Logistic Regression Customer Attrition Case Study

Niranjan

10 December 2018

Logistic Regression Customer Attrition Case Study

Loading all the required packages

Importing and examining the file

```
cust <- fread("~/Case studies/Logistic Regression/Proactive Attrition Management-Logistic Reg
ression Case Study.csv",stringsAsFactors = T)

str(cust)
```

```
## Classes 'data.table' and 'data.frame':   71047 obs. of  78 variables:
##  $ REVENUE : num  57.5 82.3 31.7 62.1 25.2 ...
##  $ MOU     : num  482.8 1312.2 25.5 97.5 2.5 ...
##  $ RECCHRGE: num  37.4 75 30 66 25 ...
##  $ DIRECTAS: num  0.25 1.24 0.25 2.48 0 2.23 0.25 0 0.74 0 ...
##  $ OVERAGE : num  22.8 0 0 0 0 ...
##  $ ROAM    : num  0 0 0 0 0 35.5 0 0 1.29 0 ...
##  $ CHANGEM : num  532.2 156.8 59.5 23.5 -2.5 ...
##  $ CHANGER : num  50.99 8.14 4.03 6.82 -0.23 ...
##  $ DROPVCE : num  8.33 52 0 0 0 9 3.33 2 2.67 1.67 ...
##  $ BLCKVCE : num  1 7.67 1 0.33 0 0 1.67 0.67 6 0.33 ...
##  $ UNANSVCE: num  61.33 76 2.33 4 0.33 ...
##  $ CUSTCARE: num  1.67 4.33 0 4 0 0.33 1 0 4.33 0.33 ...
##  $ THREEWAY: num  0.33 1.33 0 0 0 0 0 0 0 0 ...
##  $ MOUREC  : num  55.28 200.32 0 0 1.13 ...
##  $ OUTCALLS: num  46.33 370.33 0 3.67 0.33 ...
##  $ INCALLS : num  6.33 147 0 0 0 4.67 3.67 4.67 8.33 5.67 ...
##  $ PEAKVCE : num  83.67 555.67 1.67 7.67 0.67 ...
##  $ OPEAKVCE: num  157 303.67 1.67 7.33 0.67 ...
##  $ DROPBLK : num  9.33 59.67 1 0.33 0 ...
##  $ CALLFWDV: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CALLWAIT: num  5.67 22.67 0 0 0 ...
##  $ CHURN   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ MONTHS  : int  56 59 57 59 53 59 55 59 52 56 ...
##  $ UNIQSUBS: int  1 2 2 2 2 5 2 3 1 1 ...
##  $ ACTVSUBS: int  1 2 2 2 2 1 2 2 1 1 ...
##  $ CSA     : Factor w/ 774 levels "","AIRAIK803",..: 366 633 583 580 654 745 580 745 583 3
66 ...
##  $ PHONES  : int  7 9 2 3 2 10 5 6 4 4 ...
##  $ MODELS  : int  6 4 2 3 2 6 4 5 4 3 ...
##  $ EQPDAYS : int  240 458 601 464 354 199 697 48 408 253 ...
##  $ CUSTOMER: int  1000002 1000006 1000010 1000011 1000014 1000015 1000016 1000018 1000019
1000020 ...
##  $ AGE1    : int  30 30 52 46 0 30 58 46 58 30 ...
##  $ AGE2    : int  0 0 58 46 0 22 58 0 0 30 ...
##  $ CHILDREN: int  0 0 0 1 0 1 1 1 0 1 ...
##  $ CREDITA : int  0 0 1 1 1 0 0 1 1 0 ...
##  $ CREDITAA: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CREDITB : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ CREDITC : int  0 1 0 0 0 1 0 0 0 0 ...
##  $ CREDITDE: int  1 0 0 0 0 0 0 0 0 0 ...
##  $ CREDITGY: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CREDITZ : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PRIZMRUR: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PRIZMUB : int  0 0 0 0 0 1 1 0 1 0 ...
##  $ PRIZMTWN: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ REFURB  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ WEBCAP  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ TRUCK   : int  1 0 0 0 1 0 0 1 0 0 ...
##  $ RV      : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ OCCPROF : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ OCCCLER : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCCRFT : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCSTUD : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCHMKR : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCRET  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCSELF : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ OWNRENT : int  1 0 0 0 1 0 0 0 0 0 ...
##  $ MARRYUN : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ MARRYYES: int  0 0 1 0 0 0 0 0 0 0 ...
##  $ MARRYNO : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ MAILORD : int  1 1 1 1 0 0 0 1 1 1 ...
##  $ MAILRES : int  1 1 1 1 0 1 0 1 1 1 ...
##  $ MAILFLAG: int  0 0 0 0 0 0 0 1 0 0 ...
##  $ TRAVEL  : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ PCOWN   : int  0 0 0 0 0 1 0 0 0 1 ...
##  $ CREDITCD: int  1 1 1 1 0 1 1 1 1 0 ...
##  $ RETCALLS: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RETACCPT: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ NEWCELLY: int  0 1 0 1 1 1 0 0 0 1 ...
##  $ NEWCELLN: int  1 0 1 0 0 0 1 1 1 0 ...
##  $ REFER   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ INCMISS : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ INCOME  : int  5 6 9 6 7 3 1 4 3 1 ...
##  $ MCYCLE  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CREDITAD: int  1 0 1 0 0 1 1 1 0 0 ...
##  $ SETPRCM : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SETPRC  : num  149.99 9.99 29.99 29.99 29.99 ...
##  $ RETCALL : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CALIBRAT: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CHURNDEP: num  NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Removing a few variables

