STA141A Final Project

Animay Sharma, Aditya Kallepalli, Charles Chien, Shaumik Pathak

12/14/2020

1. Introduction

1.1 Background

Here, we have a set of marketing data of a banking institution. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Among the four datasets provided, we chose to utilize the "bank-full.csv" in this report, with 17 different inputs.

As mentioned in article (Lopez, Customer segmentation using machine learning 2020) [1], data science and machine learning methods are helpful when it comes to helping companies with customer segmentation. Customer targeting is the process of analyzing customer features to select those customers who are more prone to a target product or service. By making intelligent use of data, companies could make a big difference to their competitors.

Advanced analytics plays a key role when it comes to selecting potentially profitable clients, which allows the design of more effective marketing campaigns. By using the four steps of advanced analytics: descriptive, diagnostic, predictive, and prescriptive, we would be able to answer key questions such as "what happened?", "why did it happen?", "what will happen?", and "how can we make it happen?"

In this report, we would be covering most of those steps. Our primary goal is to build a predictive model to answer a simple yes or no question: to determine whether a client will sign on to a long-term deposit. A model as such would allow banks to save on marketing expense on groups of customers that have a low chance of subscription, and focus on other customers that have a high chance of success. Overall, this would improve the profitability of banks and ultimately decrease marketing deficiencies.

While our main goal is to build a classification model and assist with bank marketing efforts, we would also like to conduct an exploratory data analysis (EDA) to explore relationships between different input variables. We would report any useful insights along the way, which covers both the "descriptive" and "diagnostic" parts of the four steps of advanced analytics as mentioned in the article.

[1] Lopez, R. (2020). Customer segmentation using machine learning. Retrieved December 12, 2020, from https://www.neuraldesigner.com/blog/customer_segmentation_using_advanced_analytics

1.2 Statistical Questions of Interest

To answer the primary scientific question of interest, we would fit our model in 2 different methods. The response will be a binary yes/no variable "has the client subscribed a term deposit?" All other variables provided will then be our input variables to allow us to build this model. Here, our 2 classification methods are

1. Logistic Regression
2. Random Forest

We would then use both backward and forward stepwise model selection using a likelihood ratio test (LRT) to conduct a heauristic model selection and prune down our model. We would also use cross validation (CV) to obtain more robust results.

1.) Setup

```
library(readr)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages ------------
```

```
## v ggplot2 3.3.2     v dplyr   1.0.2
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.1     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts ---------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fastDummies)
```

```
## Warning: package 'fastDummies' was built under R version 4.0.3
```

```
library(knitr)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(dplyr)
bank_full <- read_delim("datasets/bank-full.csv",
    ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   job = col_character(),
##   marital = col_character(),
##   education = col_character(),
##   default = col_character(),
##   balance = col_double(),
##   housing = col_character(),
##   loan = col_character(),
##   contact = col_character(),
##   day = col_double(),
##   month = col_character(),
##   duration = col_double(),
##   campaign = col_double(),
##   pdays = col_double(),
##   previous = col_double(),
##   poutcome = col_character(),
##   y = col_character()
## )
```

```
#View(bank_full)
bank.data = bank_full
head(bank.data)
```

```
## # A tibble: 6 x 17
##     age job   marital education default balance housing loan  contact   day
##   <dbl> <chr> <chr>   <chr>     <chr>     <dbl> <chr>   <chr> <chr>   <dbl>
## 1    58 mana~ married tertiary  no         2143 yes     no    unknown     5
## 2    44 tech~ single  secondary no           29 yes     no    unknown     5
## 3    33 entr~ married secondary no            2 yes     yes   unknown     5
## 4    47 blue~ married unknown   no         1506 yes     no    unknown     5
## 5    33 unkn~ single  unknown   no            1 no      no    unknown     5
## 6    35 mana~ married tertiary  no          231 yes     no    unknown     5
## # ... with 7 more variables: month <chr>, duration <dbl>, campaign <dbl>,
## #   pdays <dbl>, previous <dbl>, poutcome <chr>, y <chr>
```

```
#Binary
housing.binary = ifelse(bank.data$housing=='yes',1,0)
```

