Bank Telemarketing

null

null

### **Data Loading and consolidation**

## 'data.frame': 41188 obs. of 21 variables:  
## $ AGE : int 46 44 51 28 37 34 46 42 36 54 ...  
## $ JOB : Factor w/ 12 levels "admin.","blue-collar",..: 5 8 1 8 2 10 2 2 1 2 ...  
## $ MARITAL : Factor w/ 4 levels "divorced","married",..: 2 2 3 2 2 2 2 2 2 2 ...  
## $ EDUCATION : Factor w/ 8 levels "basic.4y","basic.6y",..: 3 4 2 4 3 4 2 3 7 1 ...  
## $ DEFAULTCREDIT : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 1 1 2 1 1 2 ...  
## $ HOUSING : Factor w/ 3 levels "no","unknown",..: 1 3 1 3 1 1 3 3 3 1 ...  
## $ LOAN : Factor w/ 3 levels "no","unknown",..: 1 3 1 1 1 1 1 3 1 1 ...  
## $ CONTACT : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...  
## $ MONTH : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...  
## $ DAY\_OF\_WEEK : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ DURATION : int 128 107 303 81 270 228 240 673 233 102 ...  
## $ CAMPAIGN : int 2 1 2 1 1 1 1 2 3 1 ...  
## $ PDAYS : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ PREVIOUS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ POUTCOME : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ EMP\_VAR\_RATE : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ CONS\_PRICE\_IDX: num 94 94 94 94 94 ...  
## $ CONS\_CONF\_IDX : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ EURIBOR3M : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ NR\_EMPLOYED : num 5191 5191 5191 5191 5191 ...  
## $ Y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...

##### **Missing value and empty data imputation**

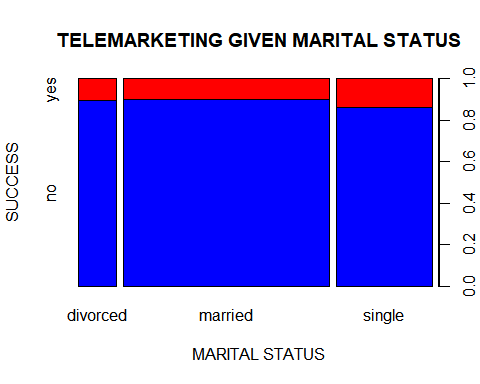
## AGE JOB MARITAL EDUCATION DEFAULTCREDIT   
## 0 0 0 0 0   
## HOUSING LOAN CONTACT MONTH DAY\_OF\_WEEK   
## 0 0 0 0 0   
## DURATION CAMPAIGN PDAYS PREVIOUS POUTCOME   
## 0 0 0 0 0   
## EMP\_VAR\_RATE CONS\_PRICE\_IDX CONS\_CONF\_IDX EURIBOR3M NR\_EMPLOYED   
## 0 0 0 0 0   
## Y   
## 0

## AGE JOB MARITAL EDUCATION DEFAULTCREDIT   
## 0 0 0 0 0   
## HOUSING LOAN CONTACT MONTH DAY\_OF\_WEEK   
## 0 0 0 0 0   
## DURATION CAMPAIGN PDAYS PREVIOUS POUTCOME   
## 0 0 0 0 0   
## EMP\_VAR\_RATE CONS\_PRICE\_IDX CONS\_CONF\_IDX EURIBOR3M NR\_EMPLOYED   
## 0 0 0 0 0   
## Y   
## 0

