DATA 621 – Business Analytics and Data Mining

Abdelmalek Hajjam/ Monu Chacko

4/26/2020

In this homework assignment, you will explore, analyze and model a dataset containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first responsevariable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash.The second responsevariable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

```
library(ggcorrplot)
library(car)
library(MASS)
library(dplyr)
library(ggplot2)
library(caret)
library(pROC)
library(pscl)
library(psych)
library(data.table)
```

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

DATA EXPLORATION

```
# Read data
train_df <- read.csv("https://raw.githubusercontent.com/monuchacko/cuny_msds/master/data_621/Homework4/

dim(train_df)
```

```
## [1] 8161   26
```

```
str(train_df)
```

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/
```

```
head(train_df)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14               No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION          JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M           PhD  Professional       14    Private  $14,230
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial  $14,940
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private   $4,010
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private  $15,440
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private  $18,000
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial  $17,430
##   TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11    Minivan     yes   $4,461        2      No       3      18
## 2   1    Minivan     yes       $0        0      No       0       1
## 3   4      z_SUV      no  $38,690        2      No       3      10
## 4   7    Minivan     yes       $0        0      No       0       6
## 5   1      z_SUV      no  $19,217        2     Yes       3      17
## 6   1 Sports Car      no       $0        0      No       0       7
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
```

```
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

```
# Exclude the INDEX column
tr <- train_df[-1]

#  Convert to numeric
tr$INCOME <- as.numeric(gsub('[$,]', '', tr$INCOME))
tr$HOME_VAL <- as.numeric(gsub('[$,]', '', tr$HOME_VAL))
tr$BLUEBOOK <- as.numeric(gsub('[$,]', '', tr$BLUEBOOK))
tr$OLDCLAIM <- as.numeric(gsub('[$,]', '', tr$OLDCLAIM))

# Remove characters that are not required
tr$MSTATUS  <- gsub("z_", "", tr$MSTATUS)
tr$SEX  <- gsub("z_", "", tr$SEX)
tr$EDUCATION  <- gsub("z_", "", tr$EDUCATION)
tr$JOB  <- gsub("z_", "", tr$JOB)
tr$CAR_USE  <- gsub("z_", "", tr$CAR_USE)
tr$CAR_TYPE  <- gsub("z_", "", tr$CAR_TYPE)
tr$URBANICITY  <- gsub("z_", "", tr$URBANICITY)


# Reorder columns -- predictor categorical, predictor numeric, target
indx <- c(8, 10:13, 15, 18:19, 22, 25, 3:7, 9, 14, 16:17, 20:21, 23:24, 1:2)
tr_ordered <- tr
setcolorder(tr_ordered,indx)
```

```
table(tr$PARENT1)
```

Examine the data

```
##
##   No  Yes
## 7084 1077
```

```
table(tr$MSTATUS)
```

```
##
##   No  Yes
## 3267 4894
```

```
table(tr$SEX)
```

```
##
##    F    M
## 4375 3786
```

```
table(tr$EDUCATION)
```

```
##
## <High School    Bachelors  High School      Masters          PhD
##         1203         2242         2330         1658          728
```

```
table(tr$JOB)
```

```
##
##               Blue Collar     Clerical       Doctor   Home Maker       Lawyer
##          526         1825         1271          246          641          835
##      Manager Professional      Student
##          988         1117          712
```

```
table(tr$CAR_USE)
```

```
##
## Commercial    Private
##       3029       5132
```

```
table(tr$CAR_TYPE)
```

```
##
##     Minivan Panel Truck      Pickup  Sports Car         SUV         Van
##        2145         676        1389         907        2294         750
```

```
table(tr$RED_CAR)
```

```
##
##   no  yes
## 5783 2378
```

```
table(tr$REVOKED)
```

```
##
##   No  Yes
## 7161 1000
```

```
table(tr$URBANICITY)
```

```
##
## Highly Rural/ Rural Highly Urban/ Urban
##                1669                6492
```

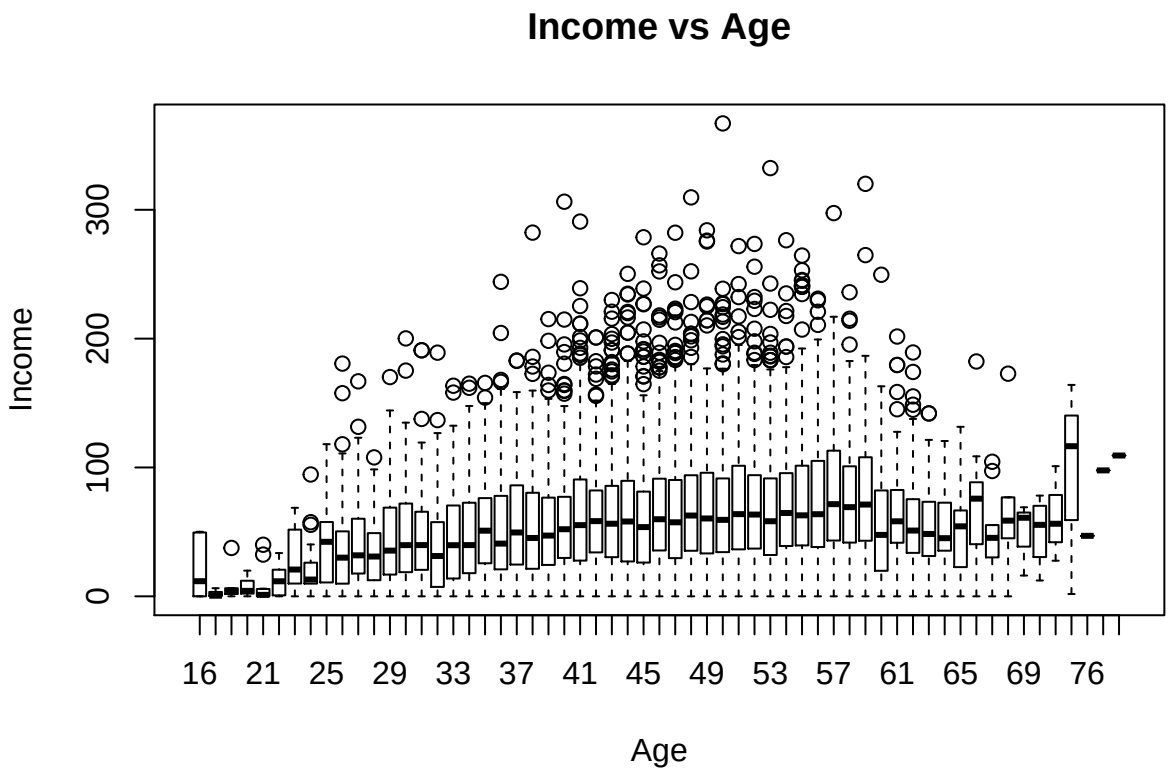
```r
summary(tr)
```

Summary of the data

```
##    PARENT1             MSTATUS              SEX             EDUCATION
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      JOB               CAR_USE            CAR_TYPE           RED_CAR
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    REVOKED            URBANICITY           KIDSDRIV           AGE
##  Length:8161        Length:8161        Min.   :0.0000   Min.   :16.00
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:39.00
##  Mode  :character   Mode  :character   Median :0.0000   Median :45.00
##                                        Mean   :0.1711   Mean   :44.79
##                                        3rd Qu.:0.0000   3rd Qu.:51.00
##                                        Max.   :4.0000   Max.   :81.00
##                                                         NA's   :6
##     HOMEKIDS           YOJ           INCOME          HOME_VAL
##  Min.   :0.0000   Min.   : 0.0   Min.   :     0   Min.   :     0
##  1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   1st Qu.:     0
##  Median :0.0000   Median :11.0   Median : 54028   Median :161160
##  Mean   :0.7212   Mean   :10.5   Mean   : 61898   Mean   :154867
##  3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986   3rd Qu.:238724
##  Max.   :5.0000   Max.   :23.0   Max.   :367030   Max.   :885282
##                   NA's   :454    NA's   :445      NA's   :464
##     TRAVTIME         BLUEBOOK          TIF            OLDCLAIM
##  Min.   :  5.00   Min.   : 1500   Min.   : 1.000   Min.   :    0
##  1st Qu.: 22.00   1st Qu.: 9280   1st Qu.: 1.000   1st Qu.:    0
##  Median : 33.00   Median :14440   Median : 4.000   Median :    0
##  Mean   : 33.49   Mean   :15710   Mean   : 5.351   Mean   : 4037
##  3rd Qu.: 44.00   3rd Qu.:20850   3rd Qu.: 7.000   3rd Qu.: 4636
##  Max.   :142.00   Max.   :69740   Max.   :25.000   Max.   :57037
##
##     CLM_FREQ         MVR_PTS          CAR_AGE        TARGET_FLAG
##  Min.   :0.0000   Min.   : 0.000   Min.   :-3.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 1.000   1st Qu.:0.0000
##  Median :0.0000   Median : 1.000   Median : 8.000   Median :0.0000
##  Mean   :0.7986   Mean   : 1.696   Mean   : 8.328   Mean   :0.2638
##  3rd Qu.:2.0000   3rd Qu.: 3.000   3rd Qu.:12.000   3rd Qu.:1.0000
##  Max.   :5.0000   Max.   :13.000   Max.   :28.000   Max.   :1.0000
##                                    NA's   :510
```

```
##    TARGET_AMT
##  Min.   :     0
##  1st Qu.:     0
##  Median :     0
##  Mean   :  1504
##  3rd Qu.:  1036
##  Max.   :107586
##
```

```
boxplot((INCOME/1000)~AGE,data=tr, main="Income vs Age", xlab="Age", ylab="Income")
```