modifying data set

```
cat_data = data.frame(bank.data$job,bank.data$marital,bank.data$education)
bin_cat_data = dummy_cols(cat_data)
bin_cat_data = bin_cat_data %>% select(4:22)
yesno_data = data.frame(bank.data$default,bank.data$housing,bank.data$loan,bank.data$y)


yesno_data$bank.data.default <- revalue(yesno_data$bank.data.default, c("yes"=1))
yesno_data$bank.data.default <- revalue(yesno_data$bank.data.default, c("no"=0))
yesno_data$bank.data.housing <- revalue(yesno_data$bank.data.housing, c("yes"=1))
yesno_data$bank.data.housing <- revalue(yesno_data$bank.data.housing, c("no"=0))
yesno_data$bank.data.loan <- revalue(yesno_data$bank.data.loan, c("yes"=1))
yesno_data$bank.data.loan <- revalue(yesno_data$bank.data.loan, c("no"=0))
yesno_data$bank.data.y <- revalue(yesno_data$bank.data.y, c("yes"=1))
yesno_data$bank.data.y <- revalue(yesno_data$bank.data.y, c("no"=0))
remaining_data = bank.data %>% select(1,6,9,10,11,12,13,14,15,16)
master_bin_data = cbind(bin_cat_data,yesno_data,remaining_data)

head(master_bin_data)
```

```
##   bank.data.job_admin. bank.data.job_blue-collar bank.data.job_entrepreneur
## 1                    0                         0                          0
## 2                    0                         0                          0
## 3                    0                         0                          1
## 4                    0                         1                          0
## 5                    0                         0                          0
## 6                    0                         0                          0
##   bank.data.job_housemaid bank.data.job_management bank.data.job_retired
## 1                       0                        1                     0
## 2                       0                        0                     0
## 3                       0                        0                     0
## 4                       0                        0                     0
## 5                       0                        0                     0
## 6                       0                        1                     0
##   bank.data.job_self-employed bank.data.job_services bank.data.job_student
## 1                           0                       0                     0
## 2                           0                       0                     0
## 3                           0                       0                     0
## 4                           0                       0                     0
## 5                           0                       0                     0
## 6                           0                       0                     0
##   bank.data.job_technician bank.data.job_unemployed bank.data.job_unknown
## 1                        0                        0                    0
## 2                        1                        0                    0
## 3                        0                        0                    0
## 4                        0                        0                    0
## 5                        0                        0                    1
## 6                        0                        0                    0
##   bank.data.marital_divorced bank.data.marital_married bank.data.marital_single
## 1                          0                        1                       0
## 2                          0                        0                       1
## 3                          0                        1                       0
## 4                          0                        1                       0
## 5                          0                        0                       1
```

```
## 6                          0                          1                          0
##   bank.data.education_primary bank.data.education_secondary
## 1                           0                             0
## 2                           0                             1
## 3                           0                             1
## 4                           0                             0
## 5                           0                             0
## 6                           0                             0
##   bank.data.education_tertiary bank.data.education_unknown bank.data.default
## 1                            1                           0                 0
## 2                            0                           0                 0
## 3                            0                           0                 0
## 4                            0                           1                 0
## 5                            0                           1                 0
## 6                            1                           0                 0
##   bank.data.housing bank.data.loan bank.data.y age balance contact day month
## 1                 1              0           0  58    2143 unknown   5   may
## 2                 1              0           0  44      29 unknown   5   may
## 3                 1              1           0  33       2 unknown   5   may
## 4                 1              0           0  47    1506 unknown   5   may
## 5                 0              0           0  33       1 unknown   5   may
## 6                 1              0           0  35     231 unknown   5   may
##   duration campaign pdays previous poutcome
## 1      261        1    -1        0  unknown
## 2      151        1    -1        0  unknown
## 3       76        1    -1        0  unknown
## 4       92        1    -1        0  unknown
## 5      198        1    -1        0  unknown
## 6      139        1    -1        0  unknown
```

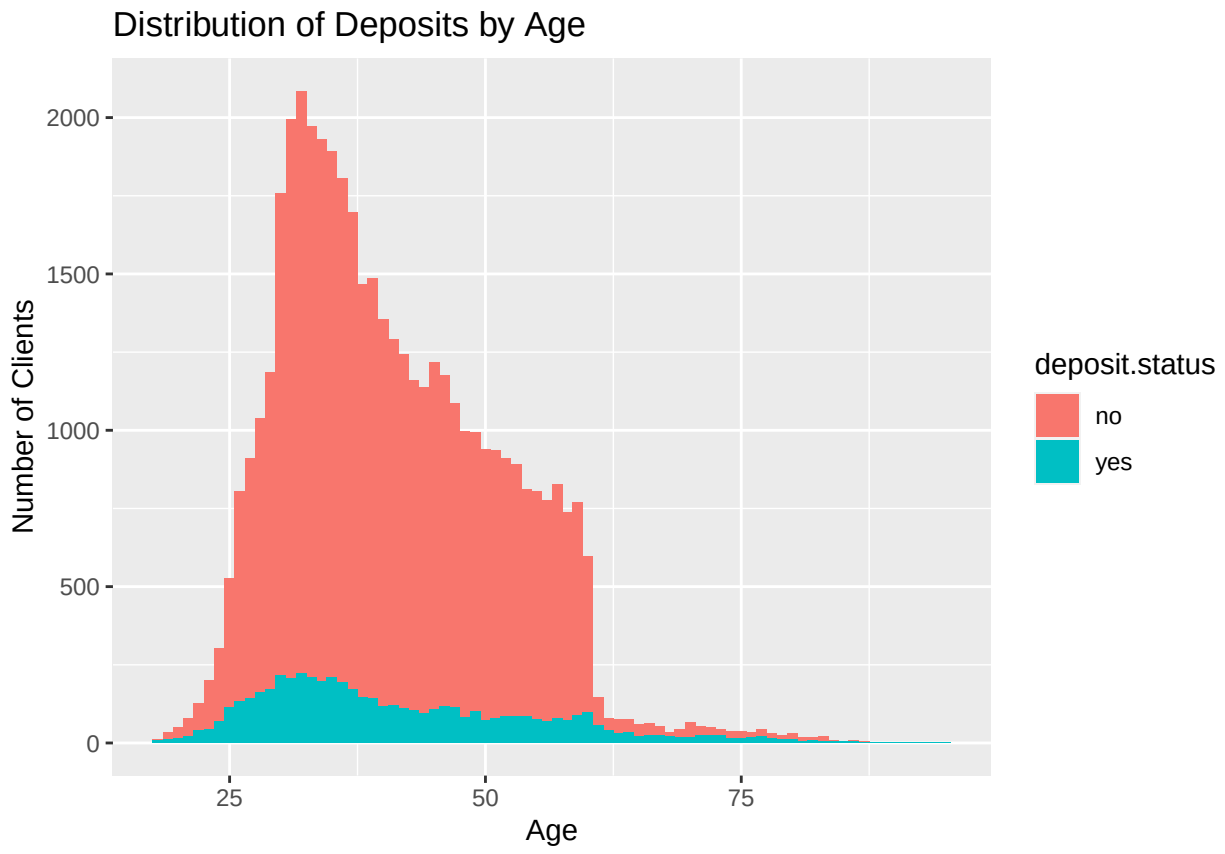
2.) Exploratory Categorical Data Analysis:

```
summary(bank.data)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.94
##  3rd Qu.:48.00
##  Max.   :95.00
##    default            balance         housing             loan
##  Length:45211       Min.   : -8019   Length:45211       Length:45211
##  Class :character   1st Qu.:    72   Class :character   Class :character
##  Mode  :character   Median :   448   Mode  :character   Mode  :character
##                     Mean   :  1362
##                     3rd Qu.:  1428
##                     Max.   :102127
##    contact               day           month              duration
##  Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0
##  Mode  :character   Median :16.00   Mode  :character   Median : 180.0
```

