Bank Telemarketing Dataset

## Read and explore the dataset

bankDf <- read.csv2("bank-additional-full.csv")  
summary(bankDf)

## age job marital education   
## Min. :17.00 Length:41188 Length:41188 Length:41188   
## 1st Qu.:32.00 Class :character Class :character Class :character   
## Median :38.00 Mode :character Mode :character Mode :character   
## Mean :40.02   
## 3rd Qu.:47.00   
## Max. :98.00   
## default housing loan contact   
## Length:41188 Length:41188 Length:41188 Length:41188   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## month day\_of\_week duration campaign   
## Length:41188 Length:41188 Min. : 0.0 Min. : 1.000   
## Class :character Class :character 1st Qu.: 102.0 1st Qu.: 1.000   
## Mode :character Mode :character Median : 180.0 Median : 2.000   
## Mean : 258.3 Mean : 2.568   
## 3rd Qu.: 319.0 3rd Qu.: 3.000   
## Max. :4918.0 Max. :56.000   
## pdays previous poutcome emp.var.rate   
## Min. : 0.0 Min. :0.000 Length:41188 Length:41188   
## 1st Qu.:999.0 1st Qu.:0.000 Class :character Class :character   
## Median :999.0 Median :0.000 Mode :character Mode :character   
## Mean :962.5 Mean :0.173   
## 3rd Qu.:999.0 3rd Qu.:0.000   
## Max. :999.0 Max. :7.000   
## cons.price.idx cons.conf.idx euribor3m nr.employed   
## Length:41188 Length:41188 Length:41188 Length:41188   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## y   
## Length:41188   
## Class :character   
## Mode :character   
##   
##   
##

glimpse(bankDf)

## Rows: 41,188  
## Columns: 21  
## $ age <int> 56, 57, 37, 40, 56, 45, 59, 41, 24, 25, 41, 25, 29, 57,…  
## $ job <chr> "housemaid", "services", "services", "admin.", "service…  
## $ marital <chr> "married", "married", "married", "married", "married", …  
## $ education <chr> "basic.4y", "high.school", "high.school", "basic.6y", "…  
## $ default <chr> "no", "unknown", "no", "no", "no", "unknown", "no", "un…  
## $ housing <chr> "no", "no", "yes", "no", "no", "no", "no", "no", "yes",…  
## $ loan <chr> "no", "no", "no", "no", "yes", "no", "no", "no", "no", …  
## $ contact <chr> "telephone", "telephone", "telephone", "telephone", "te…  
## $ month <chr> "may", "may", "may", "may", "may", "may", "may", "may",…  
## $ day\_of\_week <chr> "mon", "mon", "mon", "mon", "mon", "mon", "mon", "mon",…  
## $ duration <int> 261, 149, 226, 151, 307, 198, 139, 217, 380, 50, 55, 22…  
## $ campaign <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
## $ pdays <int> 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, …  
## $ previous <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
## $ poutcome <chr> "nonexistent", "nonexistent", "nonexistent", "nonexiste…  
## $ emp.var.rate <chr> "1.1", "1.1", "1.1", "1.1", "1.1", "1.1", "1.1", "1.1",…  
## $ cons.price.idx <chr> "93.994", "93.994", "93.994", "93.994", "93.994", "93.9…  
## $ cons.conf.idx <chr> "-36.4", "-36.4", "-36.4", "-36.4", "-36.4", "-36.4", "…  
## $ euribor3m <chr> "4.857", "4.857", "4.857", "4.857", "4.857", "4.857", "…  
## $ nr.employed <chr> "5191", "5191", "5191", "5191", "5191", "5191", "5191",…  
## $ y <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "…

## Data Wrangling

### Turn all “unknown” to NA value

bankDf <- bankDf %>% replace\_with\_na\_all(condition = ~.x == "unknown")

### Turn Bernoulli variables into 0 and 1 categories

bankDf <- rename(bankDf, telephone = contact)  
bankDf <- rename(bankDf, deposit = y)  
  
bankDf <- bankDf %>%  
 #mutate(employed = ifelse(employed == "unemployed", 0, 1)) %>%  
 mutate(default = ifelse(default == "yes", 1, 0)) %>%  
 mutate(housing = ifelse(housing == "yes", 1, 0)) %>%  
 mutate(loan = ifelse(loan == "yes", 1, 0)) %>%  
 mutate(telephone = ifelse(telephone == "telephone", 1, 0)) %>%  
 mutate(deposit = ifelse(deposit == "yes", 1, 0))

### Turn other variables into its suitable types of variables

#bankDf$employed <- as.factor(bankDf$employed)  
bankDf$job <- as.factor(bankDf$job)  
bankDf$marital <- as.factor(bankDf$marital)  
bankDf$education <- as.factor(bankDf$education)  
bankDf$default <- as.factor(bankDf$default)  
bankDf$housing <- as.factor(bankDf$housing)  
bankDf$loan <- as.factor(bankDf$loan)  
bankDf$telephone <- as.factor(bankDf$telephone)  
bankDf$poutcome <- as.factor(bankDf$poutcome)  
bankDf$month <- as.factor(bankDf$month)  
bankDf$day\_of\_week <- as.factor(bankDf$day\_of\_week)  
  
bankDf$previous <- as.numeric(bankDf$previous)  
bankDf$emp.var.rate <- as.numeric(bankDf$emp.var.rate)  
bankDf$cons.price.idx <- as.numeric(bankDf$cons.price.idx)  
bankDf$cons.conf.idx <- as.numeric(bankDf$cons.conf.idx)  
bankDf$euribor3m <- as.numeric(bankDf$euribor3m)  
bankDf$nr.employed <- as.numeric(bankDf$nr.employed)  
  
bankDf$pdays <- as.factor(bankDf$pdays)

## Data Exploration

summary(bankDf)

## age job marital   
## Min. :17.00 admin. :10422 divorced: 4612   
## 1st Qu.:32.00 blue-collar: 9254 married :24928   
## Median :38.00 technician : 6743 single :11568   
## Mean :40.02 services : 3969 NA's : 80   
## 3rd Qu.:47.00 management : 2924   
## Max. :98.00 (Other) : 7546   
## NA's : 330   
## education default housing loan telephone  
## university.degree :12168 0 :32588 0 :18622 0 :33950 0:26144   
## high.school : 9515 1 : 3 1 :21576 1 : 6248 1:15044   
## basic.9y : 6045 NA's: 8597 NA's: 990 NA's: 990   
## professional.course: 5243   
## basic.4y : 4176   
## (Other) : 2310   
## NA's : 1731   
## month day\_of\_week duration campaign pdays   
## may :13769 fri:7827 Min. : 0.0 Min. : 1.000 999 :39673   
## jul : 7174 mon:8514 1st Qu.: 102.0 1st Qu.: 1.000 3 : 439   
## aug : 6178 thu:8623 Median : 180.0 Median : 2.000 6 : 412   
## jun : 5318 tue:8090 Mean : 258.3 Mean : 2.568 4 : 118   
## nov : 4101 wed:8134 3rd Qu.: 319.0 3rd Qu.: 3.000 9 : 64   
## apr : 2632 Max. :4918.0 Max. :56.000 2 : 61   
## (Other): 2016 (Other): 421   
## previous poutcome emp.var.rate cons.price.idx   
## Min. :0.000 failure : 4252 Min. :-3.40000 Min. :92.20   
## 1st Qu.:0.000 nonexistent:35563 1st Qu.:-1.80000 1st Qu.:93.08   
## Median :0.000 success : 1373 Median : 1.10000 Median :93.75   
## Mean :0.173 Mean : 0.08189 Mean :93.58   
## 3rd Qu.:0.000 3rd Qu.: 1.40000 3rd Qu.:93.99   
## Max. :7.000 Max. : 1.40000 Max. :94.77   
##   
## cons.conf.idx euribor3m nr.employed deposit   
## Min. :-50.8 Min. :0.634 Min. :4964 Min. :0.0000   
## 1st Qu.:-42.7 1st Qu.:1.344 1st Qu.:5099 1st Qu.:0.0000   
## Median :-41.8 Median :4.857 Median :5191 Median :0.0000   
## Mean :-40.5 Mean :3.621 Mean :5167 Mean :0.1127   
## 3rd Qu.:-36.4 3rd Qu.:4.961 3rd Qu.:5228 3rd Qu.:0.0000   
## Max. :-26.9 Max. :5.045 Max. :5228 Max. :1.0000   
##