DATA PREPARATION

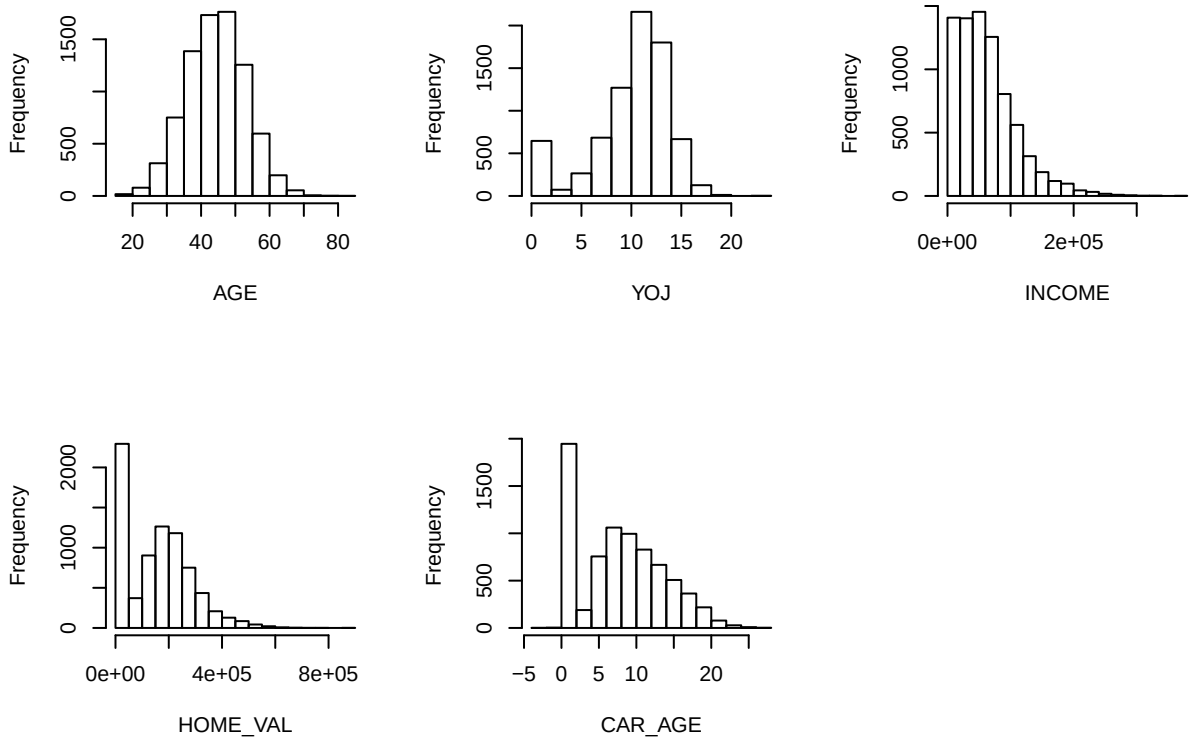
```
tr_prep <- tr_ordered
```

```
M <- sapply(tr_prep, function(x) sum(x=="") | sum(is.na(x))); names(M[(M>0)])
```

```
## [1] "JOB"      "AGE"      "YOJ"      "INCOME"   "HOME_VAL" "CAR_AGE"
```

```
x <- c(12, 14, 15, 16, 23)
par(mfrow=c(2,3))
for (val in x) {
```

```
  hist(tr_prep[,val],xlab=names(tr_prep[val]), main="")
}
par(mfrow=c(1,1))
```




```
#impute

tr_prep = tr_prep %>%
  mutate(AGE =
           ifelse(is.na(AGE),
                  mean(AGE, na.rm=TRUE), AGE)) %>%

  mutate(YOJ =
           ifelse(is.na(YOJ),
                  mean(YOJ, na.rm=TRUE), YOJ)) %>%

  mutate(INCOME =
           ifelse(is.na(INCOME),
                  median(INCOME, na.rm=TRUE), INCOME)) %>%

  mutate(HOME_VAL =
           ifelse(is.na(HOME_VAL),
                  mean(HOME_VAL, na.rm=TRUE), HOME_VAL)) %>%

  mutate(CAR_AGE =
           ifelse(is.na(CAR_AGE),
```

```
                  mean(CAR_AGE, na.rm=TRUE), CAR_AGE)) %>%

  mutate(JOB =
           ifelse((JOB == "" & EDUCATION == 'PhD'),
                  "Doctor", JOB)) %>%

  mutate(JOB =
           ifelse((JOB == "" & EDUCATION == 'Masters'),
                  "Lawyer", JOB))
```

```
M <- sapply(tr_prep, function(x) sum(x=="") | sum(is.na(x))); names(M[(M>0)])
```

```
## character(0)
```

```
# Outlier Capping

tr_prep2 <- tr_prep

id <- c(11:23)
for (val in id) {
  qnt <- quantile(tr_prep2[,val], probs=c(.25, .75), na.rm = T)
  caps <- quantile(tr_prep2[,val], probs=c(.05, .95), na.rm = T)
  H <- 1.5 * IQR(tr_prep2[,val], na.rm = T)
  tr_prep2[,val][tr_prep2[,val] < (qnt[1] - H)] <- caps[1]
  tr_prep2[,val][tr_prep2[,val] > (qnt[2] + H)] <- caps[2]

}
```

