov	300	1	40	8	failure	yes
45206	25	technician	single	secondary	no	505	no	yes	cellular	17	nov	386	2	-1	0	unknown	yes
45207	51	technician	married	tertiary	no	825	no	no	cellular	17	nov	977	3	-1	0	unknown	yes
45208	71	retired	divorced	primary	no	1729	no	no	cellular	17	nov	456	2	-1	0	unknown	yes
45209	72	retired	married	secondary	no	5715	no	no	cellular	17	nov	1127	5	184	3	success	yes
45210	57	blue-collar	married	secondary	no	668	no	no	telephone	17	nov	508	4	-1	0	unknown	no
45211	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	nov	361	2	188	11	other	no

```
      age                 job              marital          education    
 Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851  
 1st Qu.:33.00   management :9458   married :27214   secondary:23202  
 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                                     
 default        balance       housing      loan          contact     
 no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285  
 yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906  
             Median :   448                           unknown  :13020  
             Mean   :  1362                                            
             3rd Qu.:  1428                                            
             Max.   :102127                                            

      day           month              duration         campaign     
 Min.   : 1.00   Length:45211       Min.   :   0.0   Min.   : 1.000  
 1st Qu.: 8.00   Class :character   1st Qu.: 103.0   1st Qu.: 1.000  
 Median :16.00   Mode  :character   Median : 180.0   Median : 2.000  
 Mean   :15.81                      Mean   : 258.2   Mean   : 2.764  
 3rd Qu.:21.00                      3rd Qu.: 319.0   3rd Qu.: 3.000  
 Max.   :31.00                      Max.   :4918.0   Max.   :63.000  

     pdays          previous           poutcome          y            
 Min.   : -1.0   Min.   :  0.0000   failure: 4901   Length:45211      
 1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   Class :character  
 Median : -1.0   Median :  0.0000   success: 1511   Mode  :character  
 Mean   : 40.2   Mean   :  0.5803   unknown:36959                     
 3rd Qu.: -1.0   3rd Qu.:  0.0000                                     
 Max.   :871.0   Max.   :275.0000                                     
```

```
'data.frame':	45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
 $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
 $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : chr  "may" "may" "may" "may" ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ y        : chr  "no" "no" "no" "no" ...
```

- Bank dataset includes 45211 observations and 17 variables.
- There are 7 numeric variables which are age, balance, day, duration, campaign, pdays, and previous.
- There are 10 categorical variables which are job, martial, education, default, housing, loan, contact, month, poutcome and y.

1. age (numeric)
2. job : type of job (categorical: ‘admin.’,‘blue collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’)
3. marital : marital status (categorical: ‘divorced’,‘married’,‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed)
4. education (categorical: ‘basic.4y’, ‘basic.6y’, ‘basic.9y’, ‘high.school’, ‘illiterate’,‘ professional.course’, ‘university. Degree’, ‘unknown’)
5. default: has credit in default? (categorical: ‘no’, ‘yes’,‘unknown’)
6. housing: has housing loan? (categorical: ‘no’, ‘yes’, ‘unknown’)
7. loan: has personal loan? (categorical: ‘no’, ‘yes’,‘unknown’) related with the last contact of the current campaign:
8. contact: contact communication type (categorical: ‘cellular’, ‘telephone’)
9. month: last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’, …, ‘nov’, ‘dec’)
10. day_of_week: last contact day of the week (categorical: ‘mon’, ‘tue’, ‘wed’, ‘thu’, ‘fri’)
11. duration: last contact duration, in seconds (numeric)
12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. 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)
14. previous: number of contacts performed before this campaign and for this client (numeric)

15. poutcome: outcome of the previous marketing campaign (categorical: ‘failure’, ‘nonexistent’, ‘success’)

Output variable (desired target):

17. y: has the client subscribed a term deposit? (Binary: ‘yes’, ‘no’)

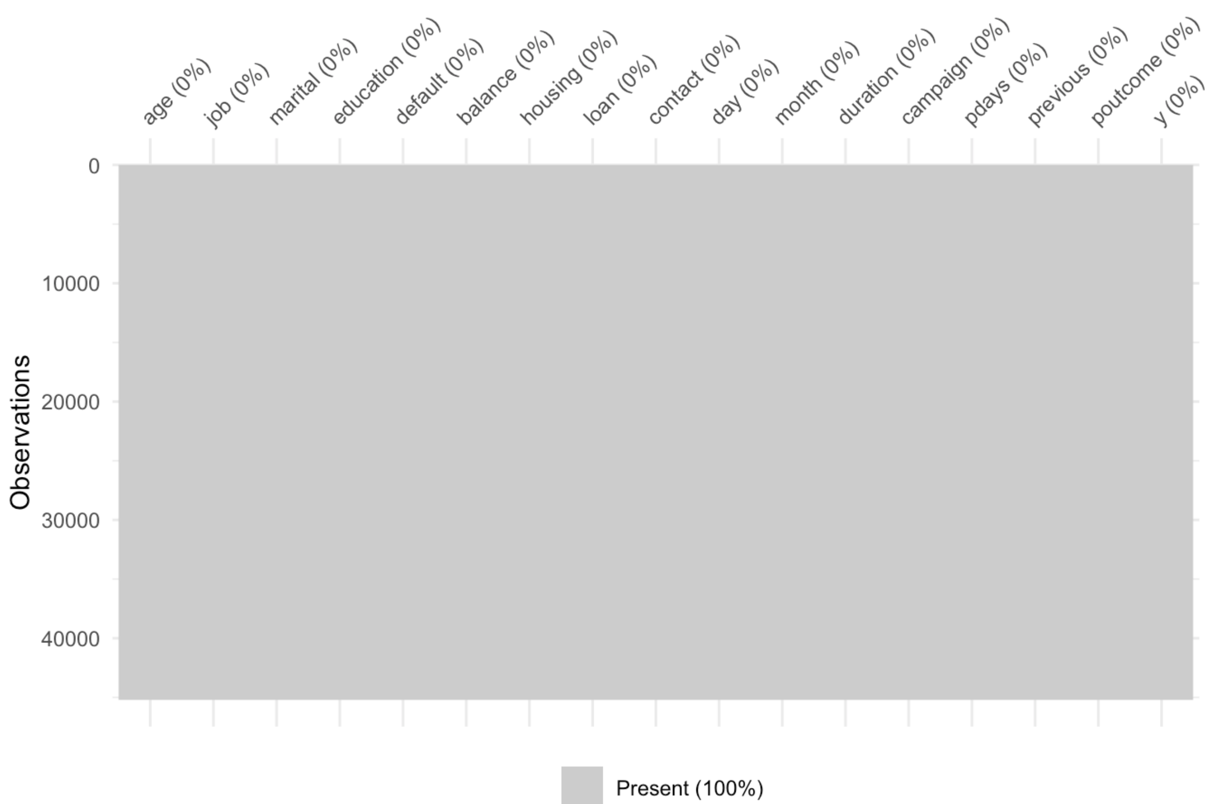
Exploratory data Analysis

Is there duplicated rows in the data?

```
# A tibble: 1 x 1
        n
    <int>
1       0
```

- As seen there is not duplicated rows in the data.

Is there any missing value in the data?

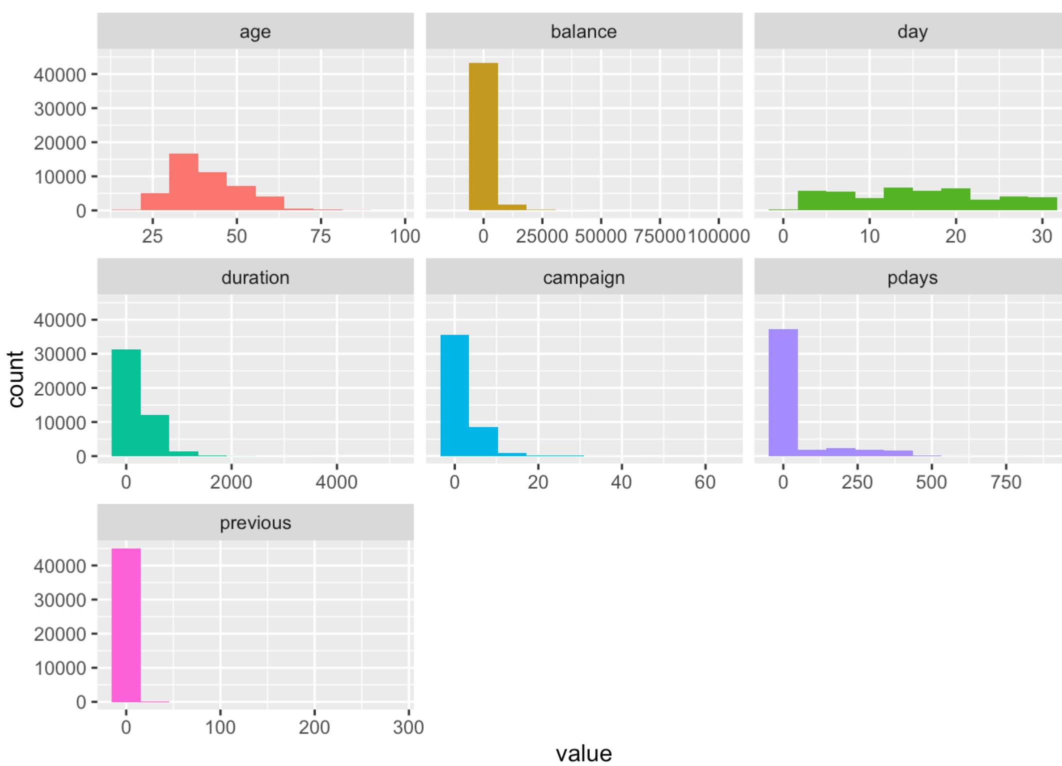


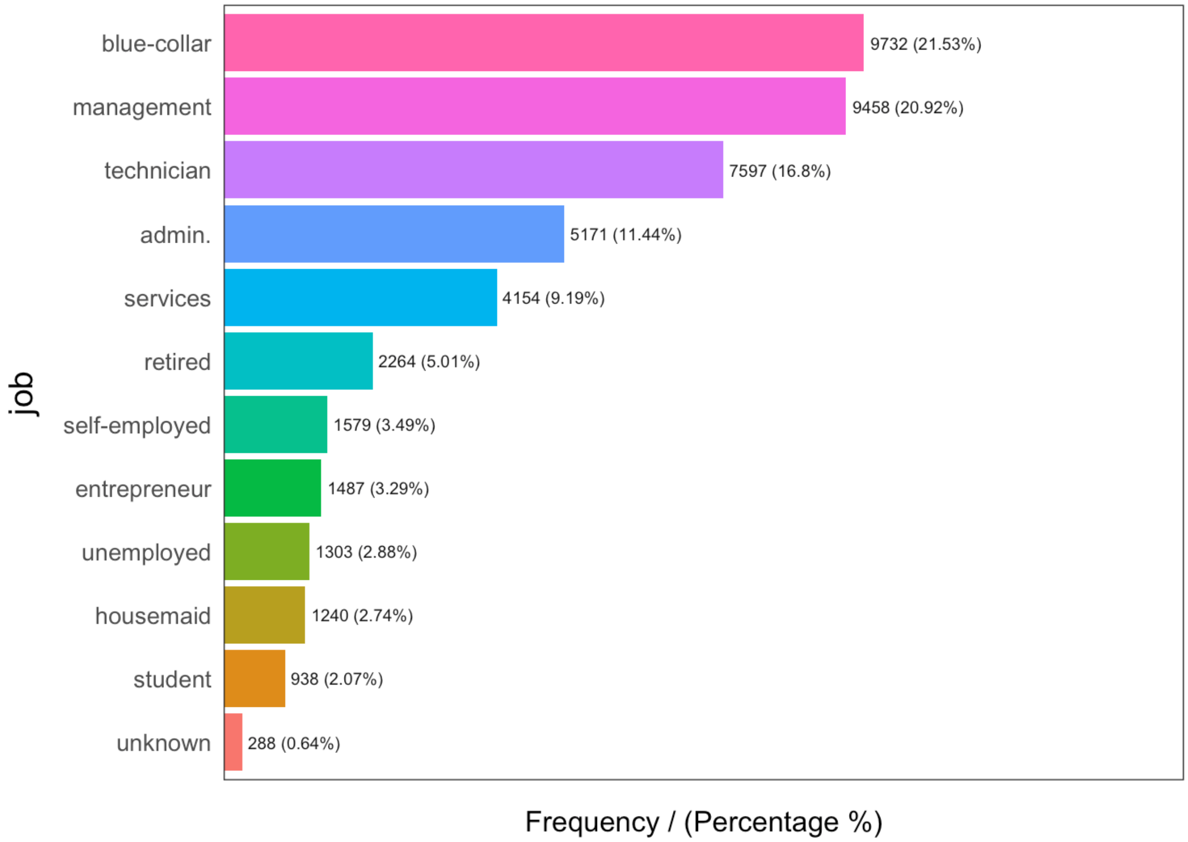
- As seen there is no missing value in the data.