### Make some decisions

We decide to remove those columns below: - default because it has too many NAs - poutcome (previous campaign’s outcome) because it has too many “nonexistent” result - pdays (days after the customers were last contacted in the previous campaigns) because most of the customers had never been contacted before

bankDf <- bankDf %>%  
 select(!default) %>%  
 select(!poutcome) %>%  
 select(!pdays)  
  
bankDf <- na.omit(bankDf)  
summary(bankDf)

## age job marital   
## Min. :17.00 admin. :9937 divorced: 4302   
## 1st Qu.:32.00 blue-collar:8560 married :23183   
## Median :38.00 technician :6380 single :10760   
## Mean :39.86 services :3716   
## 3rd Qu.:47.00 management :2728   
## Max. :98.00 retired :1577   
## (Other) :5347   
## education housing loan telephone month   
## basic.4y : 4002 0:17667 0:32286 0:24441 may :12794   
## basic.6y : 2204 1:20578 1: 5959 1:13804 jul : 6630   
## basic.9y : 5856 aug : 5822   
## high.school : 9244 jun : 4846   
## illiterate : 18 nov : 3898   
## professional.course: 5100 apr : 2436   
## university.degree :11821 (Other): 1819   
## day\_of\_week duration campaign previous   
## fri:7224 Min. : 0.0 Min. : 1.000 Min. :0.00   
## mon:7927 1st Qu.: 102.0 1st Qu.: 1.000 1st Qu.:0.00   
## thu:8011 Median : 180.0 Median : 2.000 Median :0.00   
## tue:7481 Mean : 258.2 Mean : 2.567 Mean :0.17   
## wed:7602 3rd Qu.: 319.0 3rd Qu.: 3.000 3rd Qu.:0.00   
## Max. :4918.0 Max. :43.000 Max. :7.00   
##   
## 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.:1.344   
## Median : 1.10000 Median :93.44 Median :-41.80 Median :4.857   
## Mean : 0.08286 Mean :93.57 Mean :-40.54 Mean :3.623   
## 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 deposit   
## Min. :4964 Min. :0.0000   
## 1st Qu.:5099 1st Qu.:0.0000   
## Median :5191 Median :0.0000   
## Mean :5167 Mean :0.1113   
## 3rd Qu.:5228 3rd Qu.:0.0000   
## Max. :5228 Max. :1.0000   
##

### Split into traning and test dataset

set.seed(1)  
N <- seq(38245)  
S <- sample(N,30596)  
bankTrain <- bankDf[S,]  
bankTest <- bankDf[-S,]  
summary(bankTrain)

## age job marital education   
## Min. :17.00 admin. :7937 divorced: 3427 basic.4y :3184   
## 1st Qu.:32.00 blue-collar:6856 married :18600 basic.6y :1786   
## Median :38.00 technician :5117 single : 8569 basic.9y :4696   
## Mean :39.86 services :3001 high.school :7392   
## 3rd Qu.:47.00 management :2174 illiterate : 14   
## Max. :98.00 retired :1249 professional.course:4070   
## (Other) :4262 university.degree :9454   
## housing loan telephone month day\_of\_week duration   
## 0:14180 0:25811 0:19517 may :10233 fri:5792 Min. : 0.0   
## 1:16416 1: 4785 1:11079 jul : 5316 mon:6357 1st Qu.: 102.0   
## aug : 4673 thu:6447 Median : 180.0   
## jun : 3876 tue:5954 Mean : 257.6   
## nov : 3113 wed:6046 3rd Qu.: 320.0   
## apr : 1927 Max. :4199.0   
## (Other): 1458   
## campaign previous emp.var.rate cons.price.idx   
## Min. : 1.000 Min. :0.0000 Min. :-3.40000 Min. :92.20   
## 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:-1.80000 1st Qu.:93.08   
## Median : 2.000 Median :0.0000 Median : 1.10000 Median :93.44   
## Mean : 2.556 Mean :0.1704 Mean : 0.08556 Mean :93.57   
## 3rd Qu.: 3.000 3rd Qu.:0.0000 3rd Qu.: 1.40000 3rd Qu.:93.99   
## Max. :43.000 Max. :6.0000 Max. : 1.40000 Max. :94.77   
##   
## cons.conf.idx euribor3m nr.employed deposit   
## Min. :-50.80 Min. :0.634 Min. :4964 Min. :0.0000   
## 1st Qu.:-42.70 1st Qu.:1.344 1st Qu.:5099 1st Qu.:0.0000   
## Median :-41.80 Median :4.857 Median :5191 Median :0.0000   
## Mean :-40.53 Mean :3.625 Mean :5167 Mean :0.1105   
## 3rd Qu.:-36.40 3rd Qu.:4.961 3rd Qu.:5228 3rd Qu.:0.0000   
## Max. :-26.90 Max. :5.045 Max. :5228 Max. :1.0000   
##

summary(bankTest)

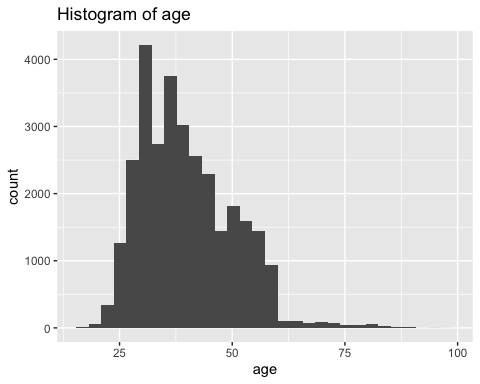
## age job marital education   
## Min. :18.00 admin. :2000 divorced: 875 basic.4y : 818   
## 1st Qu.:32.00 blue-collar:1704 married :4583 basic.6y : 418   
## Median :38.00 technician :1263 single :2191 basic.9y :1160   
## Mean :39.87 services : 715 high.school :1852   
## 3rd Qu.:47.00 management : 554 illiterate : 4   
## Max. :98.00 retired : 328 professional.course:1030   
## (Other) :1085 university.degree :2367   
## housing loan telephone month day\_of\_week duration   
## 0:3487 0:6475 0:4924 may :2561 fri:1432 Min. : 0.0   
## 1:4162 1:1174 1:2725 jul :1314 mon:1570 1st Qu.: 104.0   
## aug :1149 thu:1564 Median : 179.0   
## jun : 970 tue:1527 Mean : 260.8   
## nov : 785 wed:1556 3rd Qu.: 318.0   
## apr : 509 Max. :4918.0   
## (Other): 361   
## campaign previous emp.var.rate cons.price.idx   
## Min. : 1.000 Min. :0.0000 Min. :-3.40000 Min. :92.20   
## 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:-1.80000 1st Qu.:93.08   
## Median : 2.000 Median :0.0000 Median : 1.10000 Median :93.44   
## Mean : 2.611 Mean :0.1685 Mean : 0.07205 Mean :93.57   
## 3rd Qu.: 3.000 3rd Qu.:0.0000 3rd Qu.: 1.40000 3rd Qu.:93.99   
## Max. :42.000 Max. :7.0000 Max. : 1.40000 Max. :94.77   
##   
## cons.conf.idx euribor3m nr.employed deposit   
## Min. :-50.80 Min. :0.634 Min. :4964 Min. :0.0000   
## 1st Qu.:-42.70 1st Qu.:1.344 1st Qu.:5099 1st Qu.:0.0000   
## Median :-41.80 Median :4.857 Median :5191 Median :0.0000   
## Mean :-40.59 Mean :3.616 Mean :5167 Mean :0.1145   
## 3rd Qu.:-36.40 3rd Qu.:4.961 3rd Qu.:5228 3rd Qu.:0.0000   
## Max. :-26.90 Max. :5.045 Max. :5228 Max. :1.0000   
##

### Histogram showing the distribution

Looking at the summary table, a histogram of age, duration, campaign, previous may be worth looking at since the data seems to be skewed and needs some transformation.