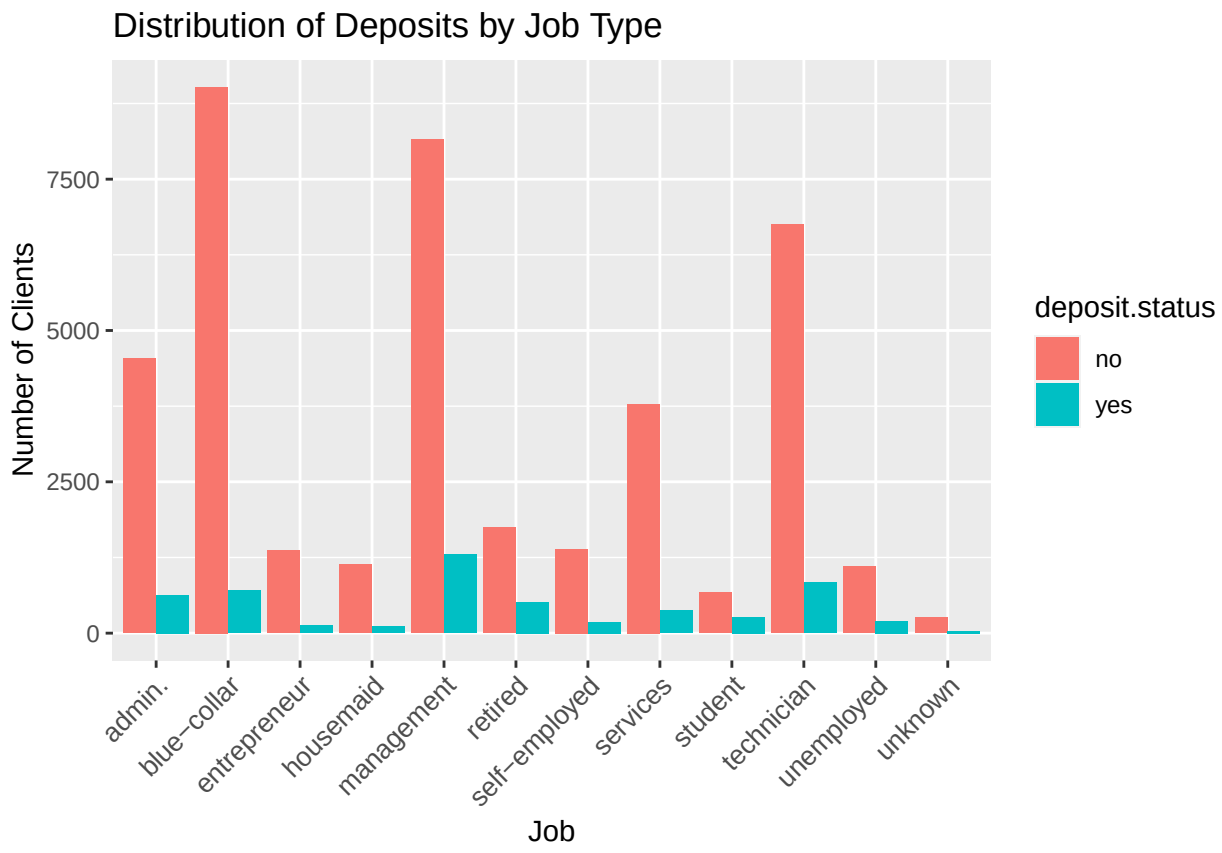
```
##                    Mean   :15.81                      Mean   : 258.2
##                    3rd Qu.:21.00                      3rd Qu.: 319.0
##                    Max.   :31.00                      Max.   :4918.0
##     campaign          pdays          previous          poutcome
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   Length:45211
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   Class :character
##  Median : 2.000   Median : -1.0   Median :  0.0000   Mode  :character
##  Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
##  3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :63.000   Max.   :871.0   Max.   :275.0000
##       y
##  Length:45211
##  Class :character
##  Mode  :character
##
##
##
```

```
deposit.status = bank.data$y
#Age
ggplot(bank.data,aes(x=bank.data$age,fill=deposit.status)) + geom_histogram(binwidth=1) +
  labs(y= "Number of Clients", x="Age", title = "Distribution of Deposits by Age")
```