Frequencies of the categorical variables and distributions of the numeric variables

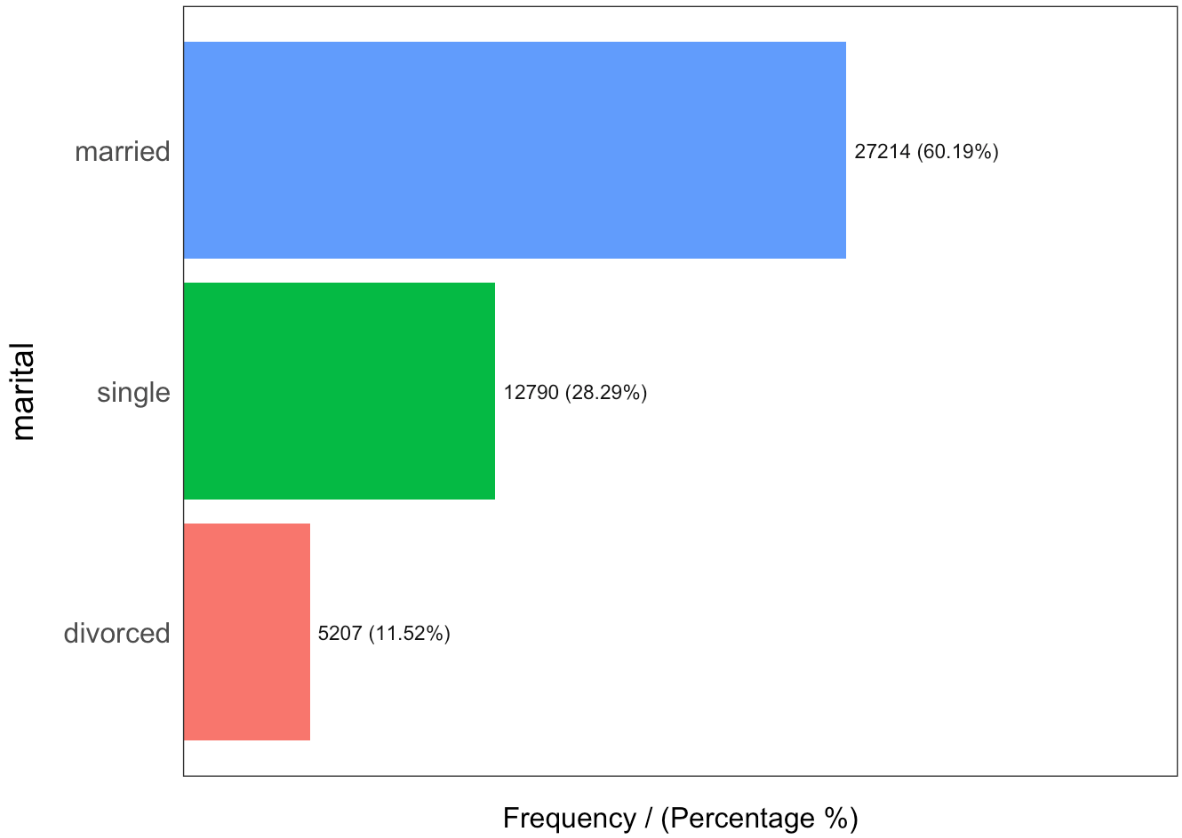
	variable	mean	std_dev	variation_coef	p_01	p_05	p_25	p_50	p_75
1	age	40.9362102	10.618762	0.2593978	23	27	33	39	48
2	balance	1362.2720577	3044.765829	2.2350644	-627	-172	72	448	1428
3	day	15.8064188	8.322476	0.5265251	2	3	8	16	21
4	duration	258.1630798	257.527812	0.9975393	11	35	103	180	319
5	campaign	2.7638407	3.098021	1.1209115	1	1	1	2	3
6	pdays	40.1978280	100.128746	2.4908994	-1	-1	-1	-1	-1
7	previous	0.5803234	2.303441	3.9692371	0	0	0	0	0

	p_95	p_99	skewness	kurtosis	iqr	range_98	range_80
1	59	71.0	0.68479520	3.319402	15	[23, 71]	[29, 56]
2	5768	13164.9	8.36003095	143.735848	1356	[-627, 13164.9]	[0, 3574]
3	29	31.0	0.09307593	1.940087	13	[2, 31]	[5, 28]
4	751	1269.0	3.14421378	21.151775	216	[11, 1269]	[58, 548]
5	8	16.0	4.89848764	42.245178	2	[1, 16]	[1, 5]
6	317	370.0	2.61562869	9.934296	0	[-1, 370]	[-1, 185]
7	3	8.9	41.84506609	4509.362118	0	[0, 8.90000000000146]	[0, 2]

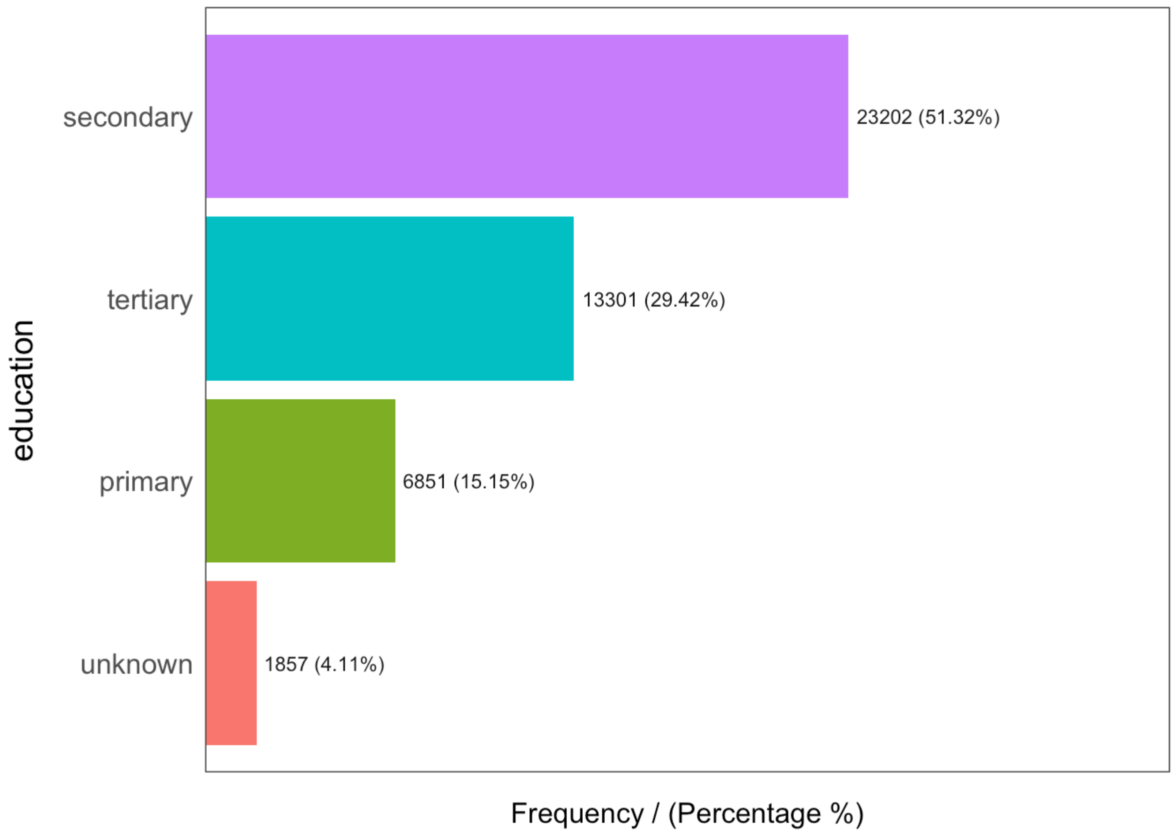




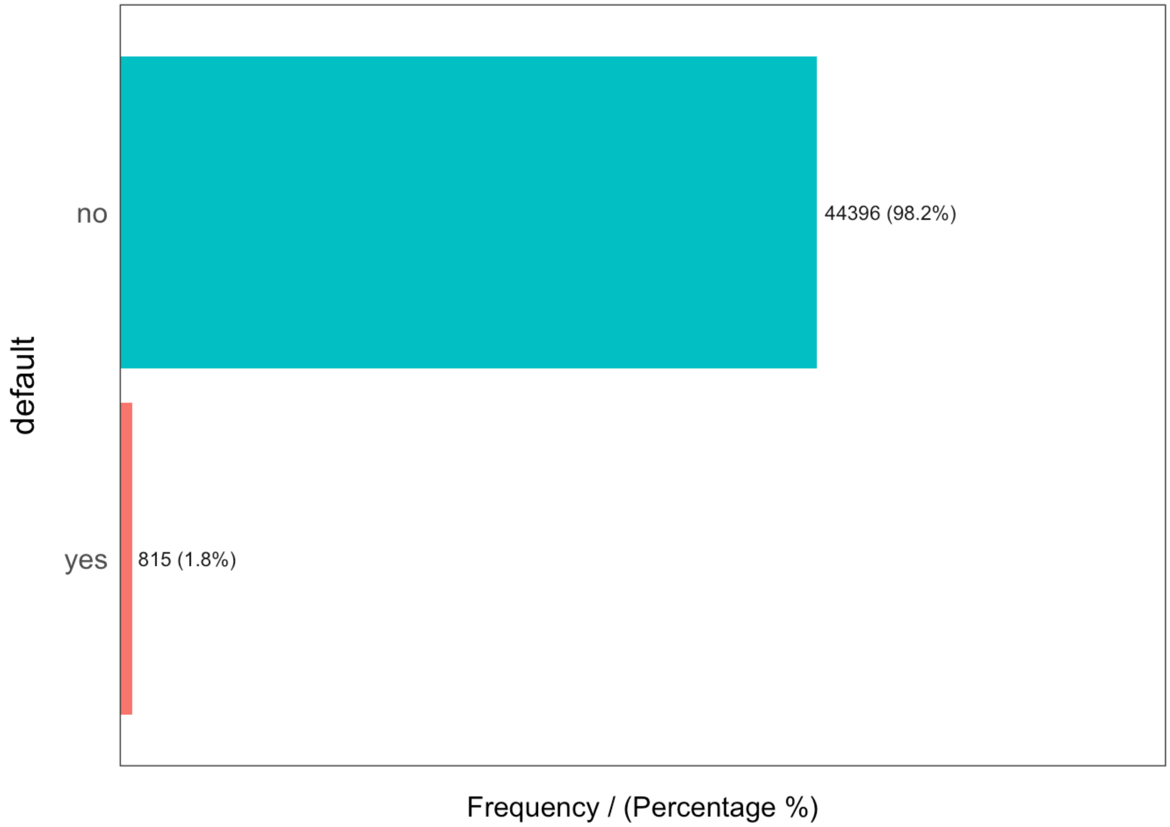

```
##              job frequency percentage cumulative_perc
## 1    blue-collar      9732      21.53           21.53
## 2     management      9458      20.92           42.45
## 3     technician      7597      16.80           59.25
## 4         admin.      5171      11.44           70.69
## 5       services      4154       9.19           79.88
## 6        retired      2264       5.01           84.89
## 7  self-employed      1579       3.49           88.38
## 8   entrepreneur      1487       3.29           91.67
## 9     unemployed      1303       2.88           94.55
## 10     housemaid      1240       2.74           97.29
## 11       student       938       2.07           99.36
## 12       unknown       288       0.64          100.00
```