```
# customerID and CSA values are removed because they can overfit the model because of the num
erous levels and to avoid overfitting.
cust$CUSTOMER <- NULL
cust$CSA <- NULL
# Variables that show value is missing or not will be removed`
cust$INCMISS <- NULL
cust$SETPRCM <- NULL
```

Getting to know the variables better

```
# Creating user generated formulas
about <- function(x){

  n = length(x)
  nmiss = sum(is.na(x))
  nmiss_pct = (mean(is.na(x)))*100
  sum = sum(x, na.rm=T)
  mean = mean(x, na.rm=T)
  median = quantile(x, p=0.5, na.rm=T)
  std = sd(x, na.rm=T)
  var = var(x, na.rm=T)
  range = max(x, na.rm=T)-min(x, na.rm=T)
  pctl = quantile(x, p=c(0, 0.01, 0.05,0.1,0.25,0.5, 0.75,0.9,0.95,0.99,1), na.rm=T)
  return(c(N=n, Nmiss =nmiss, Nmiss_pct = nmiss_pct, sum=sum, avg=mean, meidan=median, std=std, var=var, range=range, pctl=pctl))

}

about_int <- function(x){

  n = length(x)
  nmiss = sum(is.na(x))
  nmiss_pct = (mean(is.na(x)))*100
  range = max(x, na.rm=T)-min(x, na.rm=T)
  fre = table(x)
  prop = prop.table(table(x))
  return(c(N=n, Nmiss =nmiss, Nmiss_pct = nmiss_pct,range=range))

}
```

Describing the continuous numeric variables

```
integs <- names(cust)[sapply(cust, is.integer)]
abt_nums <- as.data.frame(t(sapply(cust[,!integs, with = F],about)))

cust_nums <- cust[,!integs, with = F]
churndep <- cust_nums$CHURNDEP
cust_nums$CHURNDEP <- NULL
fwrite(abt_nums,"~/Case studies/Logistic Regression/abt_nums.csv",row.names = T)
```

Plots are examined

It is seen that none of the variables are normally distributed. Percentiles will be used to cap. Percentiles have shown that all variables need to be capped with 99 percntile values

Capping with 99 percentile value

```
# A function to treat outliers
outlier_treat_99 <- function(x){
  UC1 = quantile(x, p=0.99,na.rm=T)
  LC1 = quantile(x, p=0.01,na.rm=T)

  x=ifelse(x>UC1, UC1, x)
  x=ifelse(x<LC1, LC1, x)
  return(x)
}
cust_nums <- data.table(apply(cust_nums,2, FUN = outlier_treat_99))
```

all the missing values here are less than 1% of the dataset, so they are imputed with mean

```
miss_treat_num = function(x){
  x[is.na(x)] = mean(x,na.rm=T)
  return(x)
}
cust_nums <- data.table(apply(cust_nums,2, FUN = miss_treat_num))
```

now the integers are examined

```
abt_ints <- as.data.frame(t(sapply(cust[,integs, with = F],about_int)))
# missing values in age will be imputed with mean and the rest of the missing will values be
 removed
cust_ints <- cust[,integs, with = F]
cust_ints$AGE1[is.na(cust_ints$AGE1)] <- ceiling(mean(cust_ints$AGE1, na.rm = T))
cust_ints$AGE2[is.na(cust_ints$AGE2)] <- ceiling(mean(cust_ints$AGE2, na.rm = T))
```

The numerics and integers are brought togeather

```
cust_ <- cbind(cust_ints,cust_nums)
cust_ <- cbind(cust_,churndep)
# Remaining missing values are removed
cust_ <- cust_[complete.cases(cust_[,1:25])]
# Removing few more Insignificant variables
cust_$CHURN <- NULL
cust_$marryun <- NULL
```

```
## Warning in `[<-.data.table`(x, j = name, value = value): Adding new column
## 'marryun' then assigning NULL (deleting it).
```

Creating validation and calibration data