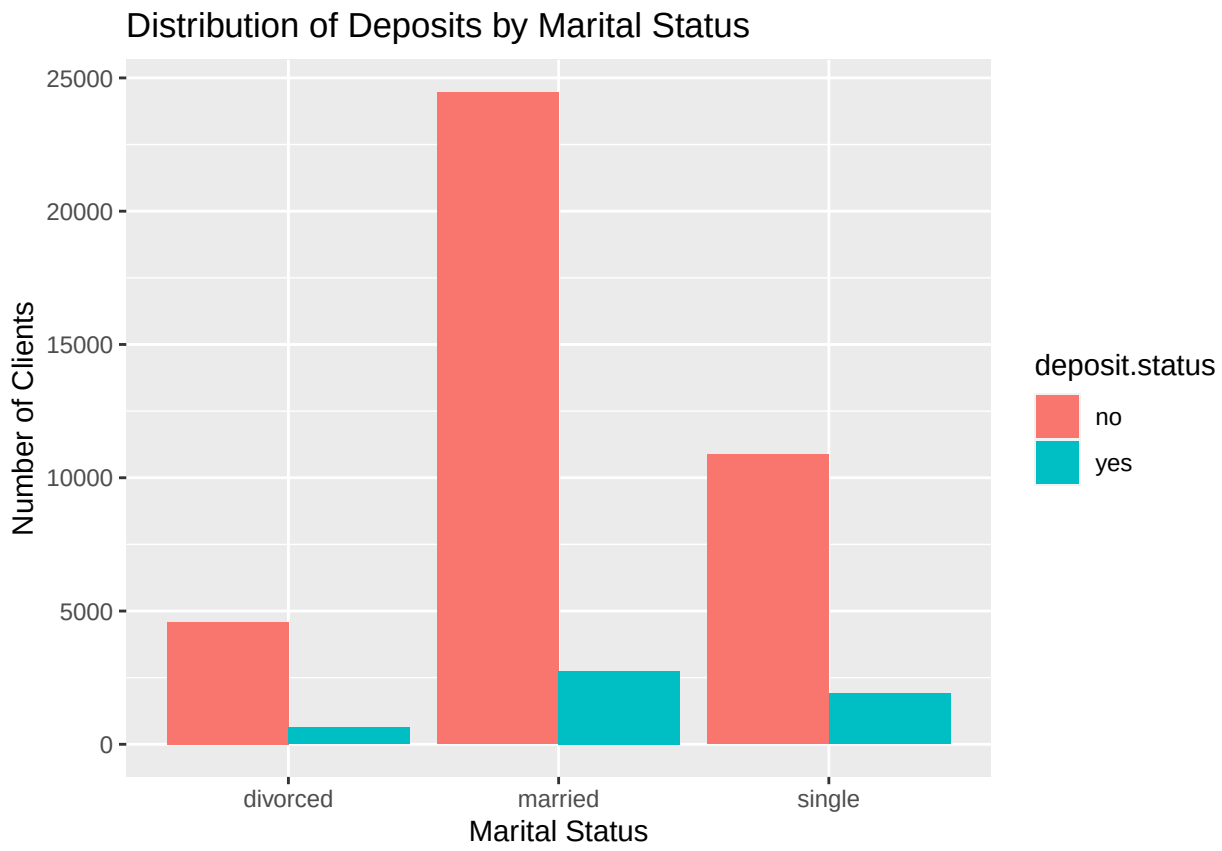



```
age.desc = bank.data %>% group_by(y) %>% summarise(age.mean = mean(age), .groups = 'drop')
```