BUILD MODELS

```
nTrain <- createDataPartition(tr_prep2$TARGET_FLAG, p=0.8, list=FALSE)
ntraining <- tr_prep2[nTrain,]
ntesting <- tr_prep2[-nTrain,]

set.seed(123)
```

```
# Logistic Regression build the model using training set
full.model_FLAG <- glm(TARGET_FLAG ~.-TARGET_AMT, data = ntraining , family = binomial)
summary(full.model_FLAG)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial,
##     data = ntraining)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2800  -0.7113  -0.3983   0.6319   3.1416
##
## Coefficients:
```

```
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.308e+00  3.204e-01  -7.202 5.92e-13 ***
## PARENT1Yes                    2.964e-01  1.255e-01   2.361 0.018228 *
## MSTATUSYes                   -5.299e-01  9.430e-02  -5.619 1.92e-08 ***
## SEXM                          1.292e-02  1.239e-01   0.104 0.916919
## EDUCATIONBachelors           -3.534e-01  1.297e-01  -2.725 0.006428 **
## EDUCATIONHigh School          4.856e-02  1.062e-01   0.457 0.647466
## EDUCATIONMasters             -5.005e-01  2.095e-01  -2.389 0.016900 *
## EDUCATIONPhD                 -1.111e-01  2.572e-01  -0.432 0.665797
## JOBClerical                   9.296e-02  1.176e-01   0.791 0.429217
## JOBDoctor                    -6.734e-01  2.807e-01  -2.399 0.016452 *
## JOBHome Maker                -1.386e-01  1.726e-01  -0.803 0.422134
## JOBLawyer                     5.765e-02  1.996e-01   0.289 0.772736
## JOBManager                   -8.095e-01  1.570e-01  -5.157 2.51e-07 ***
## JOBProfessional              -1.217e-01  1.335e-01  -0.912 0.361697
## JOBStudent                   -2.958e-01  1.474e-01  -2.007 0.044798 *
## CAR_USEPrivate               -7.926e-01  9.800e-02  -8.088 6.08e-16 ***
## CAR_TYPEPanel Truck           6.656e-01  1.775e-01   3.750 0.000177 ***
## CAR_TYPEPickup                5.070e-01  1.118e-01   4.537 5.71e-06 ***
## CAR_TYPESports Car            8.992e-01  1.464e-01   6.143 8.10e-10 ***
## CAR_TYPESUV                   6.506e-01  1.243e-01   5.234 1.66e-07 ***
## CAR_TYPEVan                   5.844e-01  1.419e-01   4.119 3.81e-05 ***
## RED_CARyes                    4.498e-02  9.573e-02   0.470 0.638446
## REVOKEDYes                    8.701e-01  1.023e-01   8.509  < 2e-16 ***
## URBANICITYHighly Urban/ Urban 2.331e+00  1.238e-01  18.830  < 2e-16 ***
## KIDSDRIV                      5.772e-01  1.096e-01   5.268 1.38e-07 ***
## AGE                          -6.159e-04  4.588e-03  -0.134 0.893209
## HOMEKIDS                      8.405e-02  4.442e-02   1.892 0.058486 .
## YOJ                          -1.369e-02  9.611e-03  -1.424 0.154417
## INCOME                       -4.078e-06  1.360e-06  -2.998 0.002714 **
## HOME_VAL                     -1.533e-06  3.856e-07  -3.974 7.06e-05 ***
## TRAVTIME                      1.609e-02  2.181e-03   7.379 1.59e-13 ***
## BLUEBOOK                     -2.863e-05  6.105e-06  -4.689 2.74e-06 ***
## TIF                          -5.665e-02  8.609e-03  -6.581 4.69e-11 ***
## OLDCLAIM                     -1.548e-05  5.376e-06  -2.879 0.003984 **
## CLM_FREQ                      1.867e-01  3.283e-02   5.687 1.29e-08 ***
## MVR_PTS                       1.286e-01  1.651e-02   7.790 6.70e-15 ***
## CAR_AGE                      -3.377e-03  8.464e-03  -0.399 0.689863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7527.5  on 6528  degrees of freedom
## Residual deviance: 5819.8  on 6492  degrees of freedom
## AIC: 5893.8
##
## Number of Fisher Scoring iterations: 5
```

```
round(exp(cbind(Estimate=coef(full.model_FLAG))),2)
```

```
##                              Estimate
## (Intercept)                      0.10
## PARENT1Yes                       1.34
```

```
## MSTATUSYes                          0.59
## SEXM                                1.01
## EDUCATIONBachelors                  0.70
## EDUCATIONHigh School                1.05
## EDUCATIONMasters                    0.61
## EDUCATIONPhD                        0.89
## JOBClerical                         1.10
## JOBDoctor                           0.51
## JOBHome Maker                       0.87
## JOBLawyer                           1.06
## JOBManager                          0.45
## JOBProfessional                     0.89
## JOBStudent                          0.74
## CAR_USEPrivate                      0.45
## CAR_TYPEPanel Truck                 1.95
## CAR_TYPEPickup                      1.66
## CAR_TYPESports Car                  2.46
## CAR_TYPESUV                         1.92
## CAR_TYPEVan                         1.79
## RED_CARyes                          1.05
## REVOKEDYes                          2.39
## URBANICITYHighly Urban/ Urban      10.29
## KIDSDRIV                            1.78
## AGE                                 1.00
## HOMEKIDS                            1.09
## YOJ                                 0.99
## INCOME                              1.00
## HOME_VAL                            1.00
## TRAVTIME                            1.02
## BLUEBOOK                            1.00
## TIF                                 0.94
## OLDCLAIM                            1.00
## CLM_FREQ                            1.21
## MVR_PTS                             1.14
## CAR_AGE                             1.00
```

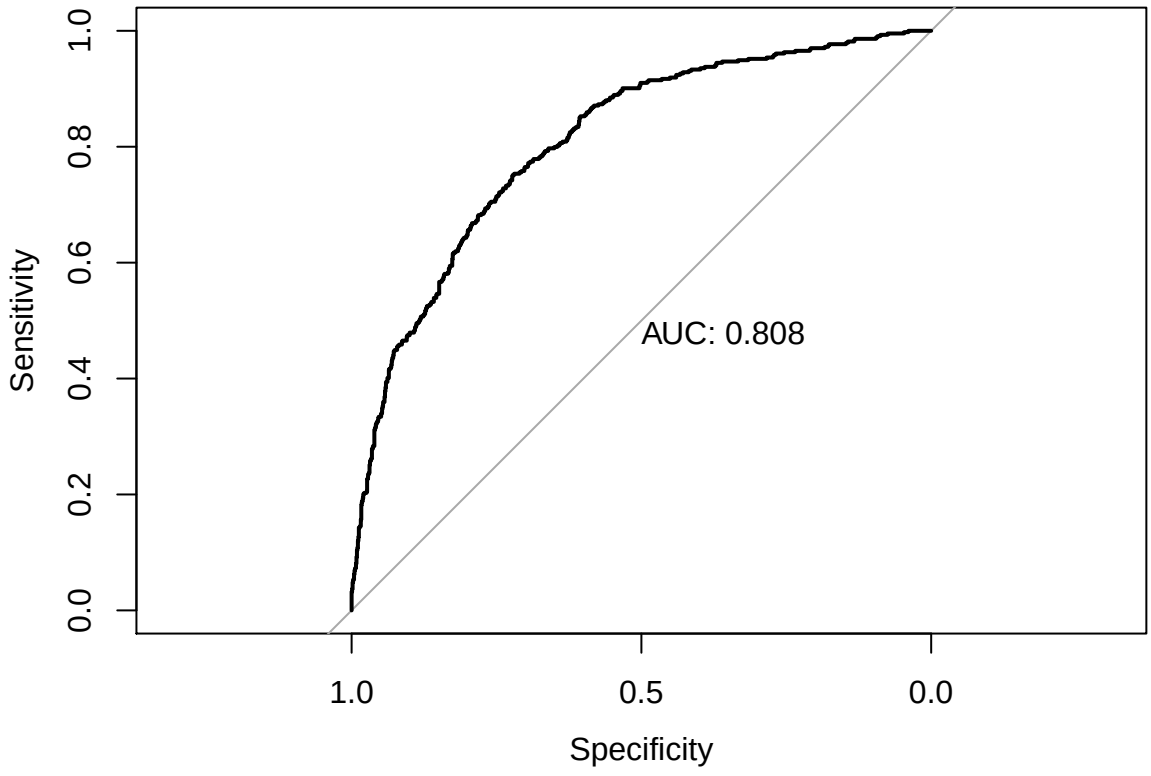
```r
# evaluate the model by predicting using the testing set
m1_prob <- predict(full.model_FLAG, ntesting, type = "response")
m1_pclass <- ifelse(m1_prob >= 0.5, 1, 0)

# create confusion matrix
pclass <- factor(m1_pclass,levels = c(1,0))
aclass <- factor(ntesting$TARGET_FLAG,levels = c(1,0))
confusionMatrix(pclass, aclass);
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  184   82
##          0  250 1116
##
##                Accuracy : 0.7966
##                  95% CI : (0.7762, 0.8159)
```

```
##     No Information Rate : 0.7341
##     P-Value [Acc > NIR] : 2.55e-09
##
##                   Kappa : 0.4056
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4240
##             Specificity : 0.9316
##          Pos Pred Value : 0.6917
##          Neg Pred Value : 0.8170
##              Prevalence : 0.2659
##          Detection Rate : 0.1127
##    Detection Prevalence : 0.1630
##       Balanced Accuracy : 0.6778
##
##        'Positive' Class : 1
##
```

```r
# plot and show area under the curve
plot(roc(ntesting$TARGET_FLAG, m1_prob),print.auc=TRUE)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



```r
# get McFadden
m1_mcFadden <- pR2(full.model_FLAG); m1_mcFadden["McFadden"]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.2268554
```