```
valid <- cust_[CALIBRAT == 0]
calib <- cust_[CALIBRAT == 1]

calib$CALIBRAT <- NULL
valid$CALIBRAT <- NULL



table(calib$churndep)
```

```
##
##     0     1
## 20000 20000
```

as both the groups are equal, there is no bias in the data and can be used for analysis

Building the model

```
fit1 <- glm(churndep~.,data = calib,family = binomial(logit),maxit=100)
summary(fit1)
```

```
## 
## Call:
## glm(formula = churndep ~ ., family = binomial(logit), data = calib, 
##     maxit = 100)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2709  -1.1383  -0.2548   1.1432   2.0312  
## 
## Coefficients: (3 not defined because of singularities)
##                Estimate  Std. Error z value             Pr(>|z|)    
## (Intercept) -0.10013836  0.09319804  -1.074             0.282613    
## MONTHS      -0.01820888  0.00193532  -9.409 < 0.0000000000000002 ***
## UNIQSUBS     0.19019438  0.01998394   9.517 < 0.0000000000000002 ***
## ACTVSUBS    -0.19845571  0.02752114  -7.211  0.00000000000055531 ***
## PHONES       0.05099706  0.01793923   2.843             0.004472 ** 
## MODELS       0.02426337  0.02736550   0.887             0.375272    
## EQPDAYS      0.00141280  0.00006920  20.415 < 0.0000000000000002 ***
## AGE1        -0.00338145  0.00081694  -4.139  0.00003485906584619 ***
## AGE2        -0.00121196  0.00067575  -1.793             0.072894 .  
## CHILDREN     0.09588476  0.02809350   3.413             0.000642 ***
## CREDITA      0.00650522  0.06130046   0.106             0.915487    
## CREDITAA     0.07135142  0.05490343   1.300             0.193745    
## CREDITB      0.09684065  0.05760256   1.681             0.092727 .  
## CREDITC     -0.09026684  0.06020785  -1.499             0.133808    
## CREDITDE    -0.27556699  0.05894511  -4.675  0.00000293987565197 ***
## CREDITGY    -0.01997071  0.08666679  -0.230             0.817757    
## CREDITZ              NA          NA      NA                   NA    
## PRIZMRUR     0.05861932  0.04918315   1.192             0.233317    
## PRIZMUB     -0.04269399  0.02411374  -1.771             0.076640 .  
## PRIZMTWN     0.03360631  0.03127347   1.075             0.282556    
## REFURB       0.24591012  0.03128956   7.859  0.00000000000000387 ***
## WEBCAP      -0.14982716  0.03723472  -4.024  0.00005725287059356 ***
## TRUCK        0.03716040  0.03585625   1.036             0.300029    
## RV           0.00344438  0.04793818   0.072             0.942721    
## OCCPROF     -0.02779220  0.03241296  -0.857             0.391201    
## OCCCLER      0.05182757  0.07477449   0.693             0.488235    
## OCCCRFT     -0.02171220  0.06281728  -0.346             0.729613    
## OCCSTUD      0.13191974  0.12155360   1.085             0.277798    
## OCCHMKR      0.26965895  0.18929695   1.425             0.154293    
## OCCRET      -0.04119810  0.09015307  -0.457             0.647686    
## OCCSELF     -0.06142688  0.08037549  -0.764             0.444719    
## OWNRENT     -0.01550344  0.04137552  -0.375             0.707883    
## MARRYUN      0.08365454  0.03190647   2.622             0.008745 ** 
## MARRYYES     0.04726416  0.03179747   1.486             0.137170    
## MARRYNO              NA          NA      NA                   NA    
## MAILORD     -0.01008829  0.08534418  -0.118             0.905904    
## MAILRES     -0.11612079  0.08573586  -1.354             0.175608    
## MAILFLAG    -0.03830477  0.08442306  -0.454             0.650028    
## TRAVEL      -0.00137267  0.04719159  -0.029             0.976795    
## PCOWN        0.03108971  0.03088874   1.007             0.314172    
## CREDITCD     0.06719629  0.04106422   1.636             0.101762    
## RETCALLS     0.12792241  0.18076459   0.708             0.479148    
## RETACCPT    -0.21506492  0.10486291  -2.051             0.040275 *  
## NEWCELLY    -0.06271686  0.02701597  -2.321             0.020261 *  
## NEWCELLN     0.00597326  0.03124518   0.191             0.848389    
## REFER       -0.04214403  0.04158918  -1.013             0.310897    
## INCOME      -0.00887801  0.00535208  -1.659             0.097157 .
## MCYCLE       0.11756345  0.08873256   1.325             0.185198
## CREDITAD    -0.07144040  0.03218642  -2.220             0.026447 *
## RETCALL      0.76622453  0.19145236   4.002  0.00006276465517728 ***
## REVENUE      0.00158016  0.00088365   1.788             0.073739 .
## MOU         -0.00030404  0.00005119  -5.940  0.00000000285642427 ***
## RECCHRGE    -0.00303904  0.00096393  -3.153             0.001617 **
## DIRECTAS    -0.00407055  0.00744838  -0.547             0.584721
## OVERAGE      0.00171978  0.00033882   5.076  0.00000038596432138 ***
## ROAM         0.01866662  0.00378488   4.932  0.00000081436358561 ***
## CHANGEM     -0.00056049  0.00005969  -9.389 < 0.0000000000000002 ***
## CHANGER      0.00229808  0.00044862   5.123  0.00000030135245052 ***
## DROPVCE     -0.00156940  0.00621447  -0.253             0.800623
## BLCKVCE     -0.00796185  0.00624404  -1.275             0.202269
## UNANSVCE     0.00071902  0.00054319   1.324             0.185606
## CUSTCARE    -0.00700046  0.00352072  -1.988             0.046772 *
## THREEWAY    -0.06585448  0.01894680  -3.476             0.000509 ***
## MOUREC       0.00007716  0.00015110   0.511             0.609595
## OUTCALLS     0.00092503  0.00067423   1.372             0.170073
## INCALLS     -0.00339497  0.00127735  -2.658             0.007864 **
## PEAKVCE     -0.00097153  0.00024390  -3.983  0.00006795394138340 ***
## OPEAKVCE    -0.00002309  0.00030344  -0.076             0.939351
## DROPBLK      0.01188970  0.00571431   2.081             0.037462 *
## CALLFWDV             NA          NA      NA                   NA
## CALLWAIT     0.00391417  0.00456370   0.858             0.391073
## SETPRC       0.00113264  0.00023878   4.744  0.00000210048243860 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55452  on 39999  degrees of freedom
## Residual deviance: 53579  on 39931  degrees of freedom
## AIC: 53717
##
## Number of Fisher Scoring iterations: 4
```

The first model has a lot of insignificant variables and they are removed