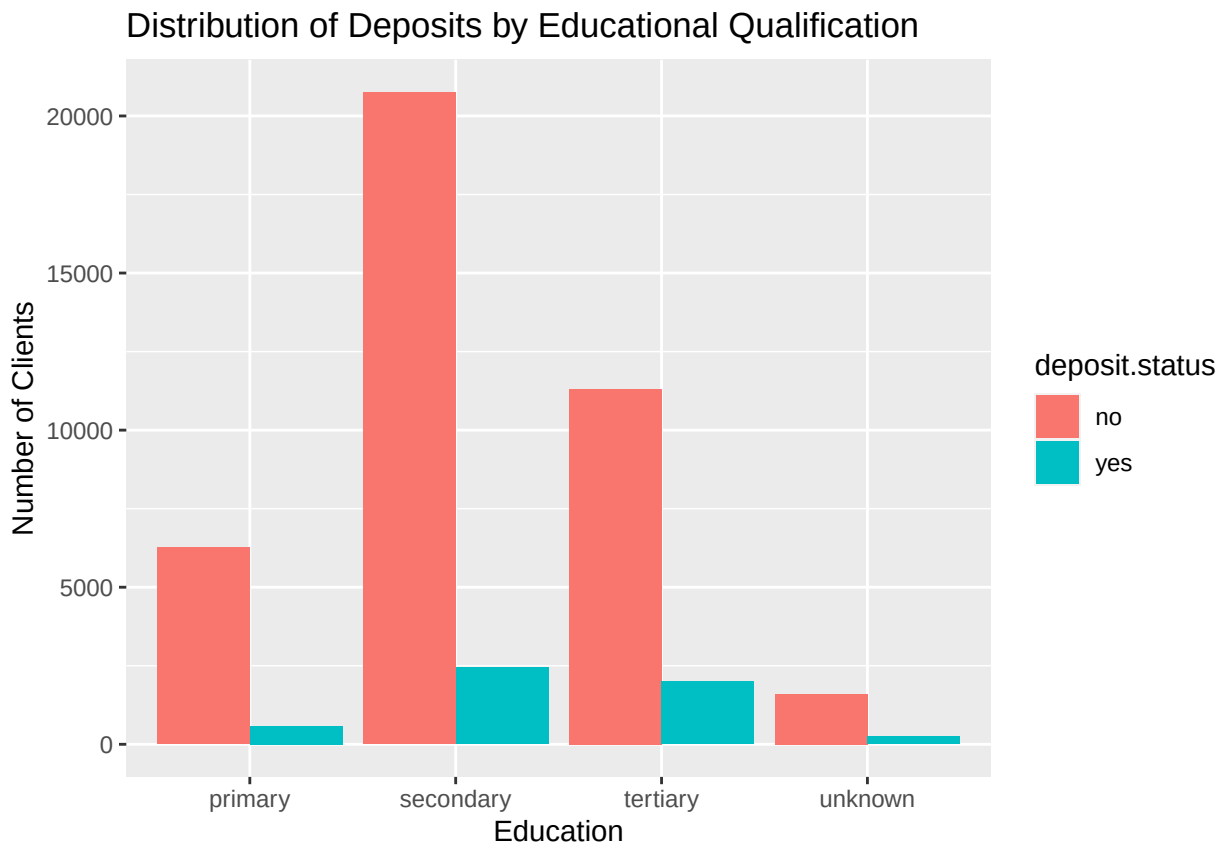
```
#Job
ggplot(bank.data, aes(x=bank.data$job,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Job", title = "Distribution of Deposits by Job Type")+
  theme(axis.text.x = element_text(size = 10, angle = 45, hjust=1, vjust=1))
```



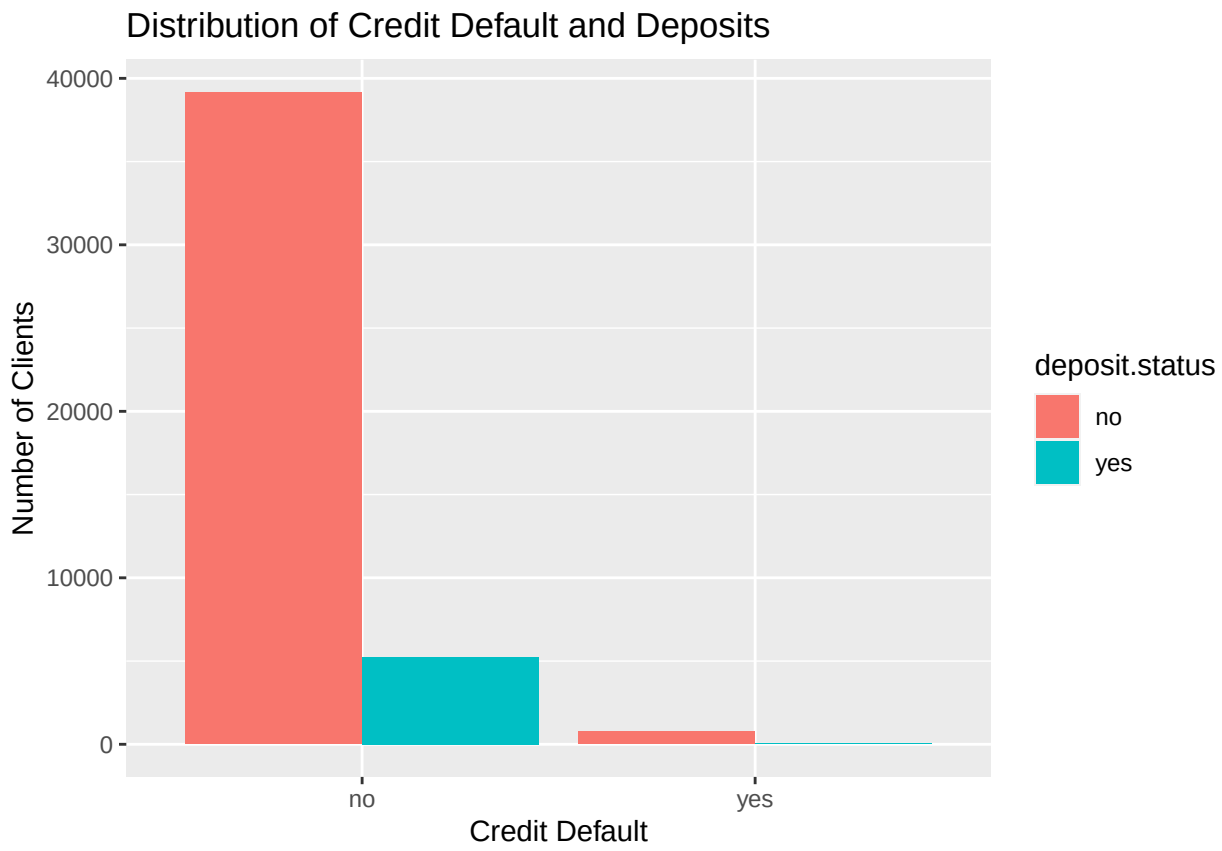

```
#Marital Status
ggplot(bank.data, aes(x=bank.data$marital,fill=deposit.status)) + geom_bar(position = position_dodge())
  labs(y= "Number of Clients", x="Marital Status", title = "Distribution of Deposits by Marital Status"
```




```
#Education
ggplot(bank.data, aes(x=bank.data$education,fill=deposit.status)) + geom_bar(position = position_dodge()
  labs(y= "Number of Clients", x="Education", title = "Distribution of Deposits by Educational Qualifica
```