```r
full.model.AMT <- lm(TARGET_AMT ~. -TARGET_FLAG,  data = tr_prep2)
summary(full.model.AMT)
```

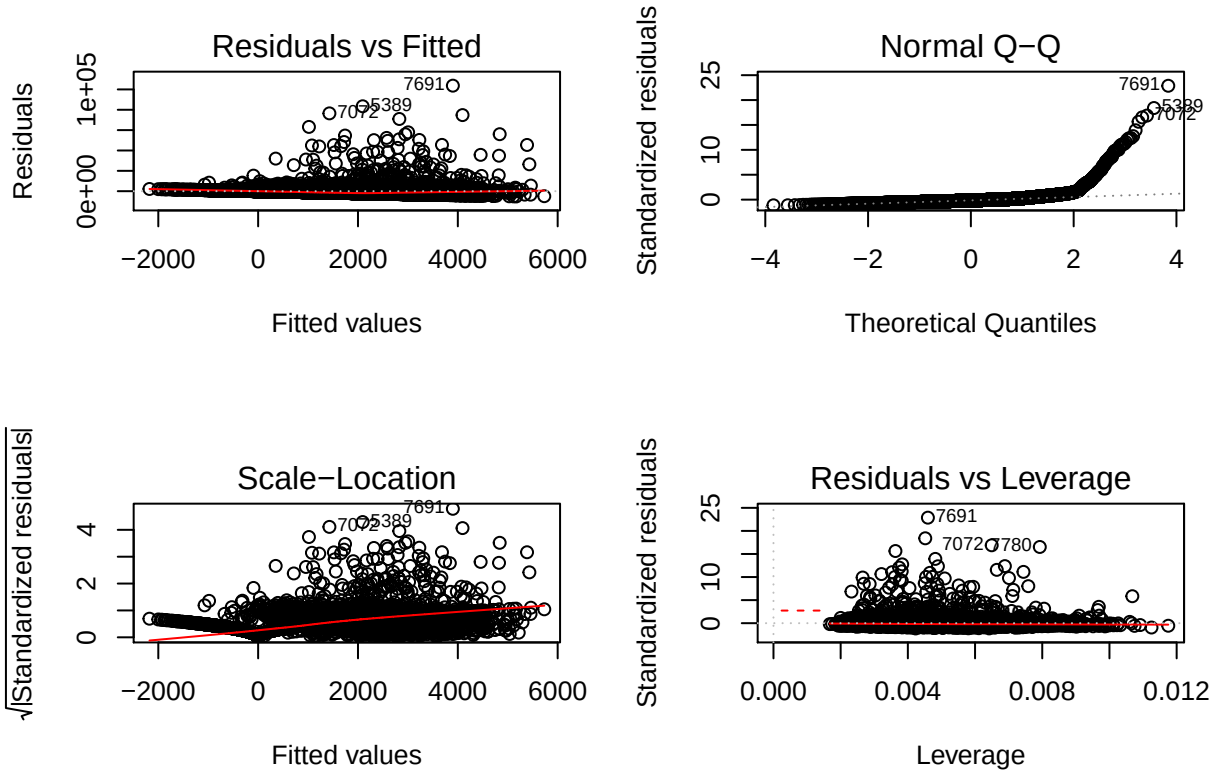
```
##
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG, data = tr_prep2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5169  -1704   -754    344 103686
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   9.821e+01  4.779e+02   0.205 0.837204
## PARENT1Yes                    5.364e+02  2.043e+02   2.626 0.008654 **
## MSTATUSYes                   -5.759e+02  1.453e+02  -3.963 7.47e-05 ***
## SEXM                          3.471e+02  1.833e+02   1.894 0.058294 .
## EDUCATIONBachelors           -2.659e+02  2.055e+02  -1.294 0.195735
## EDUCATIONHigh School         -9.582e+01  1.716e+02  -0.558 0.576571
## EDUCATIONMasters             -4.219e+01  3.061e+02  -0.138 0.890376
## EDUCATIONPhD                  3.898e+02  3.757e+02   1.038 0.299462
## JOBClerical                   1.773e+01  1.917e+02   0.093 0.926295
## JOBDoctor                    -1.046e+03  4.009e+02  -2.608 0.009117 **
## JOBHome Maker                -1.745e+02  2.684e+02  -0.650 0.515478
## JOBLawyer                    -2.484e+02  2.928e+02  -0.848 0.396242
## JOBManager                   -9.855e+02  2.328e+02  -4.233 2.33e-05 ***
## JOBProfessional              -2.882e+01  2.117e+02  -0.136 0.891724
## JOBStudent                   -2.275e+02  2.358e+02  -0.965 0.334779
## CAR_USEPrivate               -8.000e+02  1.578e+02  -5.071 4.05e-07 ***
## CAR_TYPEPanel Truck           2.668e+02  2.724e+02   0.979 0.327500
## CAR_TYPEPickup                3.707e+02  1.706e+02   2.172 0.029869 *
## CAR_TYPESports Car            1.014e+03  2.178e+02   4.656 3.28e-06 ***
## CAR_TYPESUV                   7.434e+02  1.793e+02   4.145 3.43e-05 ***
## CAR_TYPEVan                   5.220e+02  2.127e+02   2.454 0.014162 *
## RED_CARyes                   -3.741e+01  1.491e+02  -0.251 0.801877
## REVOKEDYes                    5.906e+02  1.743e+02   3.387 0.000709 ***
## URBANICITYHighly Urban/ Urban 1.675e+03  1.392e+02  12.032  < 2e-16 ***
## KIDSDRIV                      6.104e+02  1.789e+02   3.411 0.000650 ***
## AGE                           6.643e+00  7.158e+00   0.928 0.353395
## HOMEKIDS                      8.003e+01  6.981e+01   1.146 0.251677
## YOJ                          -5.048e+00  1.503e+01  -0.336 0.737007
## INCOME                       -4.652e-03  2.081e-03  -2.236 0.025408 *
## HOME_VAL                     -6.471e-04  5.961e-04  -1.085 0.277733
## TRAVTIME                      1.285e+01  3.340e+00   3.847 0.000120 ***
```

```
## BLUEBOOK                      1.357e-02  8.954e-03   1.515 0.129769
## TIF                          -5.120e+01  1.294e+01  -3.958 7.62e-05 ***
## OLDCLAIM                     -1.685e-02  9.119e-03  -1.848 0.064627 .
## CLM_FREQ                      1.667e+02  5.686e+01   2.932 0.003381 **
## MVR_PTS                       1.731e+02  2.790e+01   6.203 5.79e-10 ***
## CAR_AGE                      -2.708e+01  1.286e+01  -2.105 0.035304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4544 on 8124 degrees of freedom
## Multiple R-squared:  0.0709, Adjusted R-squared:  0.06679
## F-statistic: 17.22 on 36 and 8124 DF,  p-value: < 2.2e-16
```

```
vif(full.model.AMT)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## PARENT1     1.888802  1        1.374337
## MSTATUS     2.003847  1        1.415573
## SEX         3.302562  1        1.817295
## EDUCATION  15.755173  4        1.411490
## JOB        29.731384  7        1.274177
## CAR_USE     2.296181  1        1.515316
## CAR_TYPE    5.324276  5        1.182023
## RED_CAR     1.813330  1        1.346599
## REVOKED     1.291470  1        1.136429
## URBANICITY  1.245363  1        1.115958
## KIDSDRIV    1.338024  1        1.156730
## AGE         1.462748  1        1.209441
## HOMEKIDS    2.140216  1        1.462948
## YOJ         1.416953  1        1.190358
## INCOME      2.863846  1        1.692290
## HOME_VAL    2.143594  1        1.464102
## TRAVTIME    1.034456  1        1.017082
## BLUEBOOK    2.032225  1        1.425561
## TIF         1.006054  1        1.003022
## OLDCLAIM    1.827781  1        1.351954
## CLM_FREQ    1.714439  1        1.309366
## MVR_PTS     1.229565  1        1.108858
## CAR_AGE     1.975294  1        1.405451
```

```
par(mfrow=c(2,2))
plot(full.model.AMT)
```




```
par(mfrow=c(1,1))
```

3.2 Stepwise variable selection

```
# Logistic Regression build the model using training set
step.model_FLAG <- full.model_FLAG %>% stepAIC(trace = FALSE)
summary(step.model_FLAG)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ PARENT1 + MSTATUS + EDUCATION + JOB +
##     CAR_USE + CAR_TYPE + REVOKED + URBANICITY + KIDSDRIV + HOMEKIDS +
##     YOJ + INCOME + HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM +
##     CLM_FREQ + MVR_PTS, family = binomial, data = ntraining)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2672  -0.7115  -0.3991   0.6372   3.1479
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.302e+00  2.401e-01  -9.586  < 2e-16 ***
## PARENT1Yes                    2.955e-01  1.249e-01   2.366  0.01799 *
```