There are neither missing values nor empty data in our dataset

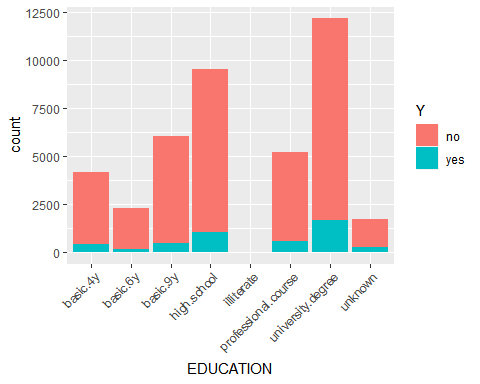
### **Exploratory analysis**

##### **Impact of Marital Status on Successful Telemarketing**

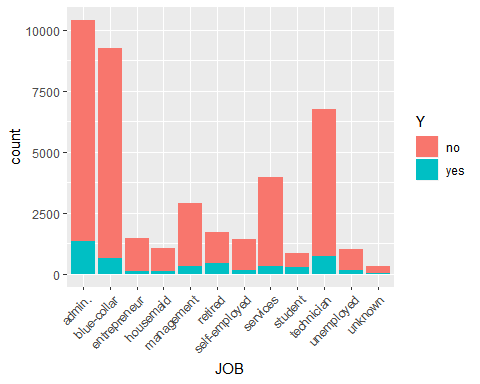


The proportion of unsuccessful telemarketing is almost similar for all categories.

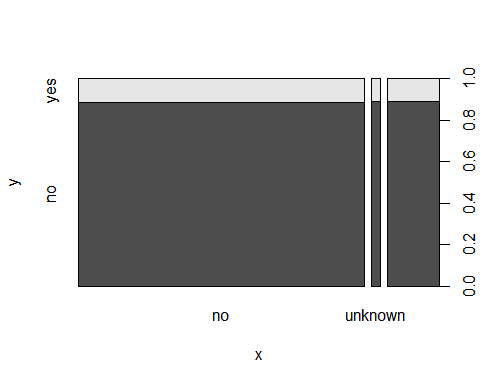
##### **Impact of Level of Education on telemarketing success**



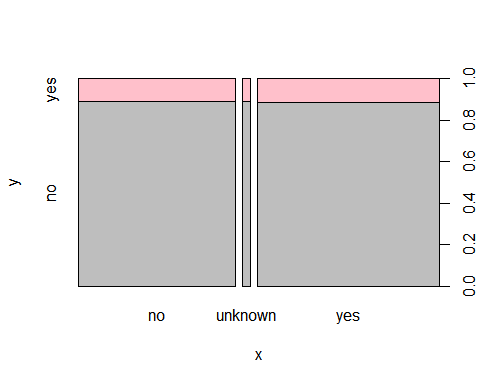
##### **Impact of Job type on telemarketing success**



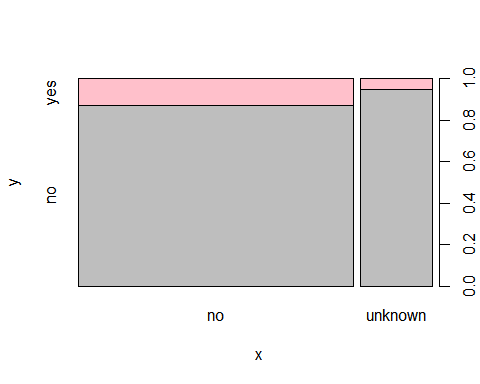
##### **Impact of Personal Loan on telemarketing success**



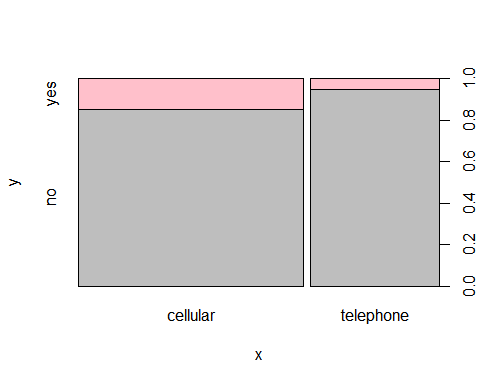
##### **Impact of Housing loan on telemarketing success**