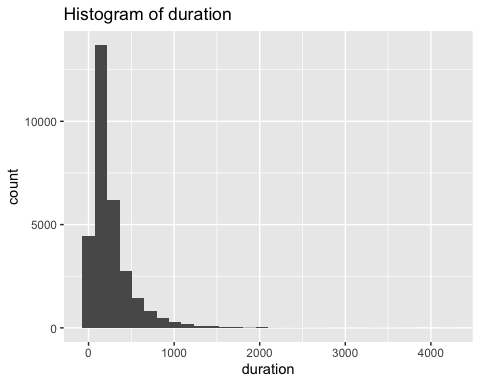
bankTrain %>%  
 ggplot(aes(age)) + geom\_histogram() +  
 labs(title = "Histogram of age")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



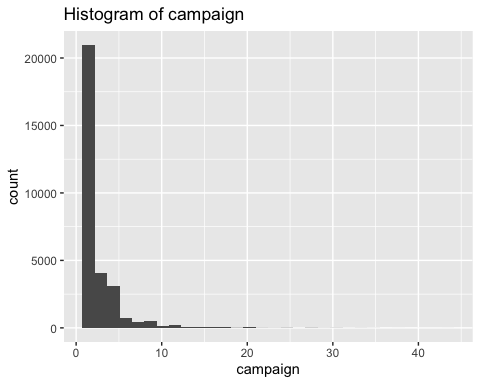
bankTrain %>%  
 ggplot(aes(duration)) + geom\_histogram() +  
 labs(title = "Histogram of duration")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



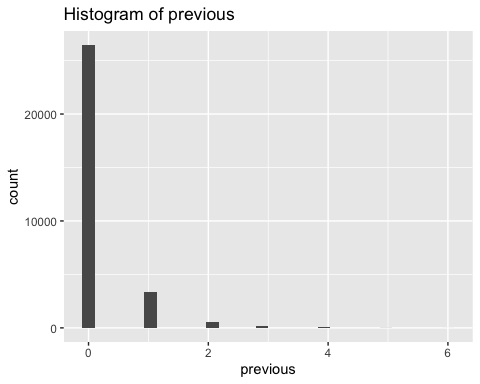
bankTrain %>%  
 ggplot(aes(campaign)) + geom\_histogram() +  
 labs(title = "Histogram of campaign")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



bankTrain %>%  
 ggplot(aes(previous)) + geom\_histogram() +  
 labs(title = "Histogram of previous")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



table(bankTrain$previous)

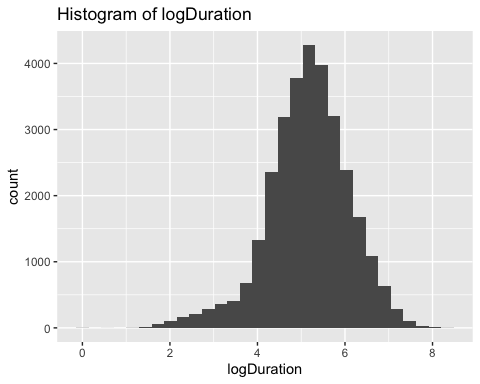
##   
## 0 1 2 3 4 5 6   
## 26451 3386 531 164 50 11 3

We will use log transformation on duration and campaign. We will also turn previous into a categorical variable.

#log transformation  
bankTrain <- bankTrain %>%  
 mutate(logDuration = log(duration))  
  
bankTrain <- bankTrain %>%  
 mutate(logCampaign = log(campaign))  
  
bankTrain %>%  
 ggplot(aes(logDuration)) + geom\_histogram() +  
 labs(title = "Histogram of logDuration")

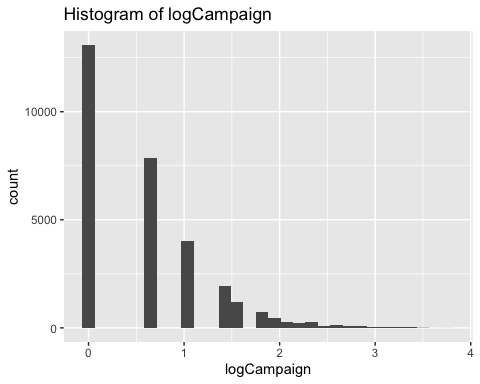
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing non-finite values (stat\_bin).



bankTrain %>%  
 ggplot(aes(logCampaign)) + geom\_histogram() +  
 labs(title = "Histogram of logCampaign")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



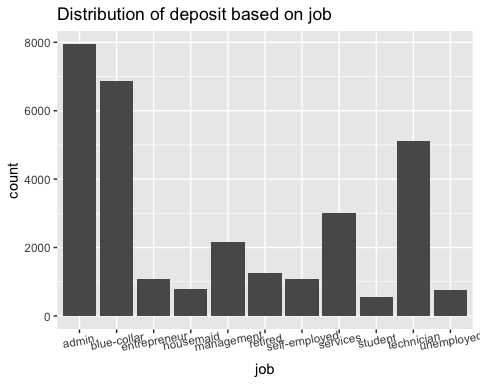
campaign seems to be worsened off so we may leave it alone instead of taking the log transformation.

bankTrain <- bankTrain %>%  
 select(!logCampaign) %>%  
 select(!duration)

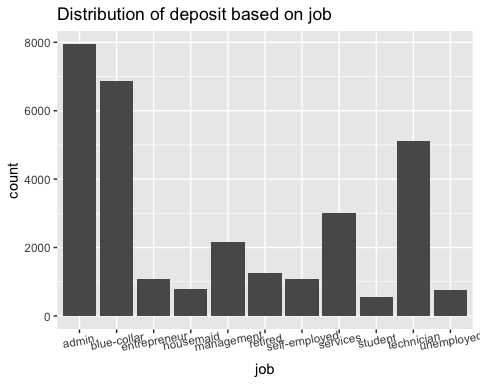
### Other plots to identify potential relationship

We will draw some more plots to see which categorical variables potentially have influence on deposit

bankTrain %>%   
 ggplot(aes(x = job, fill = deposit)) + geom\_bar() + labs(title = "Distribution of deposit based on job") + theme(axis.text.x = element\_text(angle = 10))



bankTrain %>%   
 ggplot(aes(x = job, fill = deposit)) + geom\_bar() + labs(title = "Distribution of deposit based on job") + theme(axis.text.x = element\_text(angle = 10))



## Modeling process

bankTrain <- bankTrain %>%  
 filter(logDuration != -Inf)  
  
reg <- glm(deposit ~ ., bankTrain, family = binomial)  
summary(reg)

##   
## Call:  
## glm(formula = deposit ~ ., family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.5561 -0.3189 -0.1485 -0.0632 3.7500   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.530e+02 4.478e+01 -5.651 1.59e-08 \*\*\*  
## age -2.421e-03 2.850e-03 -0.850 0.395576   
## jobblue-collar -2.300e-01 9.211e-02 -2.497 0.012529 \*   
## jobentrepreneur -1.958e-01 1.453e-01 -1.348 0.177596   
## jobhousemaid -1.132e-01 1.708e-01 -0.663 0.507475   
## jobmanagement -8.614e-03 9.853e-02 -0.087 0.930332   
## jobretired 4.055e-01 1.268e-01 3.198 0.001383 \*\*   
## jobself-employed -1.192e-01 1.356e-01 -0.879 0.379300   
## jobservices -2.502e-01 1.006e-01 -2.486 0.012911 \*   
## jobstudent 2.193e-01 1.428e-01 1.536 0.124587   
## jobtechnician 2.858e-02 8.226e-02 0.347 0.728259   
## jobunemployed 1.060e-01 1.473e-01 0.719 0.471833   
## maritalmarried -8.433e-04 7.851e-02 -0.011 0.991430   
## maritalsingle 4.821e-03 9.012e-02 0.053 0.957341   
## educationbasic.6y 1.770e-01 1.346e-01 1.315 0.188473   
## educationbasic.9y 8.958e-02 1.067e-01 0.840 0.400958   
## educationhigh.school 1.414e-01 1.051e-01 1.345 0.178665   
## educationilliterate 1.181e+00 8.270e-01 1.428 0.153161   
## educationprofessional.course 2.212e-01 1.150e-01 1.923 0.054458 .   
## educationuniversity.degree 2.847e-01 1.060e-01 2.686 0.007236 \*\*   
## housing1 1.754e-04 4.763e-02 0.004 0.997062   
## loan1 -5.632e-02 6.627e-02 -0.850 0.395396   
## telephone1 -6.686e-01 9.534e-02 -7.012 2.35e-12 \*\*\*  
## monthaug 1.193e+00 1.482e-01 8.045 8.60e-16 \*\*\*  
## monthdec 4.219e-01 2.532e-01 1.666 0.095679 .   
## monthjul 3.602e-01 1.144e-01 3.148 0.001643 \*\*   
## monthjun -3.054e-01 1.510e-01 -2.022 0.043135 \*   
## monthmar 2.478e+00 1.800e-01 13.765 < 2e-16 \*\*\*  
## monthmay -3.496e-01 9.853e-02 -3.548 0.000388 \*\*\*  
## monthnov -3.442e-01 1.403e-01 -2.454 0.014145 \*   
## monthoct 4.315e-01 1.804e-01 2.391 0.016794 \*   
## monthsep 8.618e-01 2.126e-01 4.054 5.04e-05 \*\*\*  
## day\_of\_weekmon -9.379e-02 7.695e-02 -1.219 0.222946   
## day\_of\_weekthu 6.718e-02 7.457e-02 0.901 0.367675   
## day\_of\_weektue 2.946e-02 7.739e-02 0.381 0.703414   
## day\_of\_weekwed 1.707e-01 7.610e-02 2.243 0.024916 \*   
## campaign -2.040e-02 1.329e-02 -1.535 0.124803   
## previous 1.590e-01 3.948e-02 4.027 5.64e-05 \*\*\*  
## emp.var.rate -2.003e+00 1.709e-01 -11.721 < 2e-16 \*\*\*  
## cons.price.idx 2.342e+00 2.974e-01 7.878 3.34e-15 \*\*\*  
## cons.conf.idx 2.603e-02 9.512e-03 2.737 0.006208 \*\*   
## euribor3m 5.835e-01 1.523e-01 3.830 0.000128 \*\*\*  
## nr.employed 3.442e-03 3.605e-03 0.955 0.339654   
## logDuration 2.235e+00 4.037e-02 55.354 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12182 on 30550 degrees of freedom  
## AIC: 12270  
##   
## Number of Fisher Scoring iterations: 7