```
## MSTATUSYes                     -5.310e-01  9.427e-02  -5.633 1.77e-08 ***
## EDUCATIONBachelors             -3.692e-01  1.220e-01  -3.026  0.00248 **
## EDUCATIONHigh School            4.574e-02  1.058e-01   0.432  0.66557
## EDUCATIONMasters               -5.345e-01  1.910e-01  -2.799  0.00513 **
## EDUCATIONPhD                   -1.451e-01  2.441e-01  -0.595  0.55208
## JOBClerical                     9.267e-02  1.174e-01   0.790  0.42972
## JOBDoctor                      -6.744e-01  2.806e-01  -2.403  0.01625 *
## JOBHome Maker                  -1.488e-01  1.708e-01  -0.871  0.38373
## JOBLawyer                       5.832e-02  1.995e-01   0.292  0.77009
## JOBManager                     -8.090e-01  1.568e-01  -5.159 2.48e-07 ***
## JOBProfessional                -1.214e-01  1.334e-01  -0.910  0.36279
## JOBStudent                     -2.973e-01  1.473e-01  -2.019  0.04354 *
## CAR_USEPrivate                 -7.920e-01  9.796e-02  -8.085 6.20e-16 ***
## CAR_TYPEPanel Truck             6.886e-01  1.664e-01   4.139 3.50e-05 ***
## CAR_TYPEPickup                  5.063e-01  1.117e-01   4.532 5.84e-06 ***
## CAR_TYPESports Car              8.690e-01  1.213e-01   7.162 7.93e-13 ***
## CAR_TYPESUV                     6.200e-01  9.612e-02   6.451 1.11e-10 ***
## CAR_TYPEVan                     5.983e-01  1.374e-01   4.354 1.34e-05 ***
## REVOKEDYes                      8.708e-01  1.022e-01   8.518  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.333e+00  1.237e-01  18.851  < 2e-16 ***
## KIDSDRIV                        5.732e-01  1.080e-01   5.309 1.10e-07 ***
## HOMEKIDS                        8.620e-02  4.124e-02   2.090  0.03660 *
## YOJ                            -1.384e-02  9.479e-03  -1.460  0.14442
## INCOME                         -4.109e-06  1.357e-06  -3.029  0.00246 **
## HOME_VAL                       -1.530e-06  3.843e-07  -3.981 6.85e-05 ***
## TRAVTIME                        1.608e-02  2.180e-03   7.376 1.63e-13 ***
## BLUEBOOK                       -2.965e-05  5.521e-06  -5.371 7.85e-08 ***
## TIF                            -5.667e-02  8.607e-03  -6.584 4.57e-11 ***
## OLDCLAIM                       -1.545e-05  5.375e-06  -2.875  0.00404 **
## CLM_FREQ                        1.868e-01  3.282e-02   5.693 1.25e-08 ***
## MVR_PTS                         1.286e-01  1.650e-02   7.794 6.50e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7527.5  on 6528  degrees of freedom
## Residual deviance: 5820.4  on 6496  degrees of freedom
## AIC: 5886.4
##
## Number of Fisher Scoring iterations: 5
```

```
round(exp(cbind(Estimate=coef(step.model_FLAG))),2)
```

```
##                              Estimate
## (Intercept)                      0.10
## PARENT1Yes                       1.34
## MSTATUSYes                       0.59
## EDUCATIONBachelors               0.69
## EDUCATIONHigh School             1.05
## EDUCATIONMasters                 0.59
## EDUCATIONPhD                     0.86
## JOBClerical                      1.10
## JOBDoctor                        0.51
```

```
## JOBHome Maker                      0.86
## JOBLawyer                          1.06
## JOBManager                         0.45
## JOBProfessional                    0.89
## JOBStudent                         0.74
## CAR_USEPrivate                     0.45
## CAR_TYPEPanel Truck                1.99
## CAR_TYPEPickup                     1.66
## CAR_TYPESports Car                 2.38
## CAR_TYPESUV                        1.86
## CAR_TYPEVan                        1.82
## REVOKEDYes                         2.39
## URBANICITYHighly Urban/ Urban     10.30
## KIDSDRIV                           1.77
## HOMEKIDS                           1.09
## YOJ                                0.99
## INCOME                             1.00
## HOME_VAL                           1.00
## TRAVTIME                           1.02
## BLUEBOOK                           1.00
## TIF                                0.94
## OLDCLAIM                           1.00
## CLM_FREQ                           1.21
## MVR_PTS                            1.14
```

```r
# evaluate the model by predicting using the testing set
m2_prob <- predict(step.model_FLAG, ntesting, type = "response")
m2_pclass <- ifelse(m2_prob >= 0.5, 1, 0)

# create confusion matrix
pclass <- factor(m2_pclass,levels = c(1,0))
aclass <- factor(ntesting$TARGET_FLAG,levels = c(1,0))
confusionMatrix(pclass, aclass);
```

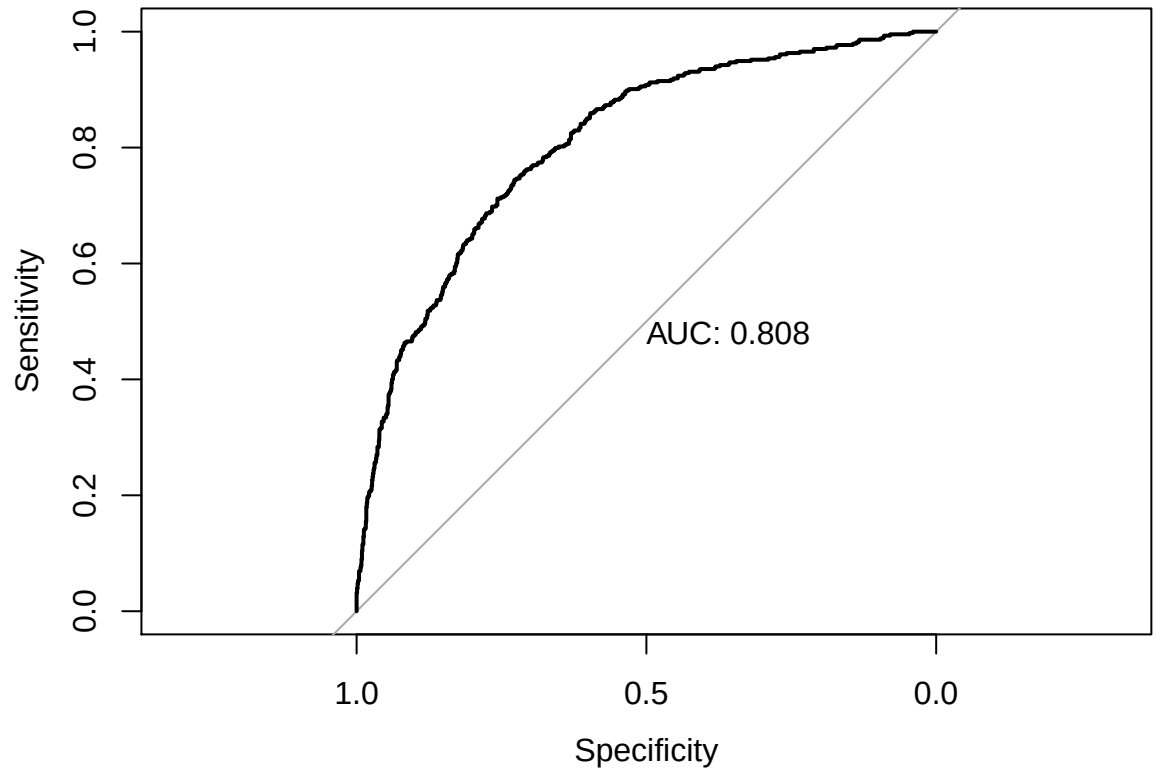
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  182   83
##          0  252 1115
##
##                Accuracy : 0.7947
##                  95% CI : (0.7743, 0.8141)
##     No Information Rate : 0.7341
##     P-Value [Acc > NIR] : 7.32e-09
##
##                   Kappa : 0.3997
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4194
##             Specificity : 0.9307
##          Pos Pred Value : 0.6868
##          Neg Pred Value : 0.8157
```

```
##              Prevalence : 0.2659
##          Detection Rate : 0.1115
##    Detection Prevalence : 0.1624
##       Balanced Accuracy : 0.6750
##
##        'Positive' Class : 1
##
```

```
# plot and show area under the curve
plot(roc(ntesting$TARGET_FLAG, m2_prob),print.auc=TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
# get McFadden
m2_mcFadden <- pR2(step.model_FLAG); m2_mcFadden["McFadden"]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.2267871
```