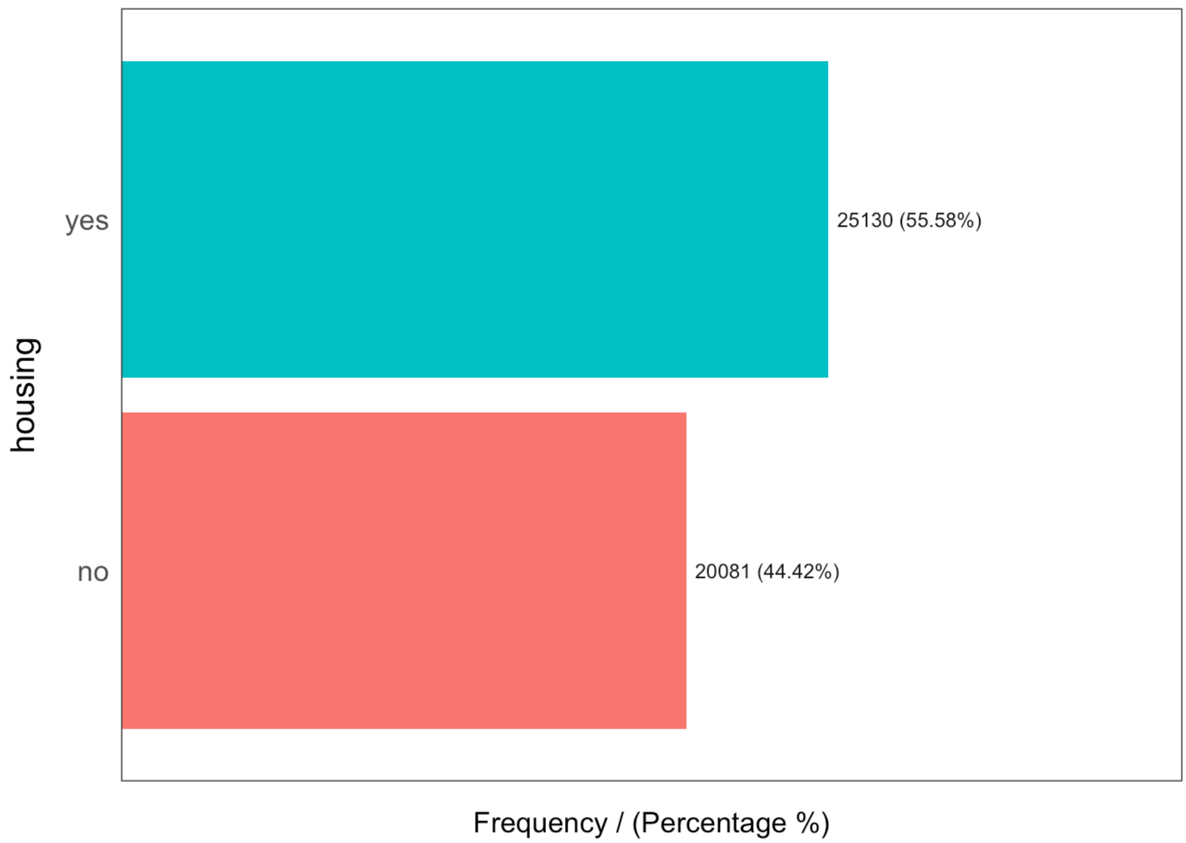

```
##    marital frequency percentage cumulative_perc
## 1  married     27214      60.19           60.19
## 2   single     12790      28.29           88.48
## 3 divorced      5207      11.52          100.00
```



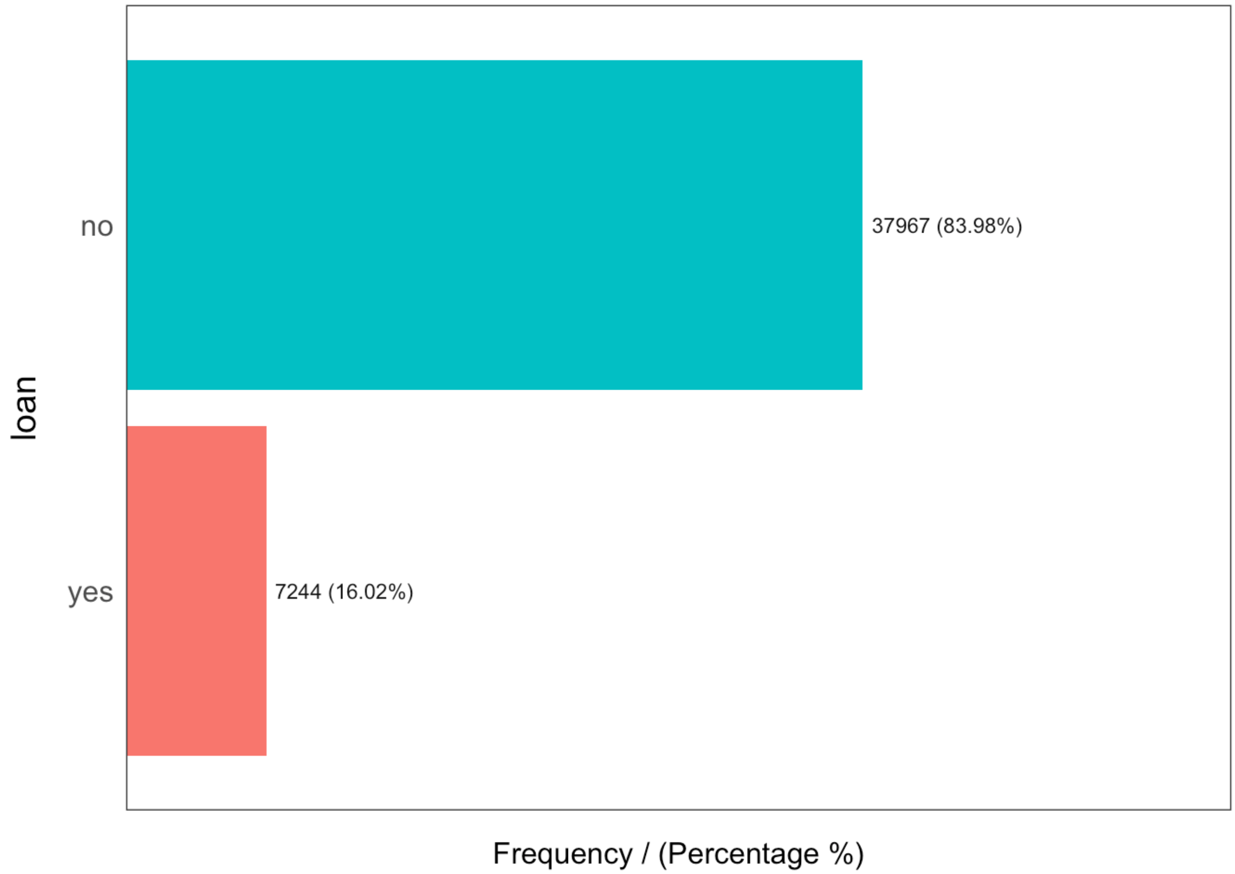

```
##   education frequency percentage cumulative_perc
## 1 secondary     23202      51.32           51.32
## 2  tertiary     13301      29.42           80.74
## 3   primary      6851      15.15           95.89
## 4   unknown      1857       4.11          100.00
```



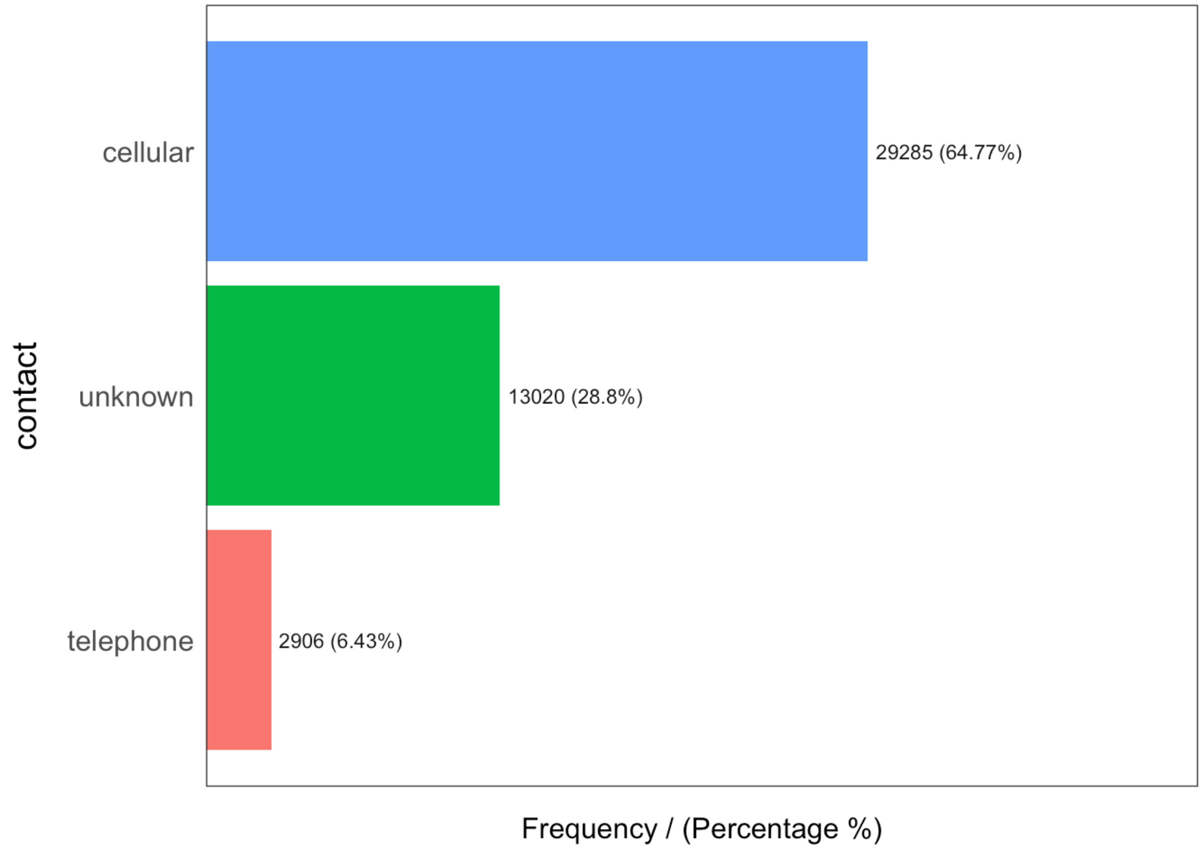

```
##   default frequency percentage cumulative_perc
## 1      no     44396       98.2            98.2
## 2     yes       815        1.8           100.0
```