We would remove age, job, marital, education, housing, loan, day\_of\_week, campaign, nr.employed

reg2 <- glm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration, bankTrain, family = binomial)  
summary(reg2)

##   
## Call:  
## glm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration,   
## family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.5786 -0.3211 -0.1504 -0.0651 3.5965   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.180e+02 1.247e+01 -17.482 < 2e-16 \*\*\*  
## telephone1 -6.623e-01 9.034e-02 -7.332 2.27e-13 \*\*\*  
## monthaug 1.248e+00 1.325e-01 9.419 < 2e-16 \*\*\*  
## monthdec 3.976e-01 2.377e-01 1.673 0.094366 .   
## monthjul 3.916e-01 1.125e-01 3.482 0.000498 \*\*\*  
## monthjun -2.219e-01 1.222e-01 -1.816 0.069418 .   
## monthmar 2.465e+00 1.514e-01 16.282 < 2e-16 \*\*\*  
## monthmay -4.003e-01 8.953e-02 -4.471 7.77e-06 \*\*\*  
## monthnov -3.411e-01 1.252e-01 -2.724 0.006446 \*\*   
## monthoct 3.845e-01 1.492e-01 2.577 0.009969 \*\*   
## monthsep 8.062e-01 1.612e-01 5.002 5.68e-07 \*\*\*  
## previous 1.597e-01 3.938e-02 4.055 5.02e-05 \*\*\*  
## emp.var.rate -1.957e+00 1.369e-01 -14.303 < 2e-16 \*\*\*  
## cons.price.idx 2.153e+00 1.291e-01 16.676 < 2e-16 \*\*\*  
## cons.conf.idx 2.292e-02 6.733e-03 3.405 0.000662 \*\*\*  
## euribor3m 7.030e-01 1.035e-01 6.794 1.09e-11 \*\*\*  
## logDuration 2.223e+00 4.009e-02 55.456 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12267 on 30577 degrees of freedom  
## AIC: 12301  
##   
## Number of Fisher Scoring iterations: 7

lrm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration, bankTrain)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration,   
## data = bankTrain)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 9005.42 R2 0.509 C 0.933   
## 0 27212 d.f. 16 g 2.764 Dxy 0.866   
## 1 3382 Pr(> chi2) <0.0001 gr 15.870 gamma 0.866   
## max |deriv| 1e-09 gp 0.165 tau-a 0.170   
## Brier 0.063   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -218.0026 12.4701 -17.48 <0.0001   
## telephone=1 -0.6623 0.0903 -7.33 <0.0001   
## month=aug 1.2484 0.1325 9.42 <0.0001   
## month=dec 0.3976 0.2377 1.67 0.0944   
## month=jul 0.3916 0.1125 3.48 0.0005   
## month=jun -0.2219 0.1222 -1.82 0.0694   
## month=mar 2.4646 0.1514 16.28 <0.0001   
## month=may -0.4003 0.0895 -4.47 <0.0001   
## month=nov -0.3411 0.1252 -2.72 0.0064   
## month=oct 0.3845 0.1492 2.58 0.0100   
## month=sep 0.8062 0.1612 5.00 <0.0001   
## previous 0.1597 0.0394 4.05 <0.0001   
## emp.var.rate -1.9575 0.1369 -14.30 <0.0001   
## cons.price.idx 2.1529 0.1291 16.68 <0.0001   
## cons.conf.idx 0.0229 0.0067 3.40 0.0007   
## euribor3m 0.7030 0.1035 6.79 <0.0001   
## logDuration 2.2232 0.0401 55.46 <0.0001   
##

We may consider remove month

reg3 <- glm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + euribor3m + cons.conf.idx + logDuration, bankTrain, family = binomial)  
summary(reg3)

##   
## Call:  
## glm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + euribor3m + cons.conf.idx + logDuration,   
## family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1289 -0.3441 -0.1633 -0.0712 3.7512   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.646e+02 7.453e+00 -22.080 < 2e-16 \*\*\*  
## telephone1 -1.146e+00 7.148e-02 -16.026 < 2e-16 \*\*\*  
## previous 1.411e-01 3.773e-02 3.739 0.000185 \*\*\*  
## emp.var.rate -1.045e+00 7.954e-02 -13.133 < 2e-16 \*\*\*  
## cons.price.idx 1.647e+00 7.786e-02 21.149 < 2e-16 \*\*\*  
## euribor3m 3.723e-02 6.222e-02 0.598 0.549589   
## cons.conf.idx 9.261e-02 4.803e-03 19.284 < 2e-16 \*\*\*  
## logDuration 2.124e+00 3.839e-02 55.336 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12803 on 30586 degrees of freedom  
## AIC: 12819  
##   
## Number of Fisher Scoring iterations: 7

lrm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + euribor3m + cons.conf.idx + logDuration, bankTrain)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + euribor3m + cons.conf.idx + logDuration,   
## data = bankTrain)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 8469.29 R2 0.483 C 0.922   
## 0 27212 d.f. 7 g 2.655 Dxy 0.843   
## 1 3382 Pr(> chi2) <0.0001 gr 14.231 gamma 0.843   
## max |deriv| 1e-09 gp 0.161 tau-a 0.166   
## Brier 0.064   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -164.5669 7.4532 -22.08 <0.0001   
## telephone=1 -1.1456 0.0715 -16.03 <0.0001   
## previous 0.1411 0.0377 3.74 0.0002   
## emp.var.rate -1.0447 0.0795 -13.13 <0.0001   
## cons.price.idx 1.6467 0.0779 21.15 <0.0001   
## euribor3m 0.0372 0.0622 0.60 0.5496   
## cons.conf.idx 0.0926 0.0048 19.28 <0.0001   
## logDuration 2.1243 0.0384 55.34 <0.0001   
##

Also, remove euribor3m.

reg4 <- glm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + cons.conf.idx + logDuration, bankTrain, family = binomial)  
summary(reg4)