```
form <- churndep~MONTHS+UNIQSUBS+ACTVSUBS+PHONES+EQPDAYS+AGE1+AGE2+CHILDREN+CREDITB+CREDITDE+
PRIZMUB+REFURB+
  WEBCAP+MARRYUN+RETACCPT+NEWCELLY+INCOME+CREDITAD+RETCALL+REVENUE+MOU+RECCHRGE+OVERAGE+ROAM+
CHANGEM+CHANGER+
  CUSTCARE+THREEWAY+INCALLS+PEAKVCE+DROPBLK+CALLWAIT+SETPRC

fit2 <- glm(form,data = calib,family = binomial(logit),maxit=100)
summary(fit2)
```

```
## 
## Call:
## glm(formula = form, family = binomial(logit), data = calib, maxit = 100)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -2.257  -1.138  -0.261   1.148   2.062  
## 
## Coefficients:
##                Estimate  Std. Error z value             Pr(>|z|)    
## (Intercept) -0.03769762  0.06780873  -0.556             0.578251    
## MONTHS      -0.01866102  0.00171413 -10.887 < 0.0000000000000002 ***
## UNIQSUBS     0.19051039  0.01969140   9.675 < 0.0000000000000002 ***
## ACTVSUBS    -0.19187257  0.02734755  -7.016  0.00000000000228179 ***
## PHONES       0.06239568  0.01237019   5.044  0.00000045581480257 ***
## EQPDAYS      0.00140676  0.00006625  21.233 < 0.0000000000000002 ***
## AGE1        -0.00340841  0.00076544  -4.453  0.00000847279023403 ***
## AGE2        -0.00128121  0.00060952  -2.102             0.035554 *  
## CHILDREN     0.09310788  0.02645527   3.519             0.000432 ***
## CREDITB      0.07398881  0.02794196   2.648             0.008098 ** 
## CREDITDE    -0.29266569  0.03386054  -8.643 < 0.0000000000000002 ***
## PRIZMUB     -0.05660242  0.02233289  -2.534             0.011261 *  
## REFURB       0.24602956  0.03099146   7.939  0.00000000000000204 ***
## WEBCAP      -0.14267956  0.03705058  -3.851             0.000118 ***
## MARRYUN      0.05917633  0.02788713   2.122             0.033838 *  
## RETACCPT    -0.18308311  0.09643542  -1.899             0.057630 .  
## NEWCELLY    -0.06534631  0.02623268  -2.491             0.012737 *  
## INCOME      -0.00319424  0.00454322  -0.703             0.482006    
## CREDITAD    -0.07322894  0.03207095  -2.283             0.022410 *  
## RETCALL      0.89038346  0.07574694  11.755 < 0.0000000000000002 ***
## REVENUE      0.00138901  0.00087070   1.595             0.110651    
## MOU         -0.00026508  0.00004312  -6.147  0.00000000078731273 ***
## RECCHRGE    -0.00304092  0.00095358  -3.189             0.001428 ** 
## OVERAGE      0.00177922  0.00033583   5.298  0.00000011712981230 ***
## ROAM         0.01928514  0.00376190   5.126  0.00000029528119040 ***
## CHANGEM     -0.00056525  0.00005945  -9.509 < 0.0000000000000002 ***
## CHANGER      0.00230290  0.00044803   5.140  0.00000027460210269 ***
## CUSTCARE    -0.00656210  0.00340173  -1.929             0.053725 .  
## THREEWAY    -0.06810424  0.01876090  -3.630             0.000283 ***
## INCALLS     -0.00214452  0.00104096  -2.060             0.039386 *  
## PEAKVCE     -0.00070177  0.00021163  -3.316             0.000913 ***
## DROPBLK      0.00736283  0.00114870   6.410  0.00000000014578637 ***
## CALLWAIT     0.00297571  0.00425007   0.700             0.483829    
## SETPRC       0.00117816  0.00023774   4.956  0.00000072097597927 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 55452  on 39999  degrees of freedom
## Residual deviance: 53651  on 39966  degrees of freedom
## AIC: 53719
## 
## Number of Fisher Scoring iterations: 4
```

The second model is relatively better, but still has some variables that can be removed