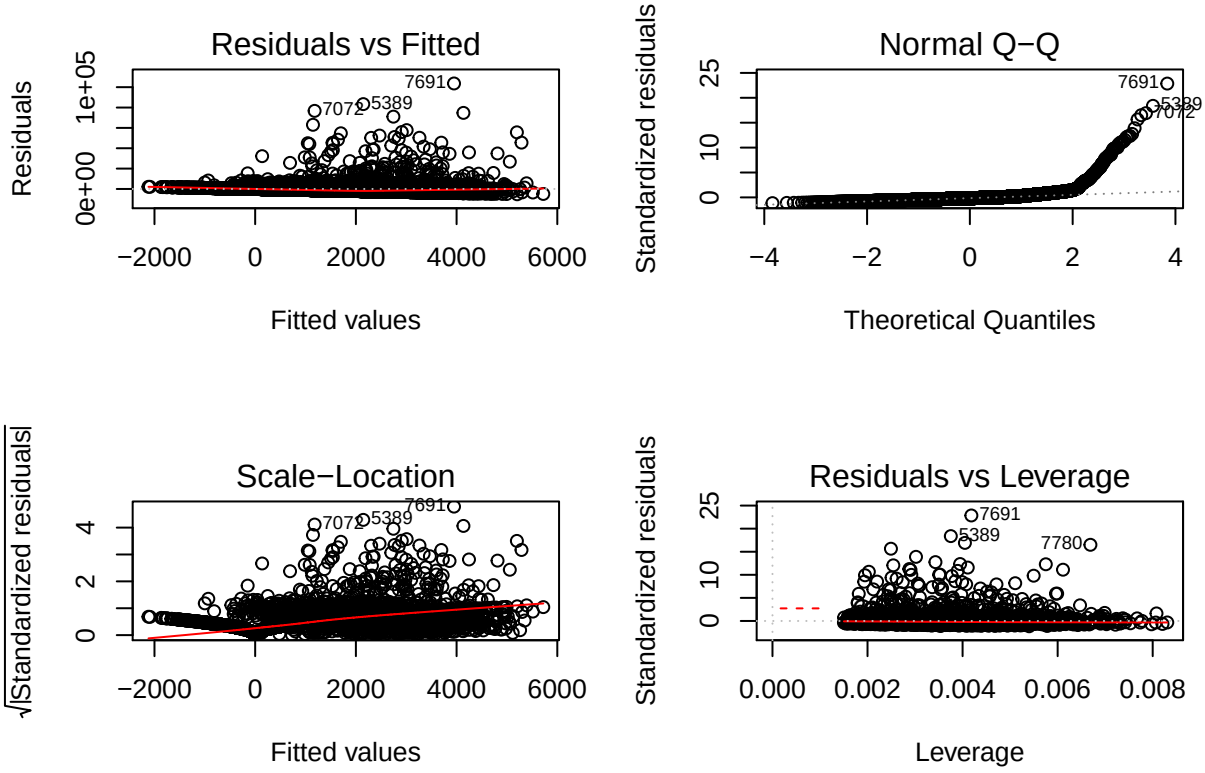
```r
# Linear Regression - TARGET_AMT
step.model.AMT <- full.model.AMT %>% stepAIC(trace = FALSE)
summary(step.model.AMT)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ PARENT1 + MSTATUS + SEX + JOB + CAR_USE +
##     CAR_TYPE + REVOKED + URBANICITY + KIDSDRIV + INCOME + TRAVTIME +
##     BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + MVR_PTS + CAR_AGE,
##     data = tr_prep2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5204  -1696   -761    339 103637
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.990e+02  3.427e+02   0.581 0.561477
## PARENT1Yes                     5.940e+02  1.783e+02   3.331 0.000869 ***
## MSTATUSYes                    -6.201e+02  1.196e+02  -5.184 2.23e-07 ***
## SEXM                           3.182e+02  1.604e+02   1.984 0.047322 *
## JOBClerical                    3.265e+01  1.907e+02   0.171 0.864042
## JOBDoctor                     -5.451e+02  2.860e+02  -1.906 0.056685 .
## JOBHome Maker                 -1.140e+02  2.465e+02  -0.462 0.643874
## JOBLawyer                     -1.610e+02  2.166e+02  -0.743 0.457219
## JOBManager                    -9.717e+02  2.116e+02  -4.592 4.46e-06 ***
## JOBProfessional               -1.250e+02  1.967e+02  -0.635 0.525251
## JOBStudent                    -1.356e+02  2.218e+02  -0.611 0.541042
## CAR_USEPrivate                -7.477e+02  1.510e+02  -4.951 7.51e-07 ***
## CAR_TYPEPanel Truck            3.094e+02  2.684e+02   1.153 0.249058
## CAR_TYPEPickup                 3.965e+02  1.693e+02   2.342 0.019200 *
## CAR_TYPESports Car             1.029e+03  2.164e+02   4.755 2.02e-06 ***
## CAR_TYPESUV                    7.487e+02  1.785e+02   4.194 2.77e-05 ***
## CAR_TYPEVan                    5.400e+02  2.114e+02   2.555 0.010648 *
## REVOKEDYes                     5.992e+02  1.742e+02   3.439 0.000587 ***
## URBANICITYHighly Urban/ Urban  1.667e+03  1.391e+02  11.985  < 2e-16 ***
## KIDSDRIV                       7.014e+02  1.620e+02   4.330 1.51e-05 ***
## INCOME                        -5.499e-03  1.838e-03  -2.991 0.002785 **
## TRAVTIME                       1.274e+01  3.338e+00   3.816 0.000137 ***
## BLUEBOOK                       1.391e-02  8.854e-03   1.571 0.116175
## TIF                           -5.065e+01  1.293e+01  -3.917 9.05e-05 ***
## OLDCLAIM                      -1.693e-02  9.114e-03  -1.858 0.063258 .
## CLM_FREQ                       1.688e+02  5.681e+01   2.972 0.002968 **
## MVR_PTS                        1.739e+02  2.786e+01   6.241 4.56e-10 ***
## CAR_AGE                       -2.765e+01  1.130e+01  -2.447 0.014435 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4544 on 8133 degrees of freedom
## Multiple R-squared:  0.06987,    Adjusted R-squared:  0.06678
## F-statistic: 22.63 on 27 and 8133 DF,  p-value: < 2.2e-16
```

```
vif(step.model.AMT)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## PARENT1    1.439404  1        1.199751
## MSTATUS    1.357795  1        1.165245
## SEX        2.529123  1        1.590322
## JOB        4.392786  7        1.111501
## CAR_USE    2.103321  1        1.450283
## CAR_TYPE   5.046346  5        1.175703
## REVOKED    1.289951  1        1.135760
## URBANICITY 1.243742  1        1.115232
## KIDSDRIV   1.096511  1        1.047144
## INCOME     2.234825  1        1.494933
## TRAVTIME   1.033216  1        1.016472
## BLUEBOOK   1.986772  1        1.409529
## TIF        1.005226  1        1.002610
## OLDCLAIM   1.825934  1        1.351271
## CLM_FREQ   1.711751  1        1.308339
## MVR_PTS    1.226375  1        1.107418
## CAR_AGE    1.524544  1        1.234724
```

```
par(mfrow=c(2,2))
plot(step.model.AMT)
```



```r
par(mfrow=c(1,1))
```

3.3 Significant value variable selection

```r
# Logistic Regression build the model using training set
select.model_FLAG <- glm(TARGET_FLAG ~.
                       -TARGET_AMT
                       -EDUCATION
                       -SEX
                       -RED_CAR
                       -KIDSDRIV
                       -AGE
                       -HOMEKIDS
                       -YOJ
                       -HOME_VAL
                       -OLDCLAIM
                       -BLUEBOOK

                         , data = ntraining , family = binomial)
summary(select.model_FLAG)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT - EDUCATION - SEX -
##     RED_CAR - KIDSDRIV - AGE - HOMEKIDS - YOJ - HOME_VAL - OLDCLAIM -
##     BLUEBOOK, family = binomial, data = ntraining)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1717  -0.7259  -0.4198   0.6630   3.0165
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.750e+00  1.961e-01 -14.025  < 2e-16 ***
## PARENT1Yes                     6.245e-01  1.014e-01   6.157 7.43e-10 ***
## MSTATUSYes                    -5.663e-01  7.483e-02  -7.568 3.78e-14 ***
## JOBClerical                    1.107e-01  1.160e-01   0.954 0.340301
## JOBDoctor                     -5.983e-01  1.880e-01  -3.183 0.001459 **
## JOBHome Maker                 -1.827e-01  1.536e-01  -1.190 0.234123
## JOBLawyer                     -2.678e-01  1.349e-01  -1.985 0.047112 *
## JOBManager                    -9.881e-01  1.411e-01  -7.003 2.51e-12 ***
## JOBProfessional               -3.374e-01  1.226e-01  -2.752 0.005928 **
## JOBStudent                    -7.074e-02  1.361e-01  -0.520 0.603137
## CAR_USEPrivate                -7.352e-01  9.188e-02  -8.002 1.23e-15 ***
## CAR_TYPEPanel Truck            3.687e-01  1.494e-01   2.468 0.013600 *
## CAR_TYPEPickup                 6.017e-01  1.082e-01   5.563 2.65e-08 ***
## CAR_TYPESports Car             9.588e-01  1.184e-01   8.095 5.72e-16 ***
## CAR_TYPESUV                    7.162e-01  9.377e-02   7.638 2.21e-14 ***
## CAR_TYPEVan                    4.380e-01  1.327e-01   3.300 0.000968 ***
## REVOKEDYes                     7.422e-01  8.792e-02   8.442  < 2e-16 ***
## URBANICITYHighly Urban/ Urban  2.246e+00  1.221e-01  18.402  < 2e-16 ***
```

```
## INCOME                         -7.766e-06  1.216e-06  -6.388 1.68e-10 ***
## TRAVTIME                        1.543e-02  2.149e-03   7.181 6.94e-13 ***
## TIF                            -5.639e-02  8.521e-03  -6.618 3.63e-11 ***
## CLM_FREQ                        1.516e-01  2.817e-02   5.381 7.41e-08 ***
## MVR_PTS                         1.270e-01  1.621e-02   7.836 4.67e-15 ***
## CAR_AGE                        -1.642e-02  7.285e-03  -2.254 0.024225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7527.5  on 6528  degrees of freedom
## Residual deviance: 5948.1  on 6505  degrees of freedom
## AIC: 5996.1
##
## Number of Fisher Scoring iterations: 5
```

```
round(exp(cbind(Estimate=coef(select.model_FLAG))),2)
```

```
##                               Estimate
## (Intercept)                       0.06
## PARENT1Yes                        1.87
## MSTATUSYes                        0.57
## JOBClerical                       1.12
## JOBDoctor                         0.55
## JOBHome Maker                     0.83
## JOBLawyer                         0.77
## JOBManager                        0.37
## JOBProfessional                   0.71
## JOBStudent                        0.93
## CAR_USEPrivate                    0.48
## CAR_TYPEPanel Truck               1.45
## CAR_TYPEPickup                    1.83
## CAR_TYPESports Car                2.61
## CAR_TYPESUV                       2.05
## CAR_TYPEVan                       1.55
## REVOKEDYes                        2.10
## URBANICITYHighly Urban/ Urban     9.45
## INCOME                            1.00
## TRAVTIME                          1.02
## TIF                               0.95
## CLM_FREQ                          1.16
## MVR_PTS                           1.14
## CAR_AGE                           0.98
```