##   
## Call:  
## glm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + logDuration, family = binomial,   
## data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1344 -0.3443 -0.1632 -0.0712 3.7661   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.618e+02 5.781e+00 -27.982 < 2e-16 \*\*\*  
## telephone1 -1.134e+00 6.899e-02 -16.444 < 2e-16 \*\*\*  
## previous 1.376e-01 3.729e-02 3.691 0.000223 \*\*\*  
## emp.var.rate -9.994e-01 2.418e-02 -41.330 < 2e-16 \*\*\*  
## cons.price.idx 1.619e+00 6.193e-02 26.134 < 2e-16 \*\*\*  
## cons.conf.idx 9.366e-02 4.468e-03 20.963 < 2e-16 \*\*\*  
## logDuration 2.124e+00 3.838e-02 55.336 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12803 on 30587 degrees of freedom  
## AIC: 12817  
##   
## Number of Fisher Scoring iterations: 7

lrm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + cons.conf.idx + logDuration, bankTrain)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + logDuration, data = bankTrain)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 8468.93 R2 0.483 C 0.922   
## 0 27212 d.f. 6 g 2.655 Dxy 0.843   
## 1 3382 Pr(> chi2) <0.0001 gr 14.221 gamma 0.843   
## max |deriv| 2e-09 gp 0.161 tau-a 0.166   
## Brier 0.064   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -161.7572 5.7807 -27.98 <0.0001   
## telephone=1 -1.1345 0.0690 -16.44 <0.0001   
## previous 0.1376 0.0373 3.69 0.0002   
## emp.var.rate -0.9994 0.0242 -41.33 <0.0001   
## cons.price.idx 1.6185 0.0619 26.13 <0.0001   
## cons.conf.idx 0.0937 0.0045 20.96 <0.0001   
## logDuration 2.1239 0.0384 55.34 <0.0001   
##

# Drop-in deviance tests  
  
pchisq(21272-12182,43,lower.tail = FALSE)

## [1] 0

pchisq(21272-12267,16,lower.tail = FALSE)

## [1] 0

pchisq(21272-12803,7,lower.tail = FALSE)

## [1] 0

pchisq(21272-12803,6,lower.tail = FALSE)

## [1] 0

We will have to now consider moving forward with one of the models. We should definitely eliminate model 1 as it has too many varriables with statistically insignificant p-values. We should also eliminate model 3 because euribor3m has a statistically insignificant p-values. We are left with model 2 and 4. The drop-in deviance tests yield 0 for all models, indicating that the probability of getting a larger or equal drop-in deviance is also statistically significant (lower than 0.05). This indicates that it’s hard to have a larger or equal drop-in deviance. Thus, our models are adequate. They are significantly better than the null model, explaining a larger amount of variation of deposit.

Since the drop-in deviance test doesn’t point out which model is better, we will take a look at the residual deviance, the AIC and Dxy.

lrm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration, bankTrain)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration,   
## data = bankTrain)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 9005.42 R2 0.509 C 0.933   
## 0 27212 d.f. 16 g 2.764 Dxy 0.866   
## 1 3382 Pr(> chi2) <0.0001 gr 15.870 gamma 0.866   
## max |deriv| 1e-09 gp 0.165 tau-a 0.170   
## Brier 0.063   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -218.0026 12.4701 -17.48 <0.0001   
## telephone=1 -0.6623 0.0903 -7.33 <0.0001   
## month=aug 1.2484 0.1325 9.42 <0.0001   
## month=dec 0.3976 0.2377 1.67 0.0944   
## month=jul 0.3916 0.1125 3.48 0.0005   
## month=jun -0.2219 0.1222 -1.82 0.0694   
## month=mar 2.4646 0.1514 16.28 <0.0001   
## month=may -0.4003 0.0895 -4.47 <0.0001   
## month=nov -0.3411 0.1252 -2.72 0.0064   
## month=oct 0.3845 0.1492 2.58 0.0100   
## month=sep 0.8062 0.1612 5.00 <0.0001   
## previous 0.1597 0.0394 4.05 <0.0001   
## emp.var.rate -1.9575 0.1369 -14.30 <0.0001   
## cons.price.idx 2.1529 0.1291 16.68 <0.0001   
## cons.conf.idx 0.0229 0.0067 3.40 0.0007   
## euribor3m 0.7030 0.1035 6.79 <0.0001   
## logDuration 2.2232 0.0401 55.46 <0.0001   
##

lrm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + cons.conf.idx + logDuration, bankTrain)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + logDuration, data = bankTrain)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 8468.93 R2 0.483 C 0.922   
## 0 27212 d.f. 6 g 2.655 Dxy 0.843   
## 1 3382 Pr(> chi2) <0.0001 gr 14.221 gamma 0.843   
## max |deriv| 2e-09 gp 0.161 tau-a 0.166   
## Brier 0.064   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -161.7572 5.7807 -27.98 <0.0001   
## telephone=1 -1.1345 0.0690 -16.44 <0.0001   
## previous 0.1376 0.0373 3.69 0.0002   
## emp.var.rate -0.9994 0.0242 -41.33 <0.0001   
## cons.price.idx 1.6185 0.0619 26.13 <0.0001   
## cons.conf.idx 0.0937 0.0045 20.96 <0.0001   
## logDuration 2.1239 0.0384 55.34 <0.0001   
##

Clearly, model 2 has a smaller residual deviance (12267) compared to model 4 (12803). Model 2 also has a larger Dxy than model 4, 0.866 compared to 0.843. This means that model 2 fits the data more than model 4, and the variables in model 2 are more significant. The only concern we have is that some of the month is not significant. We may convert month into different Bernoulli variables and eliminate those that are not significant or we may run a drop-in deviance test for month to see if it offers a greater model than the null model.

# Drop-in deviance test for month  
pchisq(12803-12267,16-7,lower.tail = FALSE)

## [1] 1.115819e-109

It appears that month is quite significant since the drop-in test yields 1.11\*10^{-109} as the p-value. After some experiment, we realize that using august as the baseline would make all the p-values significant. This can happen due to the fact that the monthaug may affect deposit different from other months a lot.

bankTrain$month <- relevel(bankTrain$month, ref = "aug")  
reg5 <- glm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration, bankTrain, family = binomial)  
summary(reg5)

##   
## Call:  
## glm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration,   
## family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.5786 -0.3211 -0.1504 -0.0651 3.5965   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.168e+02 1.242e+01 -17.447 < 2e-16 \*\*\*  
## telephone1 -6.623e-01 9.034e-02 -7.332 2.27e-13 \*\*\*  
## monthapr -1.248e+00 1.325e-01 -9.419 < 2e-16 \*\*\*  
## monthdec -8.508e-01 2.381e-01 -3.574 0.000352 \*\*\*  
## monthjul -8.568e-01 1.180e-01 -7.258 3.92e-13 \*\*\*  
## monthjun -1.470e+00 1.599e-01 -9.196 < 2e-16 \*\*\*  
## monthmar 1.216e+00 1.551e-01 7.841 4.48e-15 \*\*\*  
## monthmay -1.649e+00 1.089e-01 -15.145 < 2e-16 \*\*\*  
## monthnov -1.590e+00 1.446e-01 -10.996 < 2e-16 \*\*\*  
## monthoct -8.639e-01 1.584e-01 -5.453 4.96e-08 \*\*\*  
## monthsep -4.422e-01 1.532e-01 -2.886 0.003900 \*\*   
## previous 1.597e-01 3.938e-02 4.055 5.02e-05 \*\*\*  
## emp.var.rate -1.957e+00 1.369e-01 -14.303 < 2e-16 \*\*\*  
## cons.price.idx 2.153e+00 1.291e-01 16.676 < 2e-16 \*\*\*  
## cons.conf.idx 2.292e-02 6.733e-03 3.405 0.000662 \*\*\*  
## euribor3m 7.030e-01 1.035e-01 6.794 1.09e-11 \*\*\*  
## logDuration 2.223e+00 4.009e-02 55.456 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12267 on 30577 degrees of freedom  
## AIC: 12301  
##   
## Number of Fisher Scoring iterations: 7

lrm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration, bankTrain)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration,   
## data = bankTrain)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 9005.42 R2 0.509 C 0.933   
## 0 27212 d.f. 16 g 2.764 Dxy 0.866   
## 1 3382 Pr(> chi2) <0.0001 gr 15.870 gamma 0.866   
## max |deriv| 1e-09 gp 0.165 tau-a 0.170   
## Brier 0.063   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -216.7542 12.4238 -17.45 <0.0001   
## telephone=1 -0.6623 0.0903 -7.33 <0.0001   
## month=apr -1.2484 0.1325 -9.42 <0.0001   
## month=dec -0.8508 0.2381 -3.57 0.0004   
## month=jul -0.8568 0.1180 -7.26 <0.0001   
## month=jun -1.4703 0.1599 -9.20 <0.0001   
## month=mar 1.2161 0.1551 7.84 <0.0001   
## month=may -1.6488 0.1089 -15.15 <0.0001   
## month=nov -1.5895 0.1445 -11.00 <0.0001   
## month=oct -0.8639 0.1584 -5.45 <0.0001   
## month=sep -0.4422 0.1532 -2.89 0.0039   
## previous 0.1597 0.0394 4.05 <0.0001   
## emp.var.rate -1.9575 0.1369 -14.30 <0.0001   
## cons.price.idx 2.1529 0.1291 16.68 <0.0001   
## cons.conf.idx 0.0229 0.0067 3.40 0.0007   
## euribor3m 0.7030 0.1035 6.79 <0.0001   
## logDuration 2.2232 0.0401 55.46 <0.0001   
##