```
form <- churndep~MONTHS+UNIQSUBS+ACTVSUBS+PHONES+EQPDAYS+AGE1+AGE2+CHILDREN+CREDITB+CREDITDE+
PRIZMUB+REFURB+
  WEBCAP+MARRYUN+RETACCPT+NEWCELLY+CREDITAD+RETCALL+MOU+RECCHRGE+OVERAGE+ROAM+CHANGEM+CHANGER
+
  CUSTCARE+THREEWAY+INCALLS+PEAKVCE+DROPBLK+SETPRC

fit3 <- glm(form,data = calib,family = binomial(logit),maxit=100)
summary(fit3)
```

```
##
## Call:
## glm(formula = form, family = binomial(logit), data = calib, maxit = 100)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2426  -1.1389  -0.2519   1.1479   2.0799
##
## Coefficients:
##                Estimate  Std. Error z value             Pr(>|z|)
## (Intercept) -0.04692064  0.06691508  -0.701             0.483180
## MONTHS      -0.01852491  0.00170817 -10.845 < 0.0000000000000002 ***
## UNIQSUBS     0.19053070  0.01968888   9.677 < 0.0000000000000002 ***
## ACTVSUBS    -0.19262130  0.02733295  -7.047  0.00000000000182527 ***
## PHONES       0.06285399  0.01236813   5.082  0.00000037361566925 ***
## EQPDAYS      0.00140462  0.00006621  21.214 < 0.0000000000000002 ***
## AGE1        -0.00359033  0.00072024  -4.985  0.00000061993003171 ***
## AGE2        -0.00127365  0.00060935  -2.090             0.036601 *
## CHILDREN     0.09100727  0.02632192   3.457             0.000545 ***
## CREDITB      0.07513667  0.02792079   2.691             0.007122 **
## CREDITDE    -0.28987193  0.03379231  -8.578 < 0.0000000000000002 ***
## PRIZMUB     -0.05804302  0.02215333  -2.620             0.008791 **
## REFURB       0.24633174  0.03098863   7.949  0.00000000000000188 ***
## WEBCAP      -0.14451762  0.03703098  -3.903  0.00009515933871974 ***
## MARRYUN      0.06538844  0.02653509   2.464             0.013731 *
## RETACCPT    -0.18437658  0.09646207  -1.911             0.055955 .
## NEWCELLY    -0.06517264  0.02622738  -2.485             0.012958 *
## CREDITAD    -0.07503448  0.03206659  -2.340             0.019286 *
## RETCALL      0.89391694  0.07572810  11.804 < 0.0000000000000002 ***
## MOU         -0.00025428  0.00004192  -6.066  0.00000000131350874 ***
## RECCHRGE    -0.00190088  0.00061116  -3.110             0.001869 **
## OVERAGE      0.00223515  0.00017890  12.494 < 0.0000000000000002 ***
## ROAM         0.02144451  0.00348208   6.159  0.00000000073422095 ***
## CHANGEM     -0.00056680  0.00005941  -9.541 < 0.0000000000000002 ***
## CHANGER      0.00231065  0.00044789   5.159  0.00000024828813349 ***
## CUSTCARE    -0.00656428  0.00339736  -1.932             0.053338 .
## THREEWAY    -0.06736815  0.01874713  -3.594             0.000326 ***
## INCALLS     -0.00207598  0.00102582  -2.024             0.042997 *
## PEAKVCE     -0.00064332  0.00020333  -3.164             0.001557 **
## DROPBLK      0.00741673  0.00114681   6.467  0.00000000009978353 ***
## SETPRC       0.00118736  0.00023752   4.999  0.00000057642749538 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55452  on 39999  degrees of freedom
## Residual deviance: 53655  on 39969  degrees of freedom
## AIC: 53717
##
## Number of Fisher Scoring iterations: 4
```

fit3 is a better model to use as all variables are significant

To validae this model a few analysis is done. Firstly the VIF values are looked to see if there is any high multicollinearity