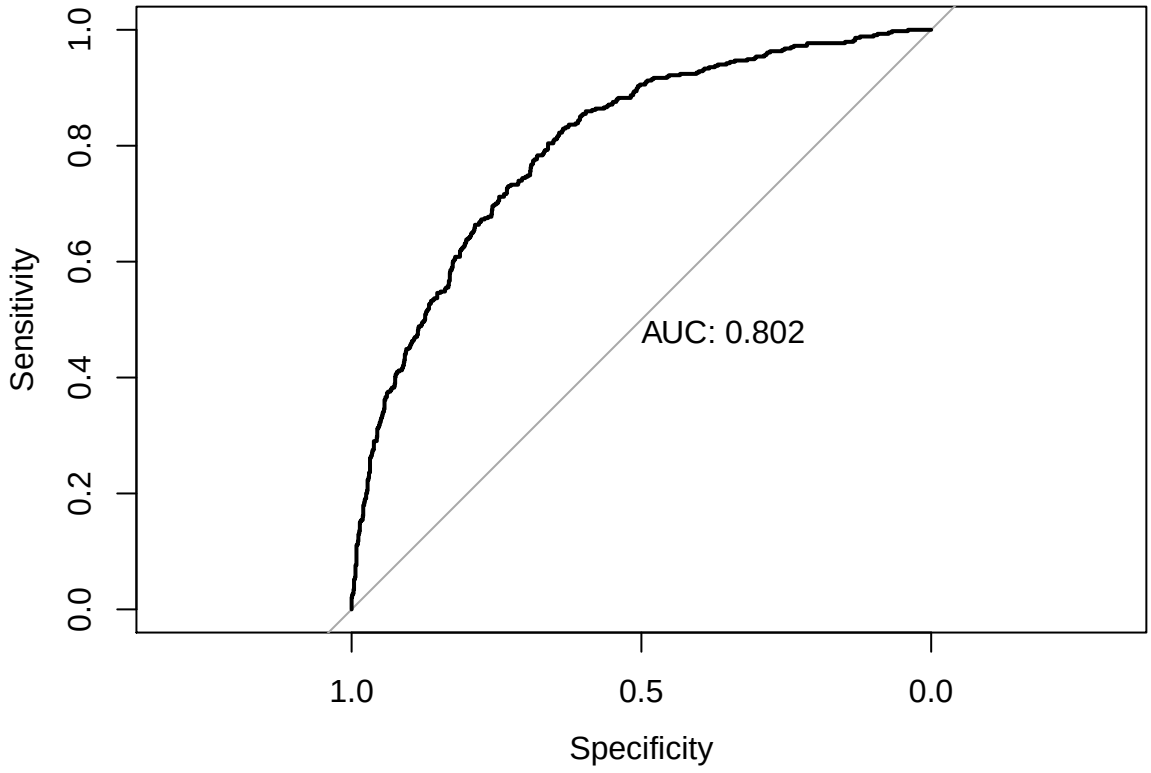
```
# evaluate the model by predicting using the testing set
m3_prob <- predict(select.model_FLAG, ntesting, type = "response")
m3_pclass <- ifelse(m3_prob >= 0.5, 1, 0)

# create confusion matrix
pclass <- factor(m3_pclass,levels = c(1,0))
aclass <- factor(ntesting$TARGET_FLAG,levels = c(1,0))
confusionMatrix(pclass, aclass);
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  166   85
##          0  268 1113
##
##                Accuracy : 0.7837
##                  95% CI : (0.7629, 0.8035)
##     No Information Rate : 0.7341
##     P-Value [Acc > NIR] : 2.069e-06
##
##                   Kappa : 0.3599
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3825
##             Specificity : 0.9290
##          Pos Pred Value : 0.6614
##          Neg Pred Value : 0.8059
##              Prevalence : 0.2659
##          Detection Rate : 0.1017
##    Detection Prevalence : 0.1538
##       Balanced Accuracy : 0.6558
##
##        'Positive' Class : 1
##
```

```
# plot and show area under the curve
plot(roc(ntesting$TARGET_FLAG, m3_prob),print.auc=TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
# get McFadden
m3_mcFadden <- pR2(select.model_FLAG); m3_mcFadden["McFadden"]
```

```
## fitting null model for pseudo-r2

##  McFadden
## 0.2098243
```

```
# Linear Regression - TARGET_AMT
select.model.AMT <- lm(TARGET_AMT ~.
                       -TARGET_FLAG
                       -EDUCATION
                       -SEX
                       -RED_CAR
                       -KIDSDRIV
                       -AGE
                       -HOMEKIDS
                       -YOJ
                       -HOME_VAL
                       -OLDCLAIM
                       -BLUEBOOK
                       ,  data = tr_prep2)
summary(select.model.AMT)
```

```
##
```