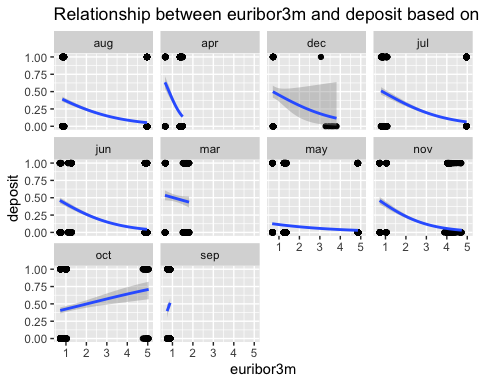
We will move forward with model 2 and try to improve the model by adding interaction terms.

## Interaction Terms

Since when we remove month, euribor3m is also affected a lot, we will try an interaction term between month and euribor3m.

bankTrain %>%   
 ggplot(aes(euribor3m, deposit)) + geom\_point() +  
 stat\_smooth(method = "glm",method.args = list(family= "binomial")) +  
 facet\_wrap(~month) +  
 labs(title = "Relationship between euribor3m and deposit based on month")

## `geom\_smooth()` using formula 'y ~ x'



reg6 <- glm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration + month\*euribor3m, bankTrain, family = binomial)  
summary(reg6)

##   
## Call:  
## glm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration +   
## month \* euribor3m, family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.0986 -0.3117 -0.1434 -0.0614 3.5207   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -194.43155 19.41146 -10.016 < 2e-16 \*\*\*  
## telephone1 -0.49594 0.10055 -4.932 8.12e-07 \*\*\*  
## monthapr 0.96046 0.49553 1.938 0.052594 .   
## monthdec -1.93622 0.47475 -4.078 4.53e-05 \*\*\*  
## monthjul -0.46212 0.23329 -1.981 0.047611 \*   
## monthjun -0.27596 0.25487 -1.083 0.278927   
## monthmar -0.71979 0.40586 -1.774 0.076145 .   
## monthmay -0.53290 0.22848 -2.332 0.019684 \*   
## monthnov -0.86790 0.24884 -3.488 0.000487 \*\*\*  
## monthoct -1.71380 0.20999 -8.161 3.31e-16 \*\*\*  
## monthsep -5.56421 2.51206 -2.215 0.026760 \*   
## previous 0.16019 0.03967 4.038 5.39e-05 \*\*\*  
## emp.var.rate -1.52489 0.23337 -6.534 6.39e-11 \*\*\*  
## cons.price.idx 1.94062 0.20144 9.634 < 2e-16 \*\*\*  
## cons.conf.idx 0.07708 0.01183 6.515 7.28e-11 \*\*\*  
## euribor3m 0.43797 0.19249 2.275 0.022885 \*   
## logDuration 2.27451 0.04097 55.511 < 2e-16 \*\*\*  
## monthapr:euribor3m -1.08968 0.40155 -2.714 0.006654 \*\*   
## monthdec:euribor3m 1.91503 0.50571 3.787 0.000153 \*\*\*  
## monthjul:euribor3m -0.01630 0.05112 -0.319 0.749770   
## monthjun:euribor3m -0.22983 0.05471 -4.201 2.66e-05 \*\*\*  
## monthmar:euribor3m 2.17512 0.35015 6.212 5.23e-10 \*\*\*  
## monthmay:euribor3m -0.28014 0.07477 -3.747 0.000179 \*\*\*  
## monthnov:euribor3m -0.06484 0.06480 -1.001 0.317035   
## monthoct:euribor3m 1.10094 0.14615 7.533 4.96e-14 \*\*\*  
## monthsep:euribor3m 6.33062 2.99808 2.112 0.034724 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12031 on 30568 degrees of freedom  
## AIC: 12083  
##   
## Number of Fisher Scoring iterations: 7

lrm(deposit ~ telephone + month + previous + emp.var.rate + cons.price.idx + cons.conf.idx + euribor3m + logDuration + month\*euribor3m + campaign, bankTrain, maxit=1000)

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + month + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + euribor3m + logDuration +   
## month \* euribor3m + campaign, data = bankTrain, maxit = 1000)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 9244.03 R2 0.520 C 0.936   
## 0 27212 d.f. 26 g 2.845 Dxy 0.872   
## 1 3382 Pr(> chi2) <0.0001 gr 17.202 gamma 0.872   
## max |deriv| 3e-09 gp 0.166 tau-a 0.171   
## Brier 0.061   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -194.0849 19.4093 -10.00 <0.0001   
## telephone=1 -0.4820 0.1009 -4.77 <0.0001   
## month=apr 0.9683 0.4961 1.95 0.0509   
## month=dec -1.9110 0.4752 -4.02 <0.0001   
## month=jul -0.4505 0.2335 -1.93 0.0537   
## month=jun -0.2720 0.2549 -1.07 0.2858   
## month=mar -0.7321 0.4061 -1.80 0.0714   
## month=may -0.5236 0.2285 -2.29 0.0219   
## month=nov -0.8586 0.2490 -3.45 0.0006   
## month=oct -1.7078 0.2101 -8.13 <0.0001   
## month=sep -5.5972 2.5127 -2.23 0.0259   
## previous 0.1596 0.0397 4.02 <0.0001   
## emp.var.rate -1.5179 0.2334 -6.50 <0.0001   
## cons.price.idx 1.9374 0.2014 9.62 <0.0001   
## cons.conf.idx 0.0772 0.0118 6.52 <0.0001   
## euribor3m 0.4369 0.1925 2.27 0.0232   
## logDuration 2.2770 0.0410 55.49 <0.0001   
## campaign -0.0238 0.0135 -1.76 0.0777   
## month=apr \* euribor3m -1.0970 0.4019 -2.73 0.0063   
## month=dec \* euribor3m 1.8950 0.5061 3.74 0.0002   
## month=jul \* euribor3m -0.0179 0.0511 -0.35 0.7265   
## month=jun \* euribor3m -0.2319 0.0547 -4.24 <0.0001   
## month=mar \* euribor3m 2.1917 0.3504 6.25 <0.0001   
## month=may \* euribor3m -0.2856 0.0748 -3.82 0.0001   
## month=nov \* euribor3m -0.0699 0.0649 -1.08 0.2816   
## month=oct \* euribor3m 1.0920 0.1463 7.46 <0.0001   
## month=sep \* euribor3m 6.3707 2.9988 2.12 0.0336   
##

Most p-values are statistically significant (< 0.05), and Dxy increases from 0.866 to 0.872 so we would move forward with this model.