```
vif(fit3)
```

```
##   MONTHS UNIQSUBS ACTVSUBS   PHONES  EQPDAYS     AGE1     AGE2 CHILDREN 
## 2.511806 2.806551 2.870435 2.488556 2.661172 2.369432 1.976684 1.225315 
##  CREDITB CREDITDE  PRIZMUB   REFURB   WEBCAP  MARRYUN RETACCPT NEWCELLY 
## 1.040380 1.116153 1.019042 1.129605 1.200934 1.598537 1.850868 1.020101 
## CREDITAD  RETCALL      MOU RECCHRGE  OVERAGE     ROAM  CHANGEM  CHANGER 
## 1.078652 1.857041 3.991526 1.676026 1.680269 1.036787 1.649035 1.633300 
## CUSTCARE THREEWAY  INCALLS  PEAKVCE  DROPBLK   SETPRC 
## 1.375668 1.271260 1.780661 3.324966 1.968691 1.648399
```

VIF of all values are below required levels

Calculating the psudo R-squared value of Logistic regression

```
(pR2 <- 1- fit3$deviance/fit3$null.deviance)
```

```
## [1] 0.03240832
```

The psudo R-squared value for this model is very low

The optimal cutoff value of the probabilities is calculated

```
predicted <- predict(fit3, calib, type="response")
cutoff <- optimalCutoff(calib$churndep,predicted)
cutoff
```

```
## [1] 0.49
```

Getting the confusion matrix

```
ConfMat <- InformationValue::confusionMatrix(calib$churndep, predicted, threshold = cutoff)
ConfMat <- as.data.frame(ConfMat,row.names = c("Actual NO","Actual YES"))
colnames(ConfMat) <- c("Predicted NO","Predicted YES")
ConfMat
```

	Predicted NO	**Predicted YES**
	<int>	<int>
Actual NO	11082	7577
Actual YES	8918	12423

2 rows

Testing the model accuracy

```
InformationValue::sensitivity(calib$churndep, predicted, threshold = cutoff)
```

```
## [1] 0.62115
```

```
InformationValue::specificity(calib$churndep, predicted, threshold = cutoff)
```

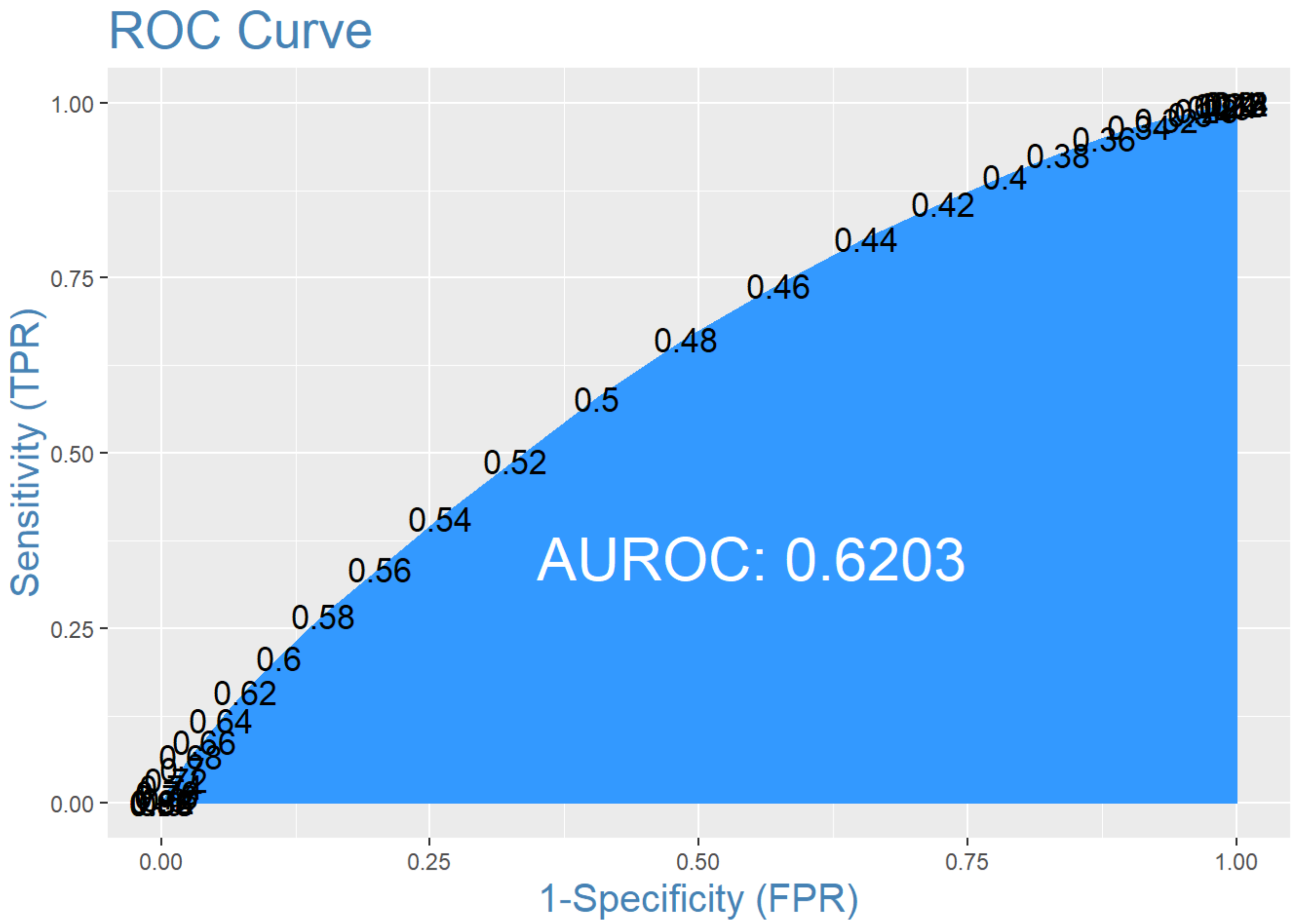
```
## [1] 0.5541
```