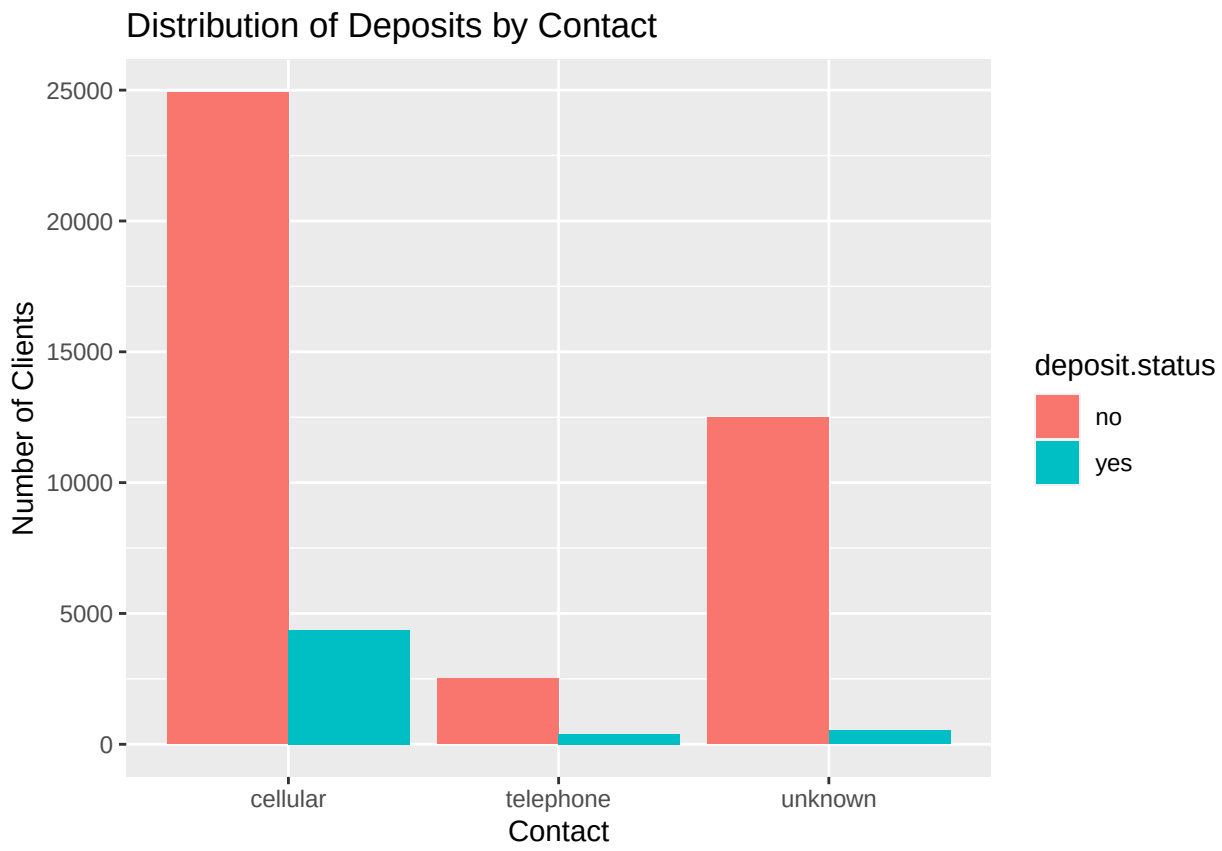



```
#Credit Default
ggplot(bank.data, aes(x=bank.data$default,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Credit Default", title = "Distribution of Credit Default and Deposits
```