## Adding one more variable

Looking back at the first model, we realize that higher education (professional.course and university.degree) is actually quite significant. We would try to create a dummy variable for higher education

table(bankTrain$education)

##   
## basic.4y basic.6y basic.9y high.school   
## 3184 1786 4695 7392   
## illiterate professional.course university.degree   
## 14 4070 9453

bankTrain <- bankTrain %>%  
 mutate(higherEd = ifelse(education == "professional.course" | education == "university.degree", 1 ,0))  
  
reg6 <- glm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + cons.conf.idx + logDuration + month\*euribor3m + higherEd, bankTrain, family = binomial)  
summary(reg6)

##   
## Call:  
## glm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + logDuration + month \* euribor3m +   
## higherEd, family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1241 -0.3106 -0.1434 -0.0606 3.4987   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.926e+02 1.942e+01 -9.914 < 2e-16 \*\*\*  
## telephone1 -4.998e-01 1.006e-01 -4.968 6.78e-07 \*\*\*  
## previous 1.606e-01 3.971e-02 4.045 5.24e-05 \*\*\*  
## emp.var.rate -1.506e+00 2.334e-01 -6.451 1.11e-10 \*\*\*  
## cons.price.idx 1.920e+00 2.016e-01 9.522 < 2e-16 \*\*\*  
## cons.conf.idx 7.659e-02 1.184e-02 6.471 9.74e-11 \*\*\*  
## logDuration 2.278e+00 4.103e-02 55.523 < 2e-16 \*\*\*  
## monthapr 9.520e-01 4.959e-01 1.920 0.054884 .   
## monthdec -1.909e+00 4.776e-01 -3.997 6.42e-05 \*\*\*  
## monthjul -4.581e-01 2.331e-01 -1.966 0.049351 \*   
## monthjun -2.874e-01 2.547e-01 -1.128 0.259281   
## monthmar -7.295e-01 4.060e-01 -1.797 0.072347 .   
## monthmay -5.148e-01 2.286e-01 -2.252 0.024325 \*   
## monthnov -8.559e-01 2.489e-01 -3.439 0.000585 \*\*\*  
## monthoct -1.697e+00 2.101e-01 -8.076 6.70e-16 \*\*\*  
## monthsep -5.287e+00 2.519e+00 -2.099 0.035806 \*   
## euribor3m 4.183e-01 1.925e-01 2.173 0.029756 \*   
## higherEd 1.903e-01 4.867e-02 3.911 9.20e-05 \*\*\*  
## monthapr:euribor3m -1.067e+00 4.019e-01 -2.655 0.007932 \*\*   
## monthdec:euribor3m 1.899e+00 5.104e-01 3.721 0.000199 \*\*\*  
## monthjul:euribor3m -4.836e-03 5.116e-02 -0.095 0.924686   
## monthjun:euribor3m -2.111e-01 5.493e-02 -3.843 0.000122 \*\*\*  
## monthmar:euribor3m 2.174e+00 3.502e-01 6.208 5.35e-10 \*\*\*  
## monthmay:euribor3m -2.673e-01 7.488e-02 -3.570 0.000357 \*\*\*  
## monthnov:euribor3m -6.019e-02 6.483e-02 -0.928 0.353176   
## monthoct:euribor3m 1.112e+00 1.461e-01 7.610 2.75e-14 \*\*\*  
## monthsep:euribor3m 6.002e+00 3.005e+00 1.997 0.045817 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12016 on 30567 degrees of freedom  
## AIC: 12070  
##   
## Number of Fisher Scoring iterations: 7

lrm <- lrm(deposit ~ telephone + previous + emp.var.rate + cons.price.idx + cons.conf.idx + logDuration + month\*euribor3m + higherEd, bankTrain, maxit = 1000)

Most p-values are statistically significant (lower than 0.05), and Dxy increases from 0.872 to 0.873 so we will decide on using this model

exp(0.4850)

## [1] 1.624175

exp(0.16)

## [1] 1.173511

exp(1.498)

## [1] 4.472735

exp(1.916)

## [1] 6.793729

exp(0.0767)

## [1] 1.079718

exp(0.4169)

## [1] 1.517251

exp(2.2811)

## [1] 9.787441

exp(0.1904)

## [1] 1.209733

summary(reg6)

##   
## Call:  
## glm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + logDuration + month \* euribor3m +   
## higherEd, family = binomial, data = bankTrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1241 -0.3106 -0.1434 -0.0606 3.4987   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.926e+02 1.942e+01 -9.914 < 2e-16 \*\*\*  
## telephone1 -4.998e-01 1.006e-01 -4.968 6.78e-07 \*\*\*  
## previous 1.606e-01 3.971e-02 4.045 5.24e-05 \*\*\*  
## emp.var.rate -1.506e+00 2.334e-01 -6.451 1.11e-10 \*\*\*  
## cons.price.idx 1.920e+00 2.016e-01 9.522 < 2e-16 \*\*\*  
## cons.conf.idx 7.659e-02 1.184e-02 6.471 9.74e-11 \*\*\*  
## logDuration 2.278e+00 4.103e-02 55.523 < 2e-16 \*\*\*  
## monthapr 9.520e-01 4.959e-01 1.920 0.054884 .   
## monthdec -1.909e+00 4.776e-01 -3.997 6.42e-05 \*\*\*  
## monthjul -4.581e-01 2.331e-01 -1.966 0.049351 \*   
## monthjun -2.874e-01 2.547e-01 -1.128 0.259281   
## monthmar -7.295e-01 4.060e-01 -1.797 0.072347 .   
## monthmay -5.148e-01 2.286e-01 -2.252 0.024325 \*   
## monthnov -8.559e-01 2.489e-01 -3.439 0.000585 \*\*\*  
## monthoct -1.697e+00 2.101e-01 -8.076 6.70e-16 \*\*\*  
## monthsep -5.287e+00 2.519e+00 -2.099 0.035806 \*   
## euribor3m 4.183e-01 1.925e-01 2.173 0.029756 \*   
## higherEd 1.903e-01 4.867e-02 3.911 9.20e-05 \*\*\*  
## monthapr:euribor3m -1.067e+00 4.019e-01 -2.655 0.007932 \*\*   
## monthdec:euribor3m 1.899e+00 5.104e-01 3.721 0.000199 \*\*\*  
## monthjul:euribor3m -4.836e-03 5.116e-02 -0.095 0.924686   
## monthjun:euribor3m -2.111e-01 5.493e-02 -3.843 0.000122 \*\*\*  
## monthmar:euribor3m 2.174e+00 3.502e-01 6.208 5.35e-10 \*\*\*  
## monthmay:euribor3m -2.673e-01 7.488e-02 -3.570 0.000357 \*\*\*  
## monthnov:euribor3m -6.019e-02 6.483e-02 -0.928 0.353176   
## monthoct:euribor3m 1.112e+00 1.461e-01 7.610 2.75e-14 \*\*\*  
## monthsep:euribor3m 6.002e+00 3.005e+00 1.997 0.045817 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21272 on 30593 degrees of freedom  
## Residual deviance: 12016 on 30567 degrees of freedom  
## AIC: 12070  
##   
## Number of Fisher Scoring iterations: 7

lrm

## Logistic Regression Model  
##   
## lrm(formula = deposit ~ telephone + previous + emp.var.rate +   
## cons.price.idx + cons.conf.idx + logDuration + month \* euribor3m +   
## higherEd, data = bankTrain, maxit = 1000)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 30594 LR chi2 9256.08 R2 0.521 C 0.936   
## 0 27212 d.f. 26 g 2.838 Dxy 0.873   
## 1 3382 Pr(> chi2) <0.0001 gr 17.079 gamma 0.873   
## max |deriv| 8e-10 gp 0.166 tau-a 0.172   
## Brier 0.061   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -192.5782 19.4250 -9.91 <0.0001   
## telephone=1 -0.4998 0.1006 -4.97 <0.0001   
## previous 0.1606 0.0397 4.04 <0.0001   
## emp.var.rate -1.5057 0.2334 -6.45 <0.0001   
## cons.price.idx 1.9197 0.2016 9.52 <0.0001   
## cons.conf.idx 0.0766 0.0118 6.47 <0.0001   
## logDuration 2.2783 0.0410 55.52 <0.0001   
## month=apr 0.9520 0.4959 1.92 0.0549   
## month=dec -1.9087 0.4776 -4.00 <0.0001   
## month=jul -0.4581 0.2331 -1.97 0.0494   
## month=jun -0.2874 0.2547 -1.13 0.2593   
## month=mar -0.7295 0.4060 -1.80 0.0723   
## month=may -0.5148 0.2286 -2.25 0.0243   
## month=nov -0.8559 0.2489 -3.44 0.0006   
## month=oct -1.6968 0.2101 -8.08 <0.0001   
## month=sep -5.2871 2.5187 -2.10 0.0358   
## euribor3m 0.4183 0.1925 2.17 0.0298   
## higherEd 0.1904 0.0487 3.91 <0.0001   
## month=apr \* euribor3m -1.0669 0.4019 -2.65 0.0079   
## month=dec \* euribor3m 1.8990 0.5104 3.72 0.0002   
## month=jul \* euribor3m -0.0048 0.0512 -0.09 0.9247   
## month=jun \* euribor3m -0.2111 0.0549 -3.84 0.0001   
## month=mar \* euribor3m 2.1740 0.3502 6.21 <0.0001   
## month=may \* euribor3m -0.2673 0.0749 -3.57 0.0004   
## month=nov \* euribor3m -0.0602 0.0648 -0.93 0.3532   
## month=oct \* euribor3m 1.1122 0.1461 7.61 <0.0001   
## month=sep \* euribor3m 6.0021 3.0055 2.00 0.0458   
##