```
InformationValue::misClassError(calib$churndep, predicted, threshold = cutoff)
```

```
## [1] 0.4124
```

```
InformationValue::plotROC(calib$churndep, predicted,Show.labels = T)
```

```
## Warning: Removed 10 rows containing missing values (geom_text).
```



In this model it is seen that both sensitivity and specificity are greater than 0.5 but the accuracy is still low as the values are still below the desierable 0.7. The misclassification error is high, but not bad to make the model invalid.

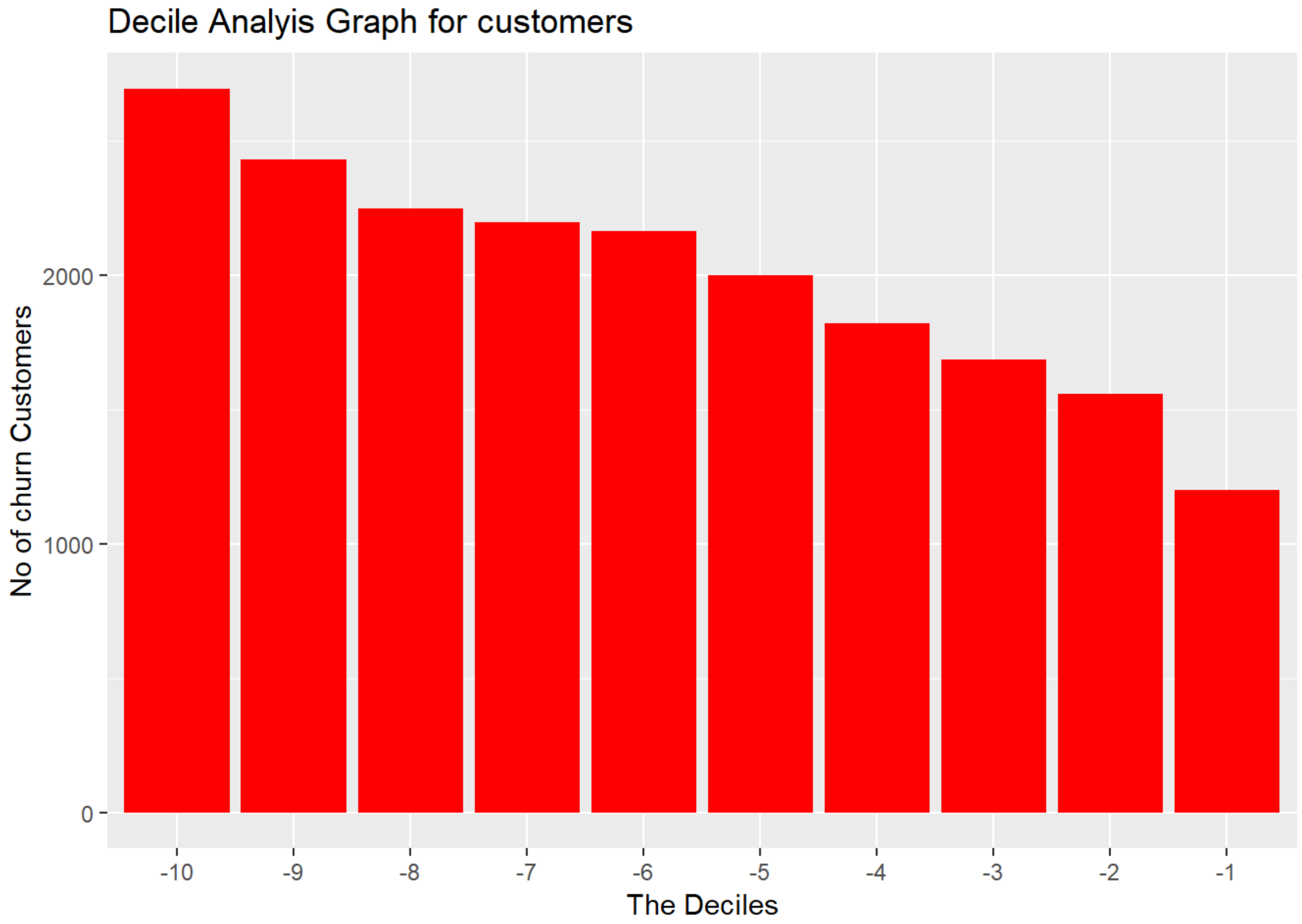
The AUROC is 0.62 which is still less than a 0.7 desierable level.

Conducting decile analysis