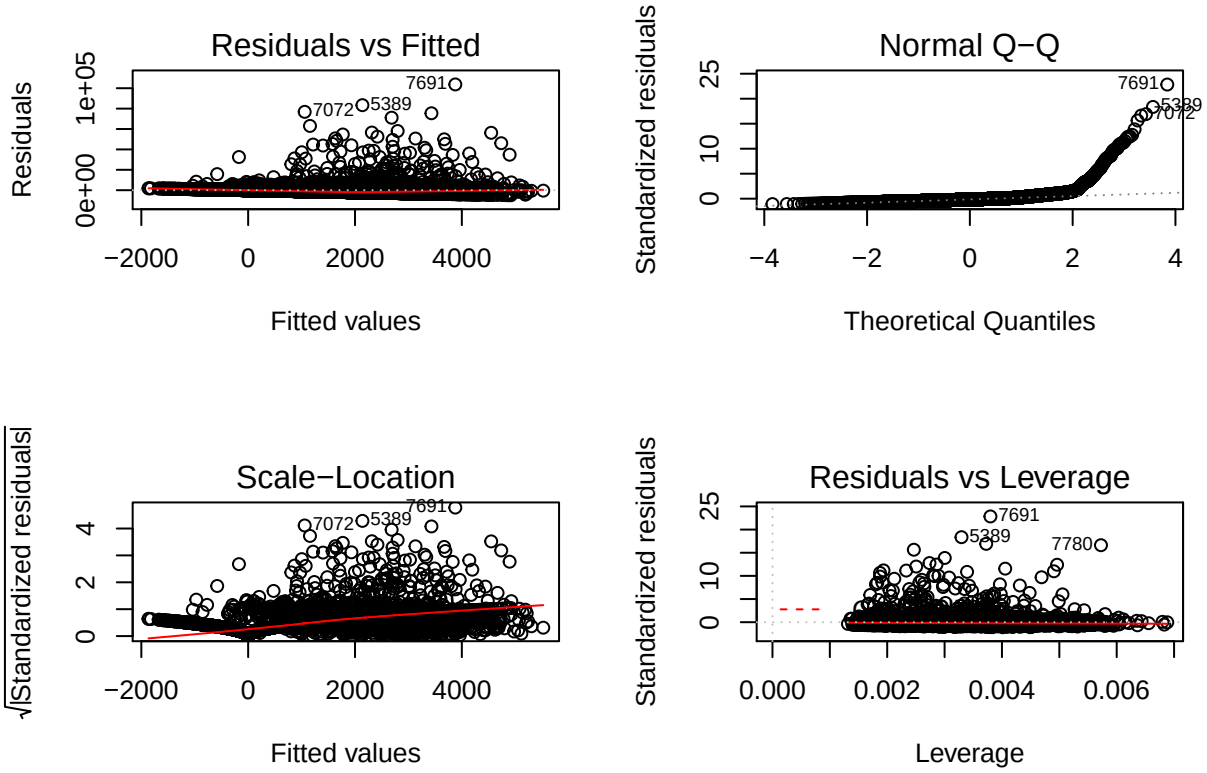
```
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG - EDUCATION - SEX -
##     RED_CAR - KIDSDRIV - AGE - HOMEKIDS - YOJ - HOME_VAL - OLDCLAIM -
##     BLUEBOOK, data = tr_prep2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4919  -1694   -764    329 103709
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.323e+02  2.756e+02   2.294 0.021833 *
## PARENT1Yes                    7.893e+02  1.716e+02   4.600 4.29e-06 ***
## MSTATUSYes                   -5.333e+02  1.180e+02  -4.521 6.23e-06 ***
## JOBClerical                   1.389e+01  1.908e+02   0.073 0.941986
## JOBDoctor                    -5.647e+02  2.861e+02  -1.974 0.048433 *
## JOBHome Maker                -1.995e+02  2.438e+02  -0.818 0.413208
## JOBLawyer                    -1.830e+02  2.167e+02  -0.844 0.398476
## JOBManager                   -9.853e+02  2.118e+02  -4.653 3.32e-06 ***
## JOBProfessional              -1.459e+02  1.967e+02  -0.742 0.458352
## JOBStudent                   -1.712e+02  2.216e+02  -0.773 0.439748
## CAR_USEPrivate               -7.361e+02  1.511e+02  -4.871 1.13e-06 ***
## CAR_TYPEPanel Truck           5.732e+02  2.388e+02   2.400 0.016396 *
## CAR_TYPEPickup                3.797e+02  1.680e+02   2.260 0.023828 *
## CAR_TYPESports Car            7.770e+02  1.831e+02   4.243 2.23e-05 ***
## CAR_TYPESUV                   5.240e+02  1.389e+02   3.771 0.000164 ***
## CAR_TYPEVan                   6.602e+02  2.038e+02   3.239 0.001202 **
## REVOKEDYes                    4.733e+02  1.550e+02   3.054 0.002266 **
## URBANICITYHighly Urban/ Urban 1.646e+03  1.392e+02  11.822  < 2e-16 ***
## INCOME                       -5.164e-03  1.805e-03  -2.860 0.004242 **
## TRAVTIME                      1.292e+01  3.341e+00   3.866 0.000111 ***
## TIF                          -5.077e+01  1.295e+01  -3.921 8.89e-05 ***
## CLM_FREQ                      1.235e+02  4.891e+01   2.524 0.011616 *
## MVR_PTS                       1.696e+02  2.775e+01   6.115 1.01e-09 ***
## CAR_AGE                      -2.778e+01  1.131e+01  -2.455 0.014105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4551 on 8137 degrees of freedom
## Multiple R-squared:  0.06681,    Adjusted R-squared:  0.06418
## F-statistic: 25.33 on 23 and 8137 DF,  p-value: < 2.2e-16
```

```
vif(select.model.AMT)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## PARENT1    1.329311  1        1.152957
## MSTATUS    1.316334  1        1.147316
## JOB        4.219698  7        1.108314
## CAR_USE    2.100340  1        1.449255
## CAR_TYPE   1.803986  5        1.060775
## REVOKED    1.017562  1        1.008743
## URBANICITY 1.242415  1        1.114637
## INCOME     2.149521  1        1.466124
## TRAVTIME   1.032640  1        1.016189
```

```
## TIF         1.005153  1         1.002573
## CLM_FREQ    1.265076  1         1.124756
## MVR_PTS     1.212759  1         1.101253
## CAR_AGE     1.524066  1         1.234531
```

```
par(mfrow=c(2,2))
plot(select.model.AMT)
```




```
par(mfrow=c(1,1))
```

SELECT MODELS

```
# Read the evaluation dataset
 eval_df <- read.csv("https://raw.githubusercontent.com/monuchacko/cuny_msds/master/data_621/Homework4/


# Remove columns not selected in 2nd model
#eval_df <- dplyr::select(eval_df, -YOJ, -MSTATUS, -RED_CAR)

#  Convert to numeric
eval_df$INCOME <- as.numeric(gsub('[$,]', '', eval_df$INCOME))
eval_df$HOME_VAL <- as.numeric(gsub('[$,]', '', eval_df$HOME_VAL))
eval_df$BLUEBOOK <- as.numeric(gsub('[$,]', '', eval_df$BLUEBOOK))
```

```r
eval_df$OLDCLAIM <- as.numeric(gsub('[$,]', '', eval_df$OLDCLAIM))

# Remove irrelevant characters
eval_df$MSTATUS  <- gsub("z_", "", eval_df$MSTATUS)
eval_df$SEX  <- gsub("z_", "", eval_df$SEX)
eval_df$EDUCATION  <- gsub("z_", "", eval_df$EDUCATION)
eval_df$JOB  <- gsub("z_", "", eval_df$JOB)
eval_df$CAR_USE  <- gsub("z_", "", eval_df$CAR_USE)
eval_df$CAR_TYPE  <- gsub("z_", "", eval_df$CAR_TYPE)
eval_df$URBANICITY  <- gsub("z_", "", eval_df$URBANICITY)

#impute
eval_df = eval_df %>%
  mutate(AGE =
           ifelse(is.na(AGE),
                  mean(AGE, na.rm=TRUE), AGE)) %>%

  mutate(YOJ =
           ifelse(is.na(YOJ),
                  mean(YOJ, na.rm=TRUE), YOJ)) %>%

  mutate(INCOME =
           ifelse(is.na(INCOME),
                  median(INCOME, na.rm=TRUE), INCOME)) %>%

  mutate(HOME_VAL =
           ifelse(is.na(HOME_VAL),
                  mean(HOME_VAL, na.rm=TRUE), HOME_VAL)) %>%

  mutate(CAR_AGE =
           ifelse(is.na(CAR_AGE),
                  mean(CAR_AGE, na.rm=TRUE), CAR_AGE)) %>%

  mutate(JOB =
           ifelse((JOB == "" & EDUCATION == 'PhD'),
                  "Doctor", JOB)) %>%

  mutate(JOB =
           ifelse((JOB == "" & EDUCATION == 'Masters'),
                  "Lawyer", JOB))


eval_prob <- predict(step.model_FLAG, eval_df, type = "response")
eval_pclass <- ifelse(eval_prob >= 0.5, 1, 0)

eval_amt <- ifelse(eval_pclass == 1, predict(step.model.AMT, eval_df, type = "response"), 0)


eval_df$TARGET_FLAG <- eval_pclass
eval_df$TARGET_AMT <- eval_amt

head(eval_df)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS      YOJ INCOME PARENT1
## 1     3           0          0        0  48        0 11.00000  52881      No
## 2     9           0          0        1  40        1 11.00000  50815     Yes
## 3    10           0          0        0  44        2 12.00000  43486     Yes
## 4    18           0          0        0  35        2 10.37909  21204     Yes
## 5    21           0          0        0  59        0 12.00000  87460      No
## 6    30           0          0        0  46        0 14.00000  51778      No
##   HOME_VAL MSTATUS SEX   EDUCATION          JOB TRAVTIME    CAR_USE BLUEBOOK
## 1        0      No   M   Bachelors      Manager       26    Private    21970
## 2        0      No   M High School      Manager       21    Private    18930
## 3        0      No   F High School  Blue Collar       30 Commercial     5900
## 4        0      No   M High School     Clerical       74    Private     9230
## 5        0      No   M High School      Manager       45    Private    15420
## 6   207519     Yes   M   Bachelors Professional        7 Commercial    25660
##   TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1   1         Van     yes        0        0      No       2      10
## 2   6     Minivan      no     3295        1      No       2       1
## 3  10         SUV      no        0        0      No       0      10
## 4   6      Pickup      no        0        0     Yes       0       4
## 5   1     Minivan     yes    44857        2      No       4       1
## 6   1 Panel Truck      no     2119        1      No       2      12
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Rural/ Rural
## 4 Highly Rural/ Rural
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

```
# Export
#  write.csv(eval_df,file="Insurance_Results.csv")
```