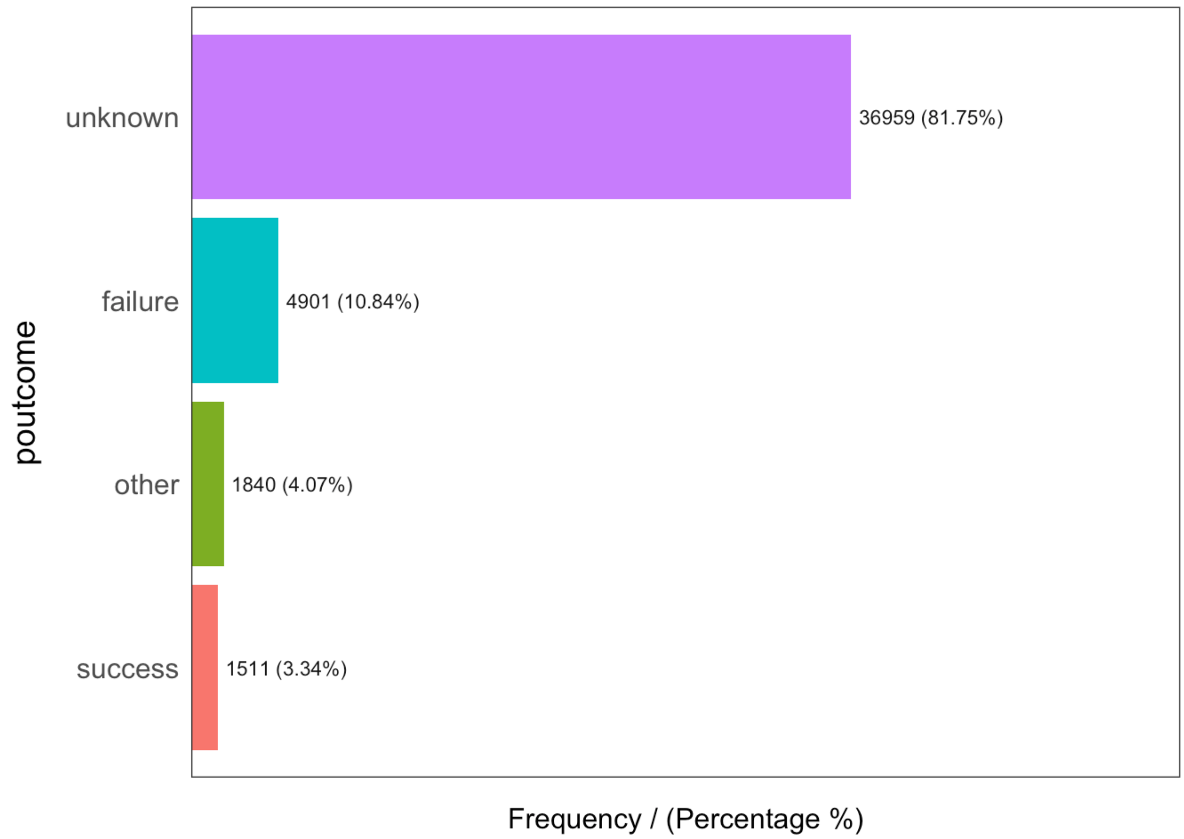

```
##   housing frequency percentage cumulative_perc
## 1     yes     25130      55.58           55.58
## 2      no     20081      44.42          100.00
```



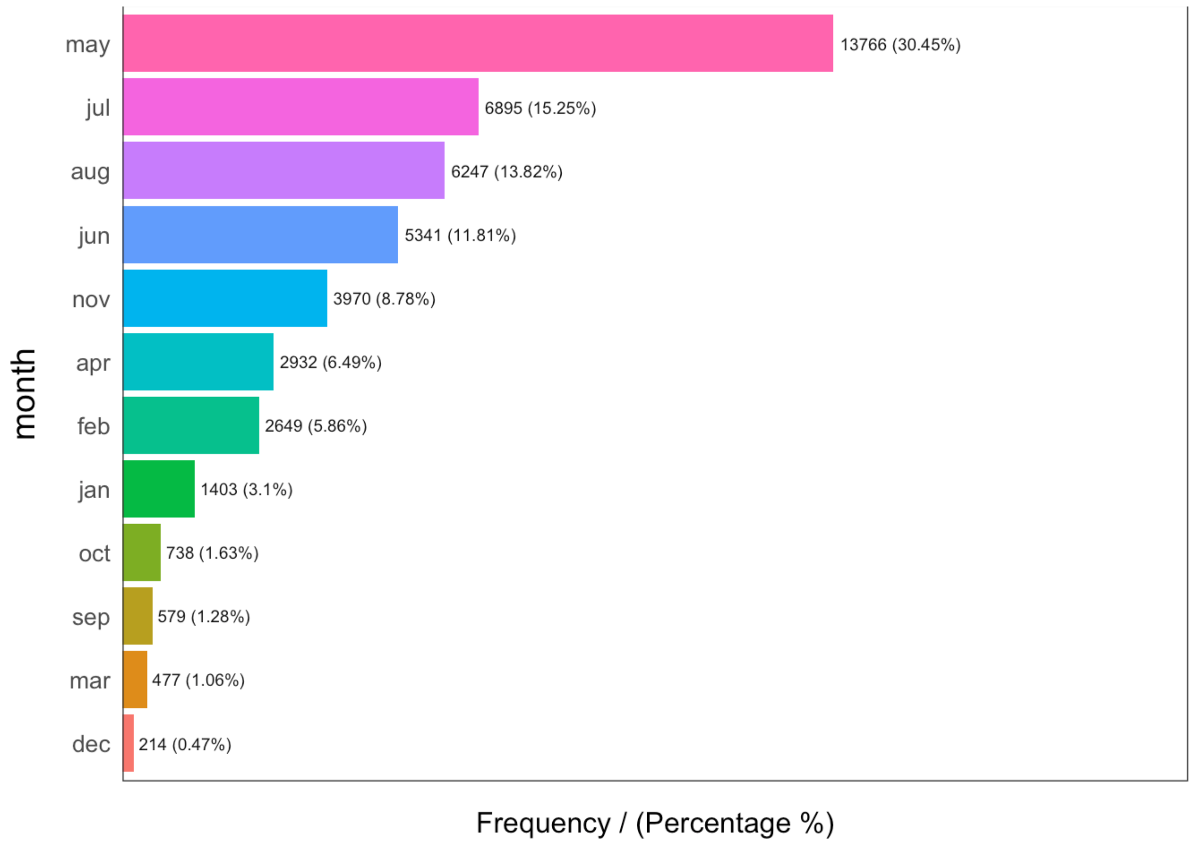

```
##   loan frequency percentage cumulative_perc
## 1   no     37967      83.98           83.98
## 2  yes      7244      16.02          100.00
```



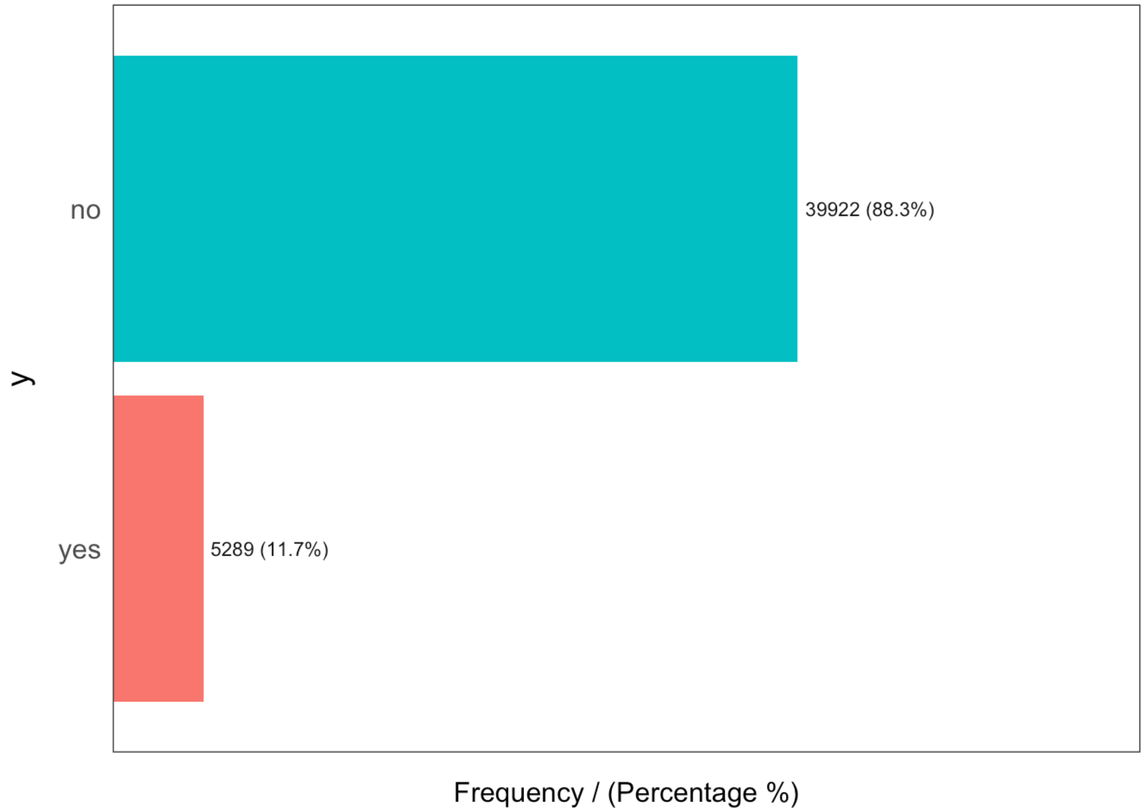

```
##     contact frequency percentage cumulative_perc
## 1  cellular     29285      64.77           64.77
## 2   unknown     13020      28.80           93.57
## 3 telephone      2906       6.43          100.00
```




```
##   poutcome frequency percentage cumulative_perc
## 1  unknown     36959      81.75           81.75
## 2  failure      4901      10.84           92.59
## 3    other      1840       4.07           96.66
## 4  success      1511       3.34          100.00
```



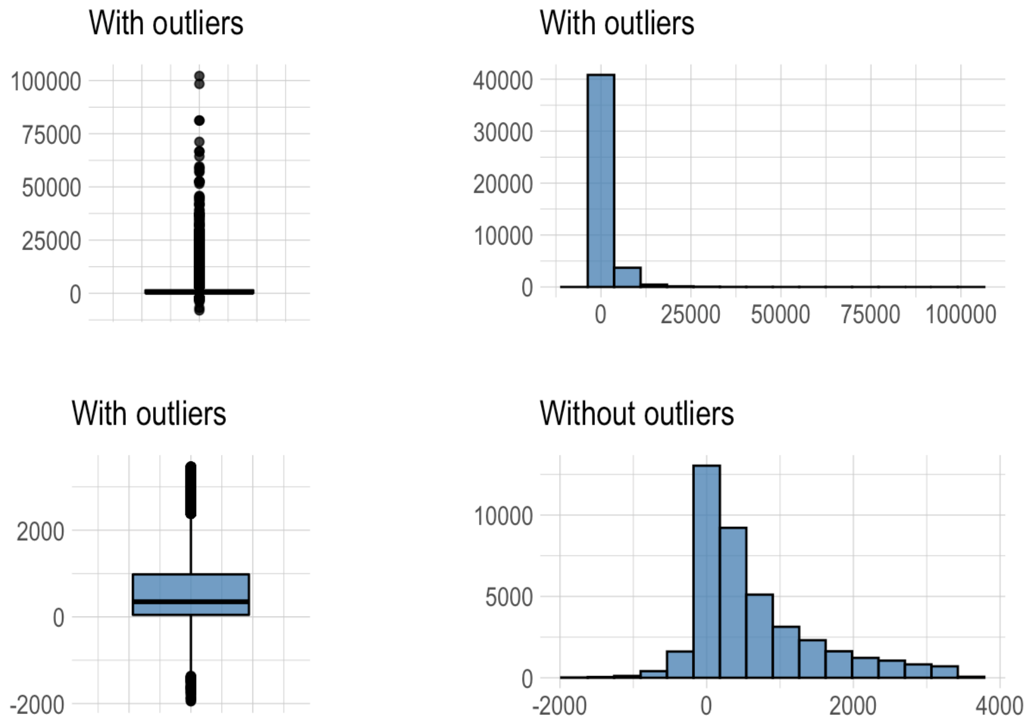

```
##    month frequency percentage cumulative_perc
## 1    may     13766      30.45           30.45
## 2    jul      6895      15.25           45.70
## 3    aug      6247      13.82           59.52
## 4    jun      5341      11.81           71.33
## 5    nov      3970       8.78           80.11
## 6    apr      2932       6.49           86.60
## 7    feb      2649       5.86           92.46
## 8    jan      1403       3.10           95.56
## 9    oct       738       1.63           97.19
## 10   sep       579       1.28           98.47
## 11   mar       477       1.06           99.53
## 12   dec       214       0.47          100.00
```