##### **Impact of Credit history on telemarketing success**



##### **Impact of Communication type on telemarketing success**

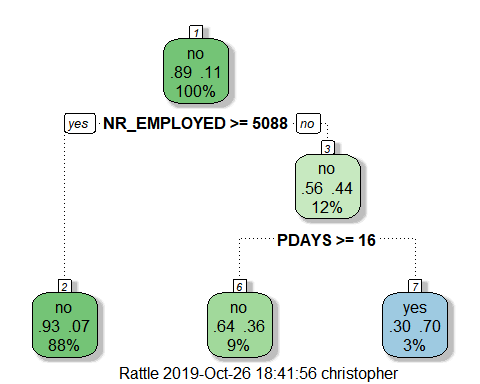


Partision the data into training and testing dataset

Data %<>% select(-DURATION)  
id <- createDataPartition(Data$Y,times = 1,p = 0.7,list = FALSE)  
train <- Data[id,]  
test <- Data[-id,]

#### **Benchmark decision tree**

create benchmark model

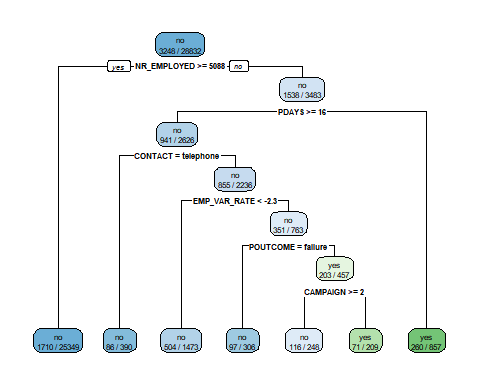


Our benchmark model is using only two features - nr\_employed(number of employees) and poutcome(previous outcome of marketing campaign) for prediction

Let’s predict using the benchmark model

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 25324 2651  
## yes 260 597  
##   
## Accuracy : 0.899   
## 95% CI : (0.8955, 0.9025)  
## No Information Rate : 0.8873   
## P-Value [Acc > NIR] : 9.791e-11   
##   
## Kappa : 0.2559   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9898   
## Specificity : 0.1838   
## Pos Pred Value : 0.9052   
## Neg Pred Value : 0.6966   
## Prevalence : 0.8873   
## Detection Rate : 0.8783   
## Detection Prevalence : 0.9703   
## Balanced Accuracy : 0.5868   
##   
## 'Positive' Class : no   
##

The accuracy of our benchmark model is 89.99%.

There is chance of overfitting in Single tree, So I will go for cross validation using n fold techinque 

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 25253 2513  
## yes 331 735  
##   
## Accuracy : 0.9014   
## 95% CI : (0.8979, 0.9048)  
## No Information Rate : 0.8873   
## P-Value [Acc > NIR] : 9.491e-15   
##   
## Kappa : 0.3019   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9871   
## Specificity : 0.2263   
## Pos Pred Value : 0.9095   
## Neg Pred Value : 0.6895   
## Prevalence : 0.8873   
## Detection Rate : 0.8759   
## Detection Prevalence : 0.9630   
## Balanced Accuracy : 0.6067   
##   
## 'Positive' Class : no   
##

Our new tree uses more features for prediction(5) which would suggest that it may be a better model than our initial benchmark model. This tree also has a slightly higher accuracy(0.9019) than our benchmark model.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 10845 1099  
## yes 119 293  
##   
## Accuracy : 0.9014   
## 95% CI : (0.896, 0.9066)  
## No Information Rate : 0.8873   
## P-Value [Acc > NIR] : 2.483e-07   
##   
## Kappa : 0.2882   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9891   
## Specificity : 0.2105   
## Pos Pred Value : 0.9080   
## Neg Pred Value : 0.7112   
## Prevalence : 0.8873   
## Detection Rate : 0.8777   
## Detection Prevalence : 0.9667   
## Balanced Accuracy : 0.5998   
##   
## 'Positive' Class : no   
##

Our model has a prediction accuracy of 89.95% on the new dataset