```
calib <- cbind(calib,pred = predicted)

decLocations <- quantile(calib$pred, probs = seq(0.1,0.9,by=0.1))
calib$decile <- findInterval(calib$pred,c(-Inf,decLocations, Inf))

calib <- data.table(calib)

fit_train_DA <- calib %>% group_by(decile) %>% dplyr::summarize(Min_prob = min(pred),
                                                                 Max_prob = max(pred),
                                                                 churn_Count = sum(churndep),
                                                                 Non_churn_Count = (length(de
cile)-sum(churndep))
                                                                ) %>% arrange(decile)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
fit_train_DA <- dplyr::arrange(fit_train_DA, desc(decile))

DecilePlot <-  ggplot2::ggplot(data = fit_train_DA,aes(x = factor(-decile), y = churn_Count))
DecilePlot <- DecilePlot + geom_bar(stat = 'identity',fill = "red")
DecilePlot <- DecilePlot + xlab("The Deciles") + ylab("No of churn Customers")
DecilePlot <- DecilePlot + ggtitle("Decile Analyis Graph for customers")
DecilePlot
```



The declile plot shows that the deciles do form step shape and can be concluded that the model is still suitable to use for prediction. It is advisable to run Machine Learning algorithms to get better accuracy for this data or new data should be collected and this model should be rebuilt.

```
calib$pred_churn <- NULL
```

```
## Warning in `[<-.data.table`(x, j = name, value = value): Adding new column
## 'pred_churn' then assigning NULL (deleting it).
```

```
calib$pred_churn <- ifelse(calib$pred>cutoff, 1,0)

valid<-cbind(valid, Pred=predict(fit3, valid, type="response"))
valid$churn <- ifelse(valid$Pred>cutoff, 1,0)
sum(valid$churn)
```

```
## [1] 13888
```

This model says that there are 13,888 customers who are about to churn.

There are around 30 factors that affect customer attrition. The top 5factors that are increase and decrease the churn are given below.

Factors that increase the odds of churn are RETCALL (No of calls made to retention team), REFURB (Refurbished handset), UNIQSUBS(Number of unique subscriptions), CHILDREN(Presense of children), CREDITB (Credt rating is B). Of these factors the relevance of REFURB and CHILDREN as strong factors is questionable as they are not logically strongly related to churning

The factors that decrease the odds of churn are CREDITAD (Credit Adjustments), WEBCAP(Web capable handset), RETACCPT(Retention offers accepted), ACTVSUBS (Number of active subscriptions), CREDITDE (Low credit rating DE). All these factors mentioned here logically make sense as a strong factor in deciding decrease in Churn.

In terms of the offers that can be made to increase revenue, customers who currently own a refurbished phone can exchange that phone and avail special discount offers to buy a new handset that is web capable. When calls are made to the retention team, a special subscription for the customer based on their use should be provided. This can be made as a flexible plan for them. Post implementing this offer, the difference in the churn amount between people who did not avail this offer and people who accepted the offer must be studied. A chi square analysis can be conducted in this case to see if those who accepted the offer and those who did not accept differ by churn category.