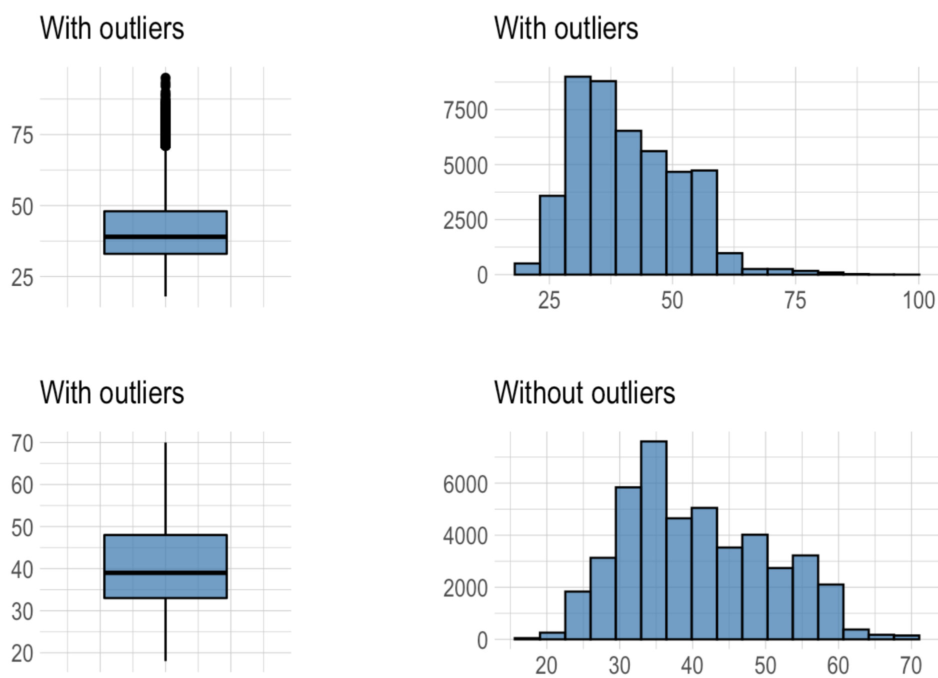

```
##     y frequency percentage cumulative_perc
## 1  no     39922       88.3            88.3
## 2 yes      5289       11.7           100.0
```

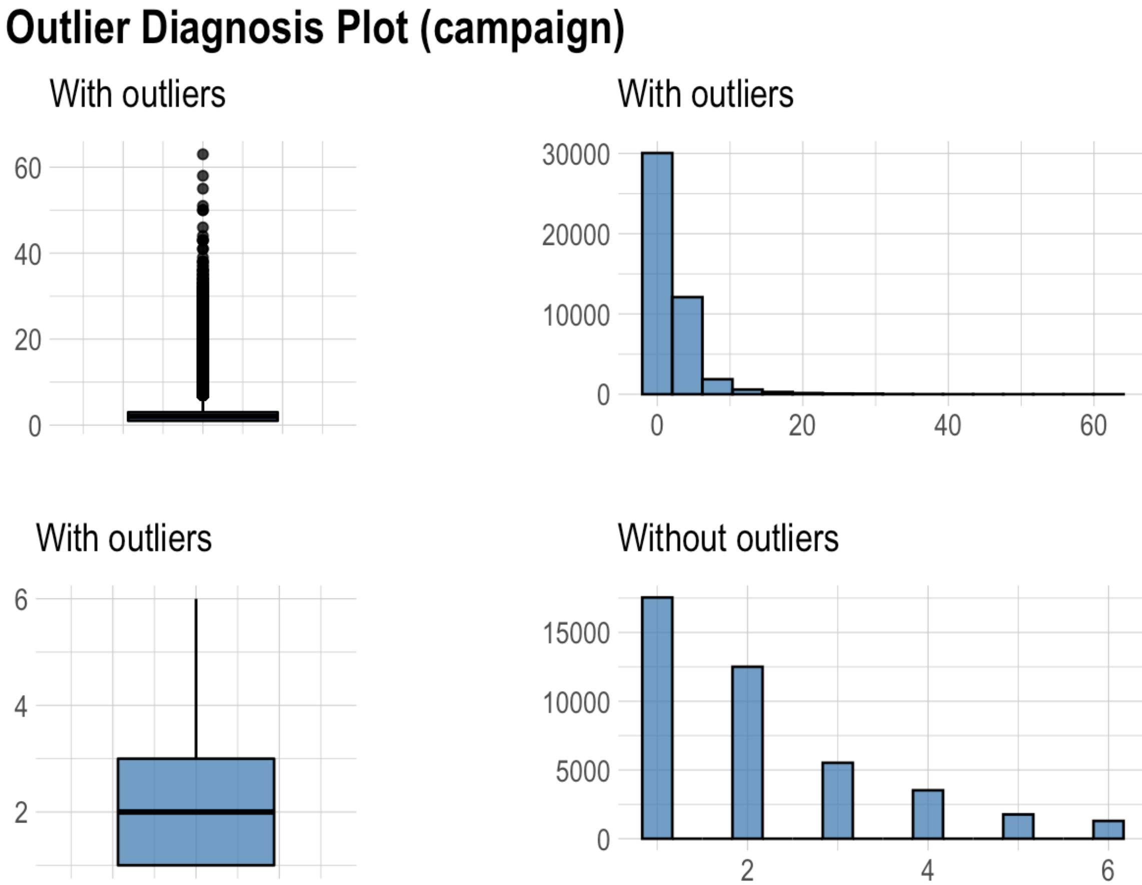
Do numeric variables have any outlier and what will be shape of the variables exclude the outliers?

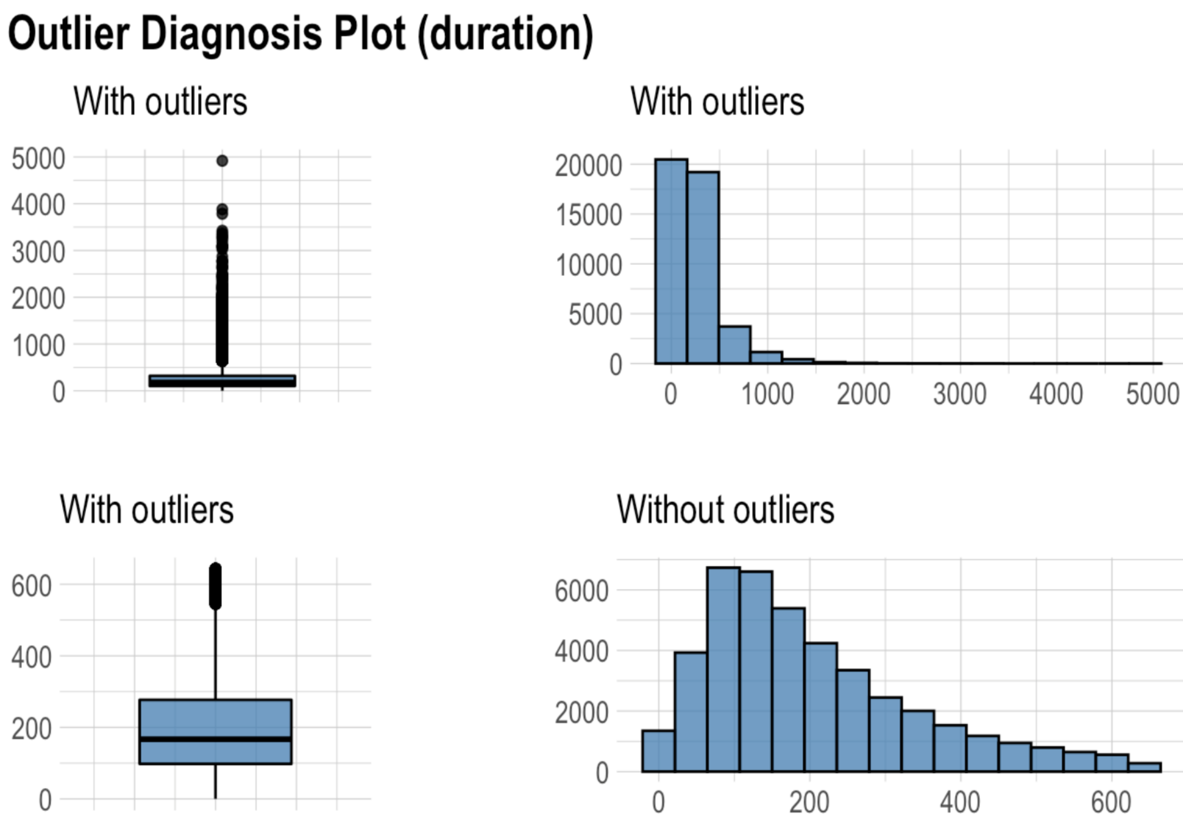
Outlier Diagnosis Plot (balance)



Outlier Diagnosis Plot (age)

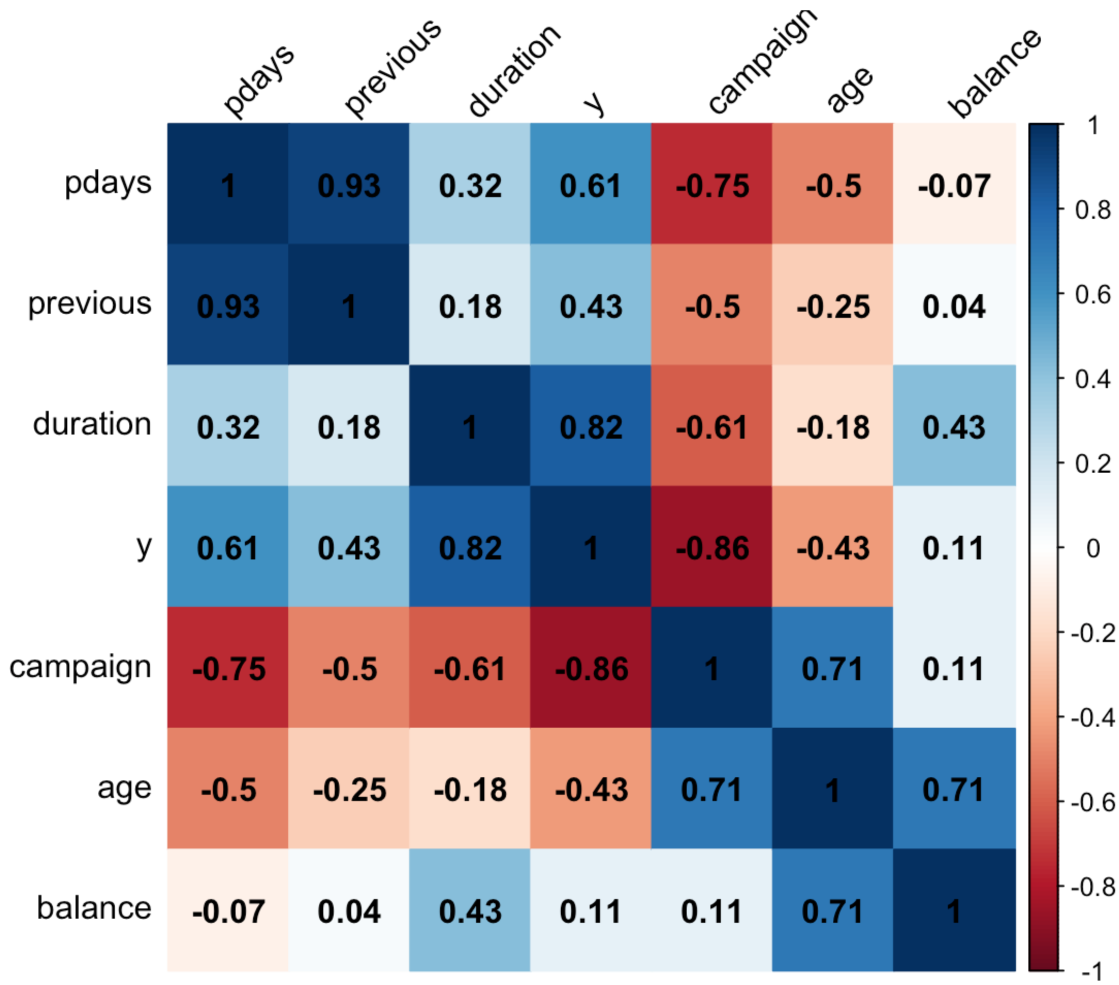



Outlier Diagnosis Plot (campaign)
With outliers
With outliers
With outliers
Without outliers
60
40
20
0
30000
20000
10000
0
0
20
40
60
6
4
2
15000
10000
5000
0
2
4
6



- As seen all of the 4 numeric variables have outlier.
- Shapes of the variables changed when outliers removed.