## Interpretation:

According to the logistic regression model, we have: p(X) =

This means that: - When all variables equal 0 and it’s August, the log odds of deposit is -192.1786, and the probability of the customer subscribing to the term deposit is 3.54\*10^(-84), indicating that the customer will not deposit. - telephone: A shift from using cell phone to telephone is associated with a decrease in the log odds of deposit by 0.4850 units, indicating that if the call method is telephone instead of cell phone, the odds of getting the customers to deposit go down by 1.62 times (or 162%). This makes sense as the dataset also include calls of which customers contact the help center. There will be people who need help and will be frustrated if they don’t get what they need immediately but some introduction to a term deposit that they don’t care about instead. Using cellphone will also create a sense of personal relationship between the caller and the customer instead of a sense of being a part of just a telemarketing campaign. - previous: When previous increases by 1, the log odds of deposit increases by 0.16 units, indicating that the odds of successfully having customers deposit go up by 1.17 times (or 117%). As mention aboved in the variable description, previous means number of contacts performed before this campaign and for this client so the more contacts are performed before, the more experience and familiarity the telemarketing team will possess and hence will persuade the customers better. - emp.var.rate: When emp.var.rate increases by 1, the log odds of deposit decreases by -1.498 units, indicating that the odds of successfully having customers deposit go down by 4.47 times (or 447%). emp.var.rate measures the variation of employment rate. High emp.var.rate indicates an unstable economy. - cons.price.idx: When cons.price.idx increases by 1, the log odds of deposit increases by 1.916 units, indicating that the odds of sucessfully having customers deposit go up by 6.79 times (or 679%). Higher cons.price.idx means higher inflation rate -> higher nominal interest rate so it makes sense that the higher the cons.price.idx, the higher chance the customers want to subscribe to a term deposit. - cons.conf.idx: When cons.conf.idx increases by 1, the log odds of deposit increases by 0.0767 units, indicating that the odds of successfully having customers deposit go up by 1.08 times (or 108%). It makes sense that the higher the consumer confidence level is, the higher chance they want to subscribe to a term deposit. - euribor3m: When euribo3m increases by 1, the log odds of deposit increases by 0.4169 units, indicating that the odds of successfully having customers deposit go up by 1.52 times (or 152%). euribor3m is the average interest rates from a panel of large European banks that are used for lending to one another. - logDuration: When logDuration increases by 1, the log odds of deposit increases by 2.2811 units, indicating that the odds of successfully having customers deposit go up by 9.79 times (or 979%). This makes sense since normally, if customers receive a call that they don’t care about, they will try to end the call as soon as possible. Longer duration also indicates that the company has more time to persuade the customer to subscribe to the term deposit. - higherEd: A shift from having high education level to not having high education level is associated with an increase in the log odds of deposit by 0.1904 units, indicating that if the customer has university degree or professional courses, the odds of getting the customers to deposit go up by 1.21 times (or 121%). - month: Of all the months, december, november and october are associated with a decrease in the log odds of deposit. This might indicate that customers tend to keep money by themselves or use it at the end of the year.

* C is 0.936. This indicates that 93.6% of pairs of 0 and 1 fit the model (0.5->1 - random guessing).
* Dxy is 0.873. This is a rescale of C to make it range from 0 to 1 instead of 0.5 to 1. 0.873 is still a good number. This model explains about 87.3% of our data.

## Validate the model

### Drop-in deviance test

pchisq(21272 - 12016, 27, lower.tail = FALSE)

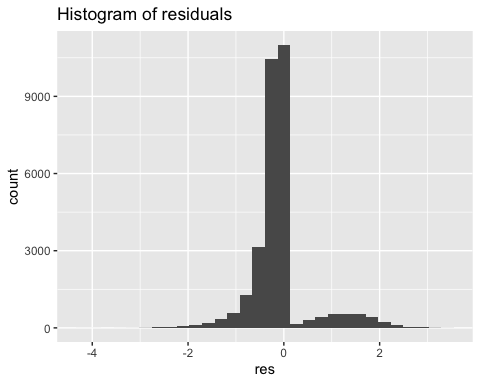
## [1] 0

The drop-in deviance tests yield 0, indicating that the probability of getting a larger or equal drop-in deviance is also statistically significant (lower than 0.05). This indicates that it’s hard to have a larger or equal drop-in deviance. Thus, our model is adequate. This is significantly better than the null model, explaining a larger amount of variation of deposit.

### Check for binormiality and normal distribution

bankTrain2 <- bankTrain %>%  
 mutate(res = resid(reg6), fit = fitted(reg6))  
bankTrain2 %>% ggplot(aes(res)) + geom\_histogram() + labs(title = "Histogram of residuals")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The histogram of residual is unbalanced. This may be due to the unbalanced dataset that we have (more ‘no’ than ‘yes’ for deposit). However, the distribution still centers at 0 so this is still acceptable. We will also run the Hosmer-Lem test to check for binormiality.

### So do we want binormial or normal distribution?

hoslem.test(bankTrain2$deposit, bankTrain2$fit,g=3000)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: bankTrain2$deposit, bankTrain2$fit  
## X-squared = 2891.9, df = 2998, p-value = 0.9159

We change the number of bins to 3000 to fit the size of the dataset. The p-value is 0.9159. It is not statistically significant (higher than 0.05) so we fail to reject the null hypothesis that the residuals have a binormal distribution.

## Prediction

### Random Prediction

We will start with predicting 3 random observations from the test set

bankTest2 <- bankTest %>%  
 mutate(logDuration = log(duration)) %>%  
 mutate(higherEd = ifelse(education == "professional.course" | education == "university.degree", 1 ,0)) %>%  
 filter(logDuration != -Inf)  
  
set.seed(4)  
N <- seq(7649)  
random <- sample(N, 3)  
bankTestPt <- bankTest2[random, ]  
  
bankTestPt <- bankTestPt %>%  
 mutate(result = 0)  
bankTestPt[1, 'result'] <- predict(reg6, bankTestPt[1,], type = "response")  
bankTestPt[2, 'result'] <- predict(reg6, bankTestPt[2,], type = "response")  
bankTestPt[3, 'result'] <- predict(reg6, bankTestPt[3,], type = "response")  
bankTestPt %>%  
 dplyr::select(deposit, result)

## # A tibble: 3 × 2  
## deposit result  
## <dbl> <dbl>  
## 1 1 0.271   
## 2 0 0.0109  
## 3 0 0.0621

The result when the actual deposit is 1 is quite higher than the results when the actual deposits are 0 but still not high enough. We will make some calculations to better assess the result. ### Recall, Precision, and Accuracy

bankTest3 <- bankTest2 %>%  
 mutate(predict = predict(reg6, bankTest2, type = "response")) %>%  
 mutate(predictDeposit = ifelse(predict < 0.5,0,1))  
  
N <- seq(nrow(bankTest3))  
  
for (i in N) {  
 if ((bankTest3[i,'predictDeposit'] == 0) && (bankTest3[i, 'deposit'] == 1)) {  
 bankTest3[i,'result'] = "FN"  
 }  
 else if ((bankTest3[i,'predictDeposit'] == 1) && (bankTest3[i, 'deposit'] == 1)) {  
 bankTest3[i,'result'] = "TP"  
 }  
 else if ((bankTest3[i,'predictDeposit'] == 1) && (bankTest3[i, 'deposit'] == 0)) {  
 bankTest3[i,'result'] = "FP"  
 }  
 else {  
 bankTest3[i,'result'] = "TN"  
 }  
}

FN <- sum(bankTest3$result == "FN")  
FP <- sum(bankTest3$result == "FP")  
TN <- sum(bankTest3$result == "TN")  
TP <- sum(bankTest3$result == "TP")  
recall <- TP/(TP + FN)  
precision <- TP/(TP + FP)  
accuracy <- (TP + TN)/(FP + FN + TP + TN)  
  
recall

## [1] 0.4383562

precision

## [1] 0.6748682

accuracy

## [1] 0.9114685

* Recall: From all the customers subscribe to the term deposit, we predict 43.8% of them.
* Precision: From all the customers we predict that they subscribe, 67.4% of them actually do.
* Accuracy: Fromm all predictions, 91.14% is correct.