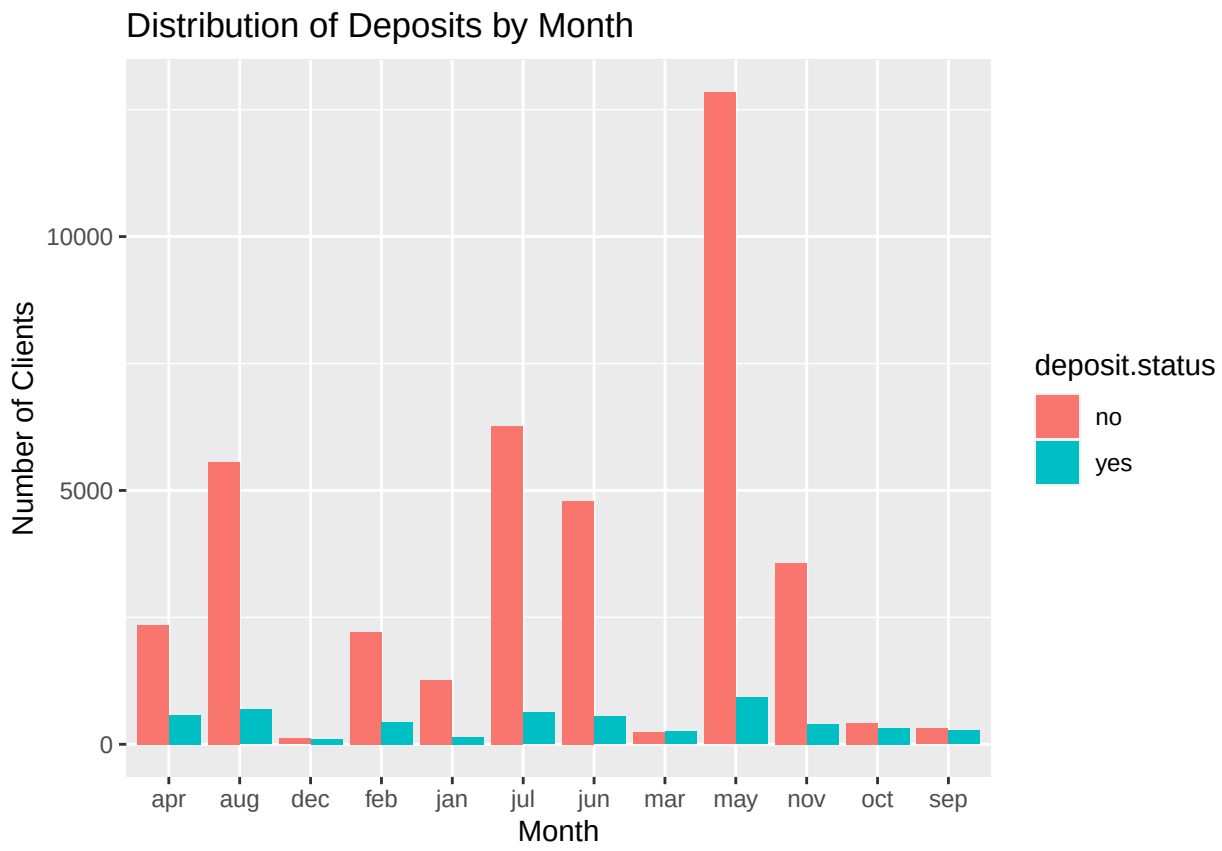

```
#Housing Loan
ggplot(bank.data, aes(x=bank.data$housing,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Contact", title = "Distribution of Client Having a Housing Loan and D
```


```
#Contact
ggplot(bank.data, aes(x=bank.data$contact,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Contact", title = "Distribution of Deposits by Contact")
```

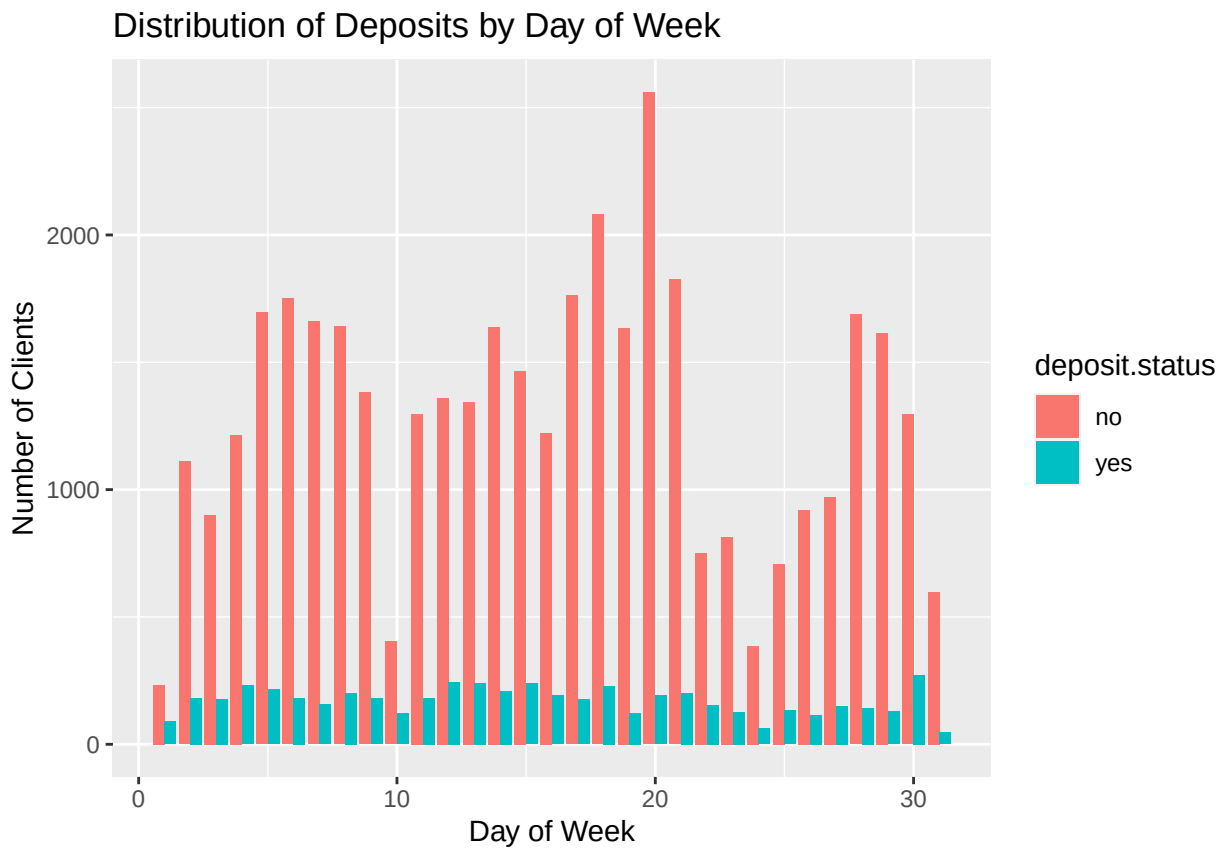



```
#Loans
ggplot(bank.data, aes(x=bank.data$loan,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Loan", title = "Distribution of Clients with Loans and Deposit")
```