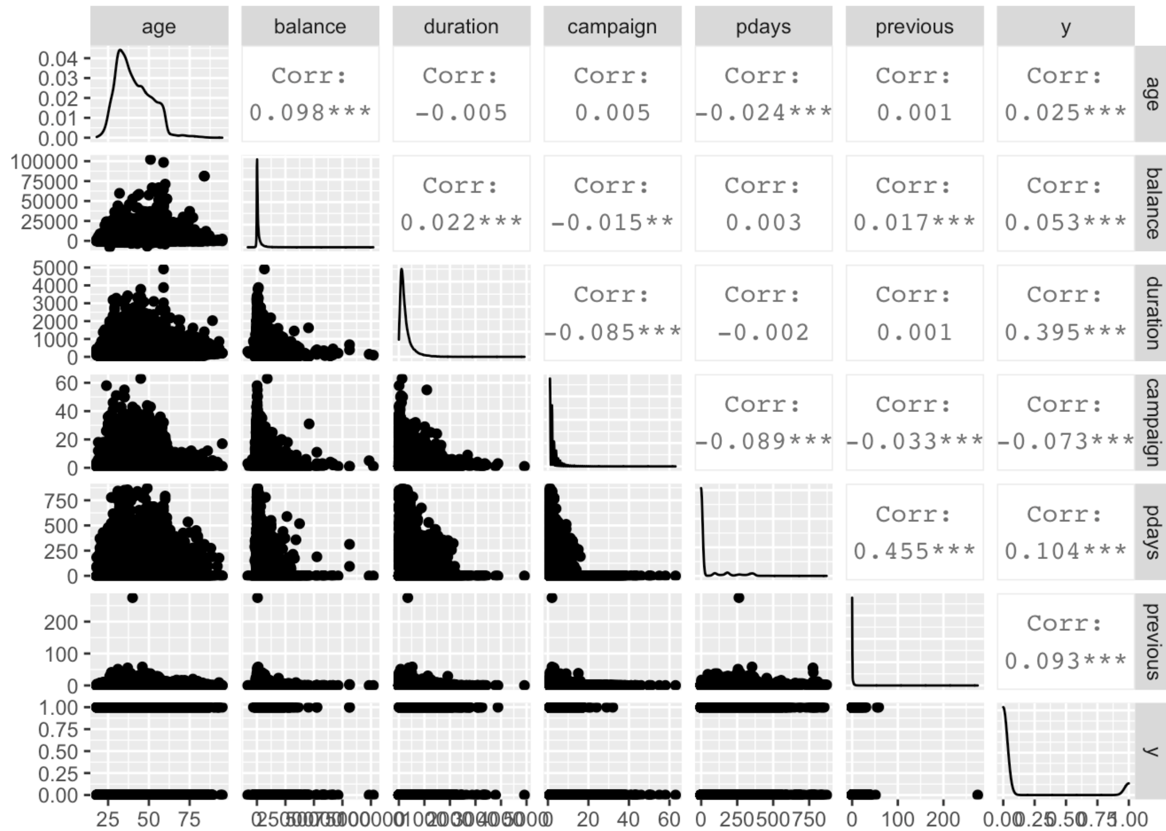
Is there any significant relationship between numeric variables and y, if y taken as numeric (1:no, 0: yes)?



Positive correlations are shown in blue and negative correlations in red color. Color intensity is proportional to the correlation coefficients. Let's look at the correlation matrix to examine which variables have strong relationship with response variable y.

- Between y and duration, there is strong positive relationship.
- Between y and campaign, there is strong negative relationship.

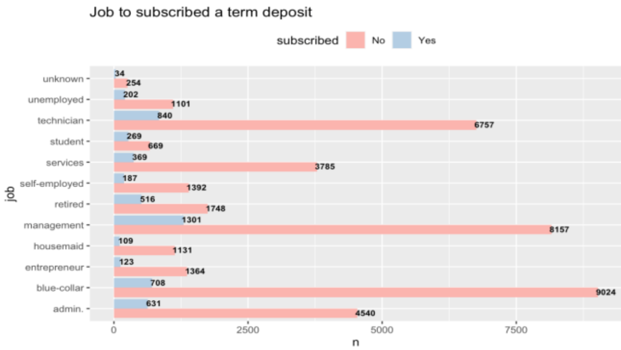
We can also see the relationship between other variables (covariates).

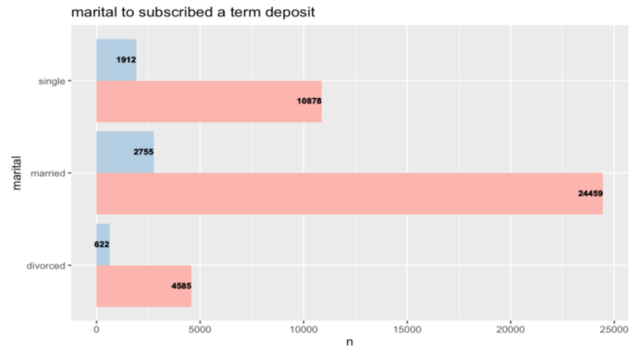


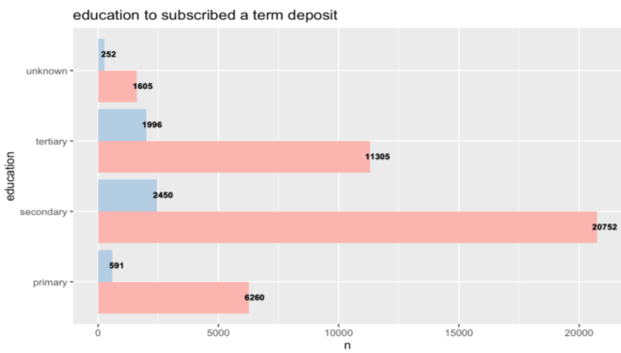
We can see from this plot we can see that if the relationship between variables significant or not. We can see that all the relationship between covariates and y are significant.

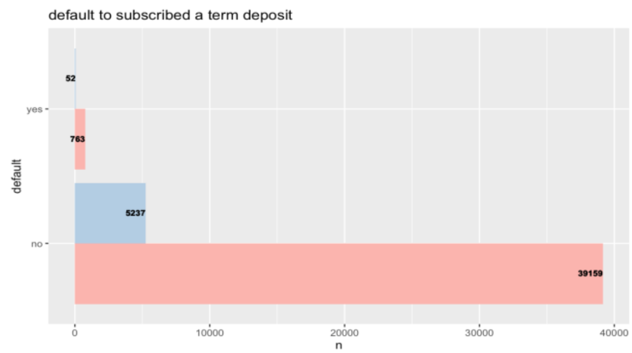
Is there any significant relationship between categorical variables and y?

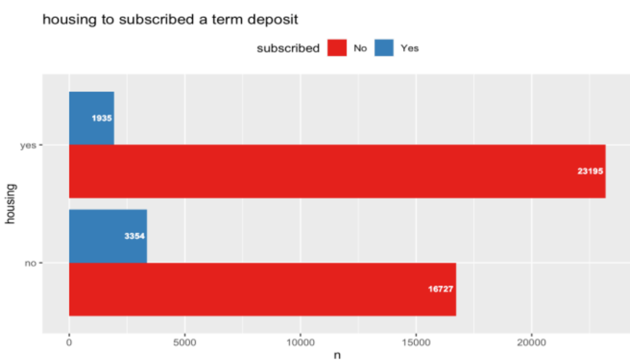
H_0: $There\ is\ not\ significant\ relationship\ between\ variables\ (Variables\ are\ independent)$

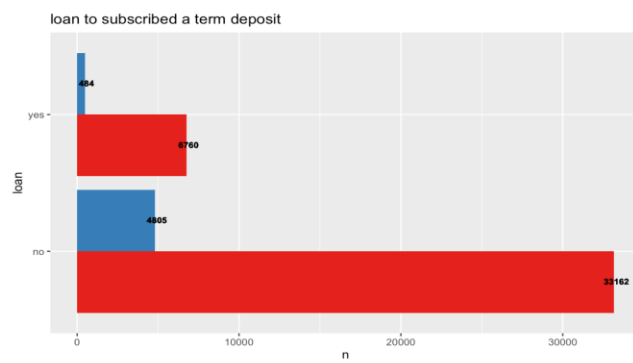

Job to subscribed a term deposit
subscribed
No
Yes
job
unknown
unemployed
technician
student
services
self-employed
retired
management
housemaid
entrepreneur
blue-collar
admin.
34
254
202
1101
840
6757
269
669
369
3785
187
1392
516
1748
1301
8157
109
1131
123
1364
708
9024
631
4540
0
2500
5000
7500
n

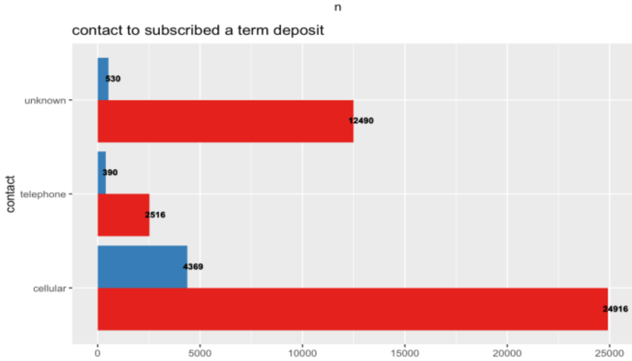

marital to subscribed a term deposit
marital
single
married
divorced
1912
10878
2755
24459
622
4585
0
5000
10000
15000
20000
25000
n

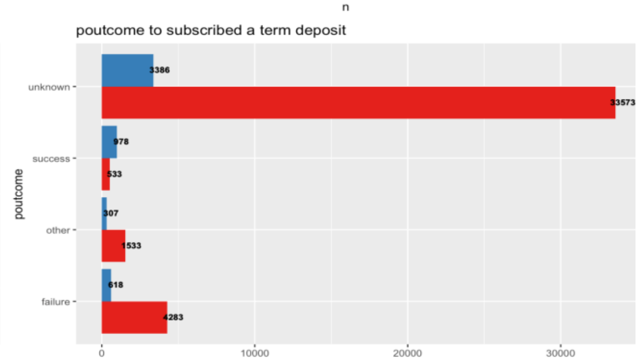

education to subscribed a term deposit
education
unknown
tertiary
secondary
primary
252
1605
1996
11305
2450
20752
591
6260
0
5000
10000
15000
20000
n


default to subscribed a term deposit
default
yes
no
52
763
5237
39159
0
10000
20000
30000
40000
n


housing to subscribed a term deposit
subscribed
No
Yes
housing
yes
no
1935
23195
3354
16727
0
5000
10000
15000
20000
n


loan to subscribed a term deposit
loan
yes
no
484
6760
4805
33162
0
10000
20000
30000
n


contact to subscribed a term deposit
contact
unknown
telephone
cellular
530
12490
390
2516
4369
24916
0
5000
10000
15000
20000
25000
n


poutcome to subscribed a term deposit
poutcome
unknown
success
other
failure
3386
33573
978
533
307
1533
618
4283
0
10000
20000
30000
n

	statistic	p.value
job	836.1055	0.000499
marital	196.4959	0.000499
education	238.9235	0.000499
default	22.7235	0.000499
housing	875.6937	0.000499
loan	210.1949	0.000499
contact	1035.714	0.000499
poutcome	4391.507	0.000499

- As seen all p-values are smaller than the significance level of 0.05, so there is significant relationship between categorical variables and y.

MODELING

Data Preparation

We see that in EDA part, in response variable, "no" class proportion is 88.3 while "yes" class proportion is 11.7. There is huge difference difference between two class. Thus, we have imbalance data, and it causes reduction in accuracy of ML algorithms.

What are the methods to deal with imbalanced data sets?

The methods are widely known as 'Sampling Methods'. Generally, these methods aim to modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by altering the size of original data set and provide the same proportion of balance.

Below are the methods used to treat imbalanced datasets:

- Undersampling
- Oversampling
- Synthetic Data Generation
- Cost Sensitive Learning
- I applied both under sampling and oversampling since you we’ve lost significant information from the sample when doing undersampling.
- In this case, the minority class is oversampled with replacement and majority class is under sampled without replacement.
- After under and oversampling number of response class be:

No	Yes
22628 (0.5004%)	22583 (0.499%)

- After over and undersampling data divided into two parts: training and test set.
- 80% of the data used as training set and 20% of the data used as test set.

Nrow train set	Nrow test set
36169	9042

- In train data, proportion of class of the response variable is:

No	**Yes**
0.50062%	0.4993%

- In all of models, y taken as response variable (by taking 0: No, 1: Yes), all of the other variables taking as covariate.

LOGISTIC REGRESSION

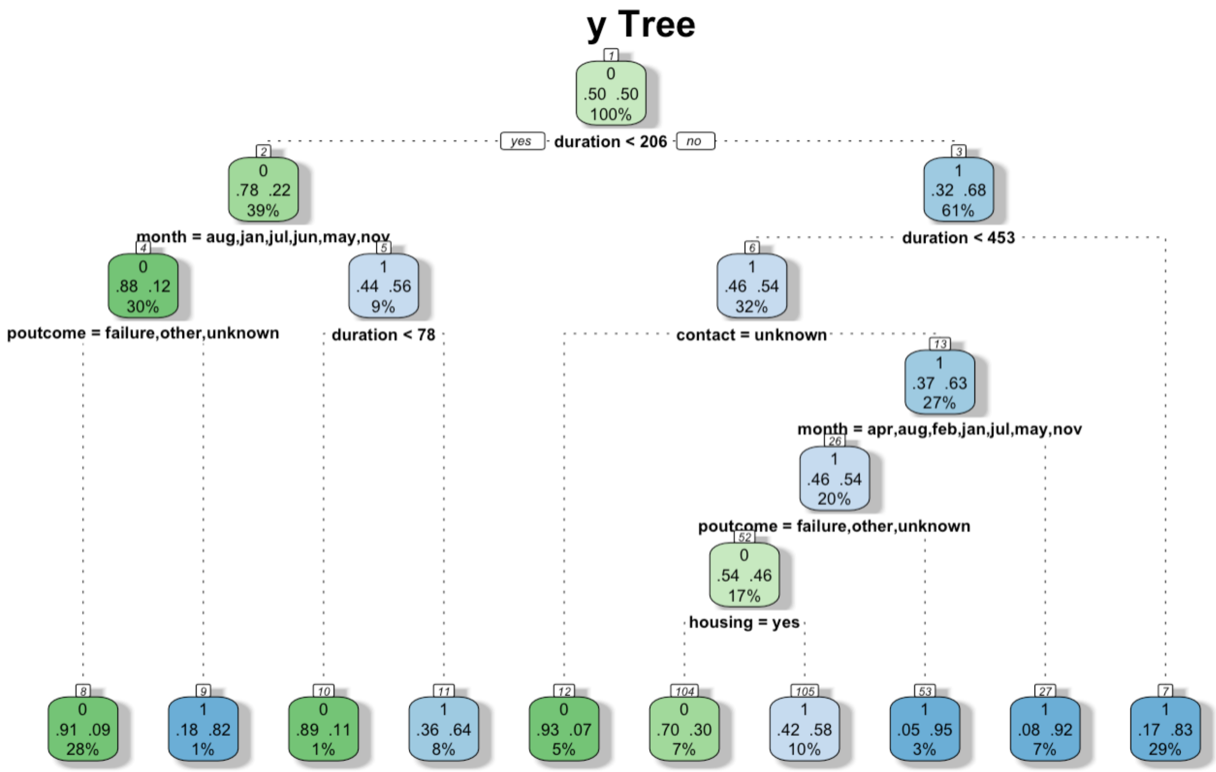
```
Call:
glm(formula = y ~ ., family = binomial(link = "logit"), data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.1313  -0.5893  -0.0601   0.5988   2.9451

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.460e-01  1.503e-01  -4.964 6.89e-07 ***
age                -9.109e-04  1.798e-03  -0.507  0.61236
jobblue-collar     -3.345e-01  5.796e-02  -5.771 7.87e-09 ***
jobentrepreneur    -3.902e-01  9.889e-02  -3.946 7.94e-05 ***
jobhousemaid       -4.098e-01  1.049e-01  -3.909 9.29e-05 ***
jobmanagement      -9.117e-02  6.011e-02  -1.517  0.12935
jobretired          4.064e-01  8.213e-02   4.949 7.47e-07 ***
jobself-employed   -2.232e-01  9.052e-02  -2.466  0.01365 *
jobservices        -2.869e-01  6.739e-02  -4.258 2.07e-05 ***
jobstudent          7.235e-01  9.822e-02   7.366 1.76e-13 ***
jobtechnician      -1.004e-01  5.569e-02  -1.803  0.07134 .
jobunemployed      -8.228e-02  9.312e-02  -0.884  0.37694
jobunknown         -2.553e-01  1.915e-01  -1.333  0.18240
maritalmarried     -1.878e-01  4.783e-02  -3.927 8.62e-05 ***
maritalsingle       1.205e-01  5.497e-02   2.192  0.02838 *
educationsecondary  2.280e-01  5.191e-02   4.391 1.13e-05 ***
educationtertiary   4.179e-01  6.126e-02   6.823 8.93e-12 ***
educationunknown    2.802e-01  8.518e-02   3.290  0.00100 **
defaultyes          1.003e-01  1.218e-01   0.824  0.41003
balance             2.341e-05  5.007e-06   4.675 2.94e-06 ***
housingyes         -7.052e-01  3.480e-02 -20.267  < 2e-16 ***
loanyes            -5.427e-01  4.672e-02 -11.615  < 2e-16 ***
contacttelephone   -4.067e-02  6.089e-02  -0.668  0.50417
contactunknown     -1.718e+00  5.404e-02 -31.787  < 2e-16 ***
day                 4.870e-03  1.976e-03   2.465  0.01371 *
```

```
day                  4.870e-03  1.976e-03   2.465  0.01371 *  
monthaug            -9.246e-01  6.292e-02 -14.696  < 2e-16 ***
monthdec             6.884e-01  1.789e-01   3.847  0.00012 ***
monthfeb            -1.044e-01  7.124e-02  -1.465  0.14297    
monthjan            -1.302e+00  9.564e-02 -13.613  < 2e-16 ***
monthjul            -1.078e+00  6.312e-02 -17.071  < 2e-16 ***
monthjun             3.044e-01  7.397e-02   4.116 3.86e-05 ***
monthmar             1.715e+00  1.202e-01  14.264  < 2e-16 ***
monthmay            -6.591e-01  6.013e-02 -10.962  < 2e-16 ***
monthnov            -1.025e+00  6.912e-02 -14.826  < 2e-16 ***
monthoct             1.241e+00  1.022e-01  12.138  < 2e-16 ***
monthsep             9.476e-01  1.161e-01   8.161 3.33e-16 ***
duration             5.698e-03  7.200e-05  79.143  < 2e-16 ***
campaign            -1.067e-01  7.754e-03 -13.760  < 2e-16 ***
pdays               -4.373e-04  2.416e-04  -1.810  0.07027 .  
previous             1.928e-02  8.812e-03   2.188  0.02864 *  
poutcomeother        1.138e-01  7.449e-02   1.527  0.12665    
poutcomesuccess      2.504e+00  8.438e-02  29.677  < 2e-16 ***
poutcomeunknown     -2.512e-01  7.915e-02  -3.173  0.00151 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50141  on 36168  degrees of freedom
Residual deviance: 28827  on 36126  degrees of freedom
AIC: 28913
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3857  813
         1  664 3708

               Accuracy : 0.8367
                 95% CI : (0.8289, 0.8442)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6733

 Mcnemar's Test P-Value : 0.0001176

            Sensitivity : 0.8531
            Specificity : 0.8202
         Pos Pred Value : 0.8259
         Neg Pred Value : 0.8481
             Prevalence : 0.5000
         Detection Rate : 0.4266
   Detection Prevalence : 0.5165
      Balanced Accuracy : 0.8367

       'Positive' Class : 0
```

	metrics
Sensitivity	0.8531
Specificity	0.8202
Pos Pred Value	0.8259
Neg Pred Value	0.8481
Precision	0.8259
Recall	0.8531
F1	0.8393
Prevalence	0.5000
Detection Rate	0.4266
Detection Prevalence	0.5165
Balanced Accuracy	0.8367

DECISION TREE



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3377  475
         1 1144 4046

               Accuracy : 0.8209
                 95% CI : (0.8129, 0.8288)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6419

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7470
            Specificity : 0.8949
         Pos Pred Value : 0.8767
         Neg Pred Value : 0.7796
             Prevalence : 0.5000
         Detection Rate : 0.3735
   Detection Prevalence : 0.4260
      Balanced Accuracy : 0.8209