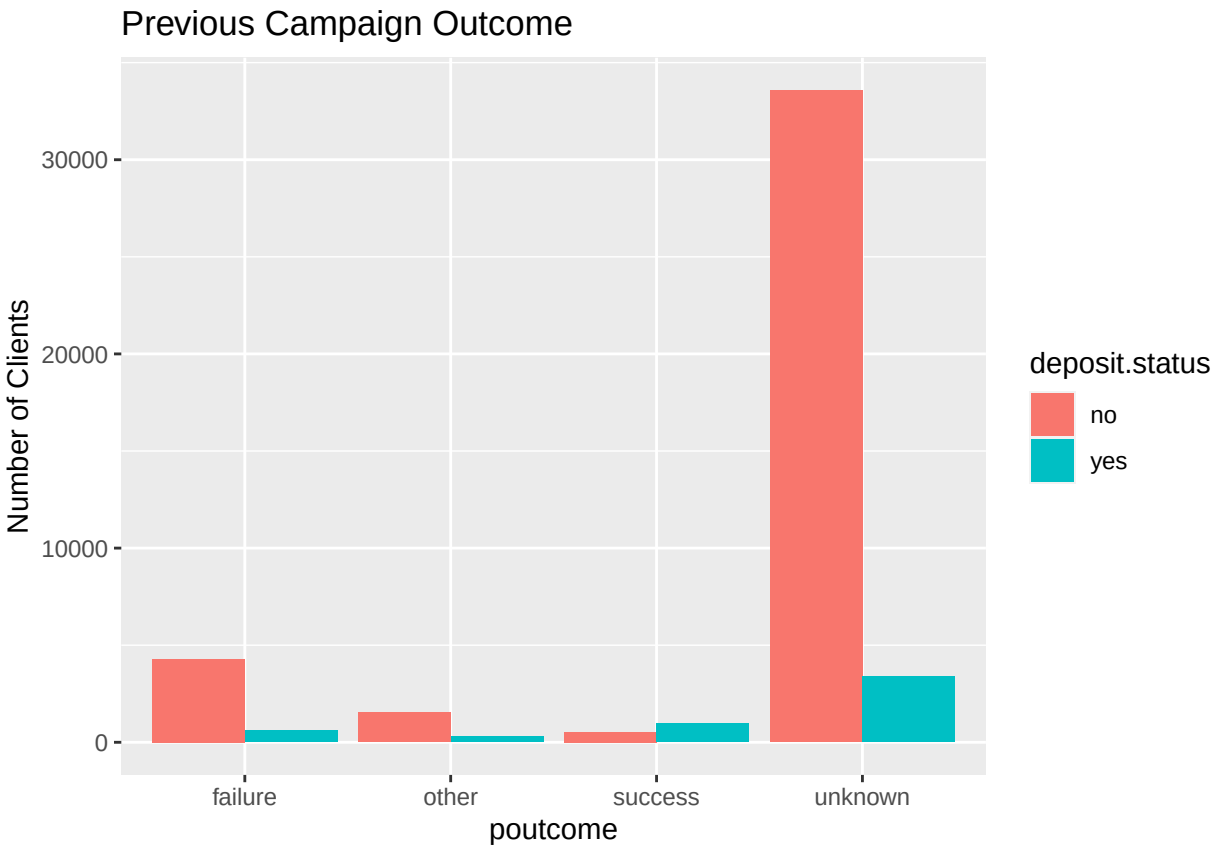

```
#month
ggplot(bank.data, aes(x=bank.data$month,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Month", title = "Distribution of Deposits by Month")
```




```
#Day
ggplot(bank.data, aes(x=bank.data$day,fill=deposit.status)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Clients", x="Day of Week", title = "Distribution of Deposits by Day of Week")
```




```
#
ggplot(bank.data, aes(x=bank.data$poutcome,fill=deposit.status)) + geom_bar(position = position_dodge()
  labs(y= "Number of Clients", x="poutcome", title = "Previous Campaign Outcome")
```



- Add interpretations One interesting observation we can make is that the percentage of yes for those without housing loans is noticeably greater than for those with a housing loan

yes percentages