       'Positive' Class : 0
```

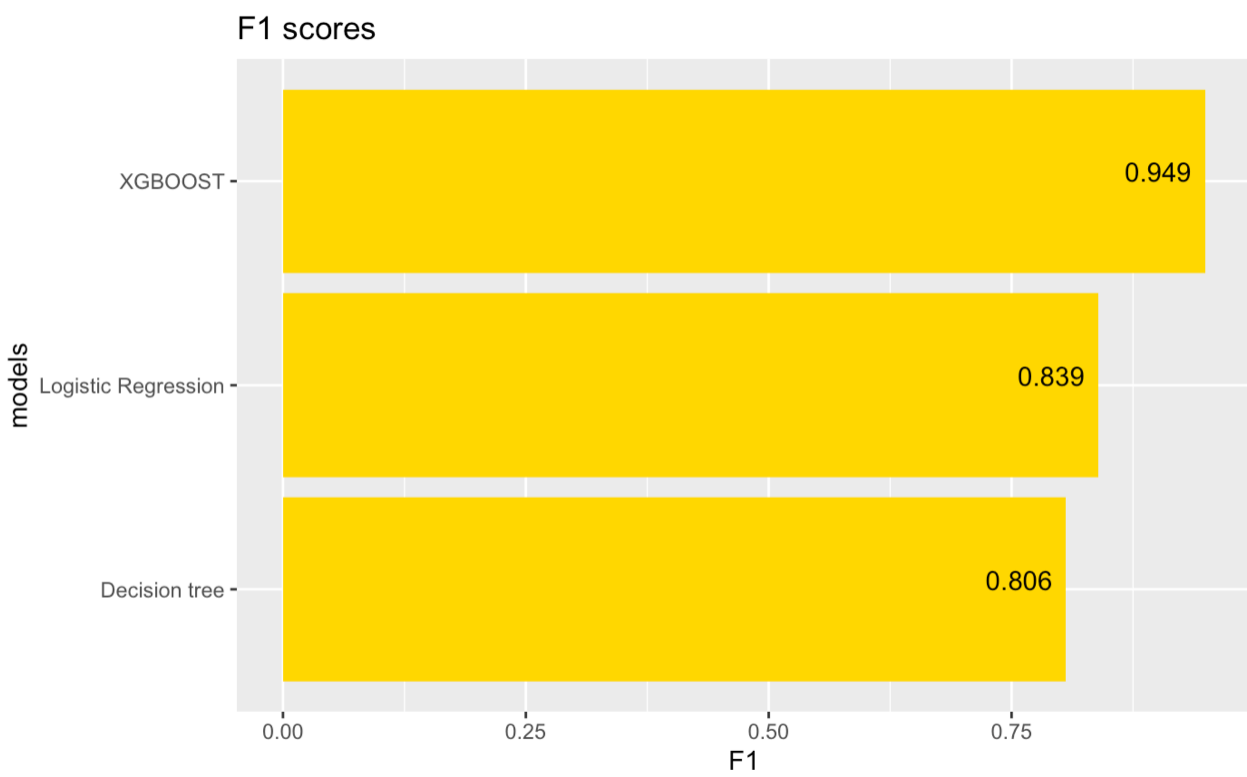
	metrics
Sensitivity	0.7470
Specificity	0.8949
Pos Pred Value	0.8767
Neg Pred Value	0.7796
Precision	0.8767
Recall	0.7470
F1	0.8066
Prevalence	0.5000
Detection Rate	0.3735
Detection Prevalence	0.4260
Balanced Accuracy	0.8209

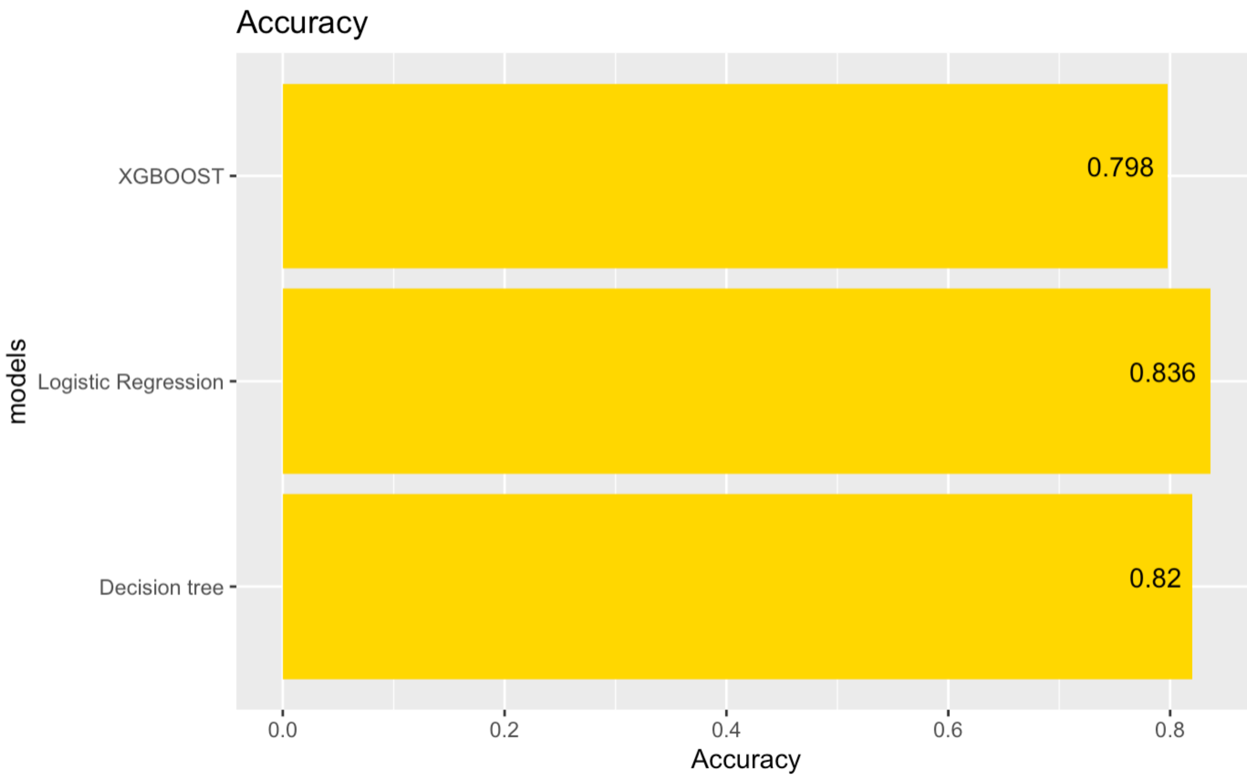
XGBOOST

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7784  200
         1  644  413

               Accuracy : 0.9066
                 95% CI : (0.9005, 0.9126)
    No Information Rate : 0.9322
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4472

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9236
            Specificity : 0.6737
         Pos Pred Value : 0.9749
         Neg Pred Value : 0.3907
             Prevalence : 0.9322
         Detection Rate : 0.8610
   Detection Prevalence : 0.8831
      Balanced Accuracy : 0.7987

       'Positive' Class : 0
```

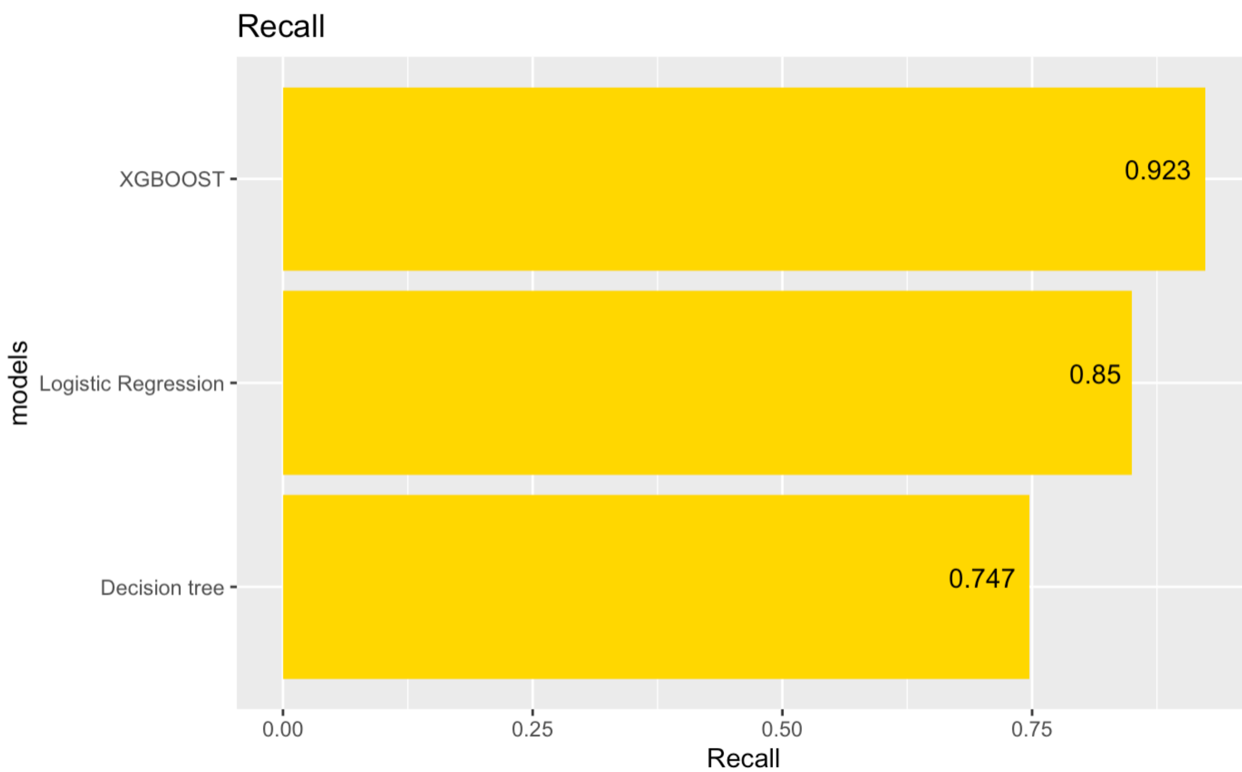
	metrics
Sensitivity	0.9236
Specificity	0.6737
Pos Pred Value	0.9749
Neg Pred Value	0.3907
Precision	0.9749
Recall	0.9236
F1	0.9486
Prevalence	0.9322
Detection Rate	0.8610
Detection Prevalence	0.8831
Balanced Accuracy	0.7987

Model Selection

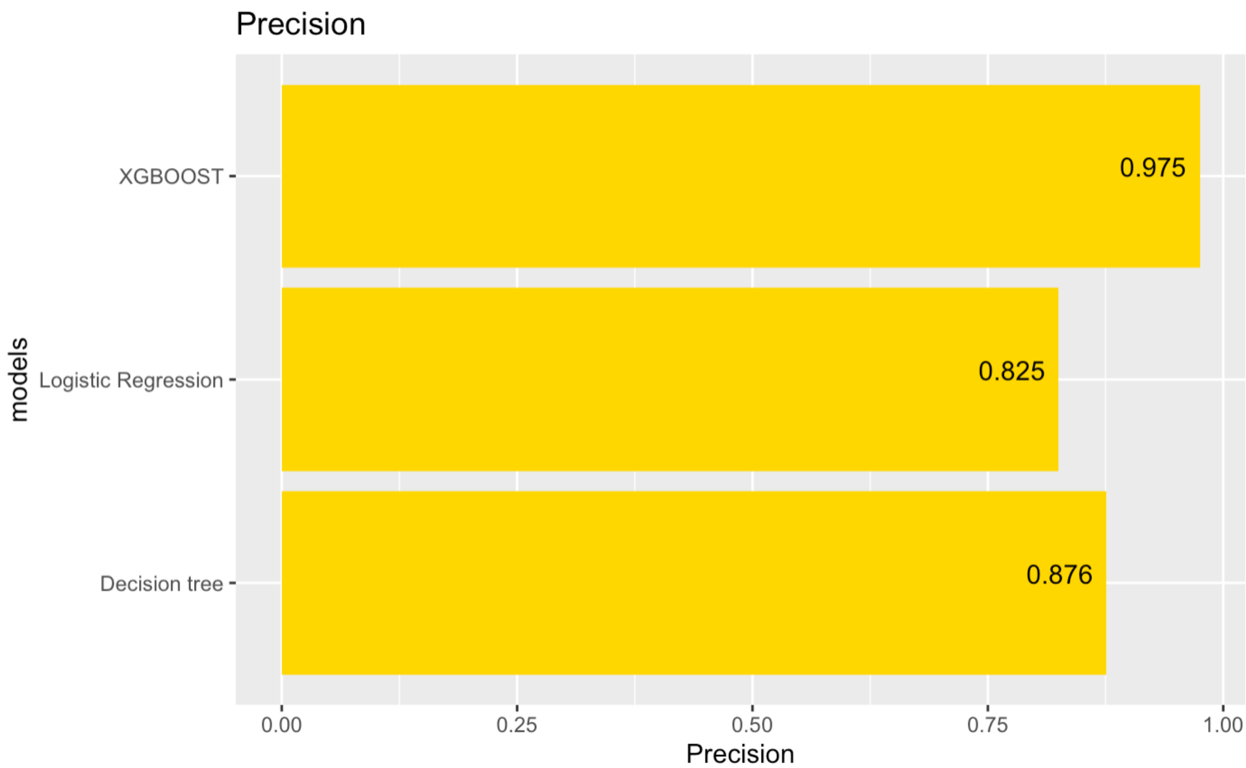




Accuracy
XGBOOST
Logistic Regression
Decision tree
models
0.798
0.836
0.82
0.0
0.2
0.4
0.6
0.8
Accuracy



Recall
XGBOOST
Logistic Regression
Decision tree
models
0.923
0.85
0.747
0.00
0.25
0.50
0.75
Recall



- As seen from the plots, F1 score, recall and precision of the model conducted with XGBOOST is the highest.
- Thus, F1 score, recall and precision suggest that XGBOOST model is better model among 3 models.