DATA621_HW4

Calvin Wong, Sudhan Maharjan, Ravi Itwaru

11/4/2019

Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, 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 response variable 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. 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).

Write Up:

1. DATA EXPLORATION (25 Points) Describe the size and the variables in the insurance training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median

b. Bar Chart or Box Plot of the data

c. Is the data correlated to the target variable (or to other variables?)

d. Are any of the variables missing and need to be imputed "fixed"?

```
#skim(train)
train %>%
  skim()
```

```
## Skim summary statistics
##  n obs: 8161
##  n variables: 26
##
## ── Variable type:character ──────────────────────────────────────────────
────────────────────────────────────────────────────────────
##    variable missing complete    n min max empty n_unique
##    BLUEBOOK       0     8161 8161   6   7     0     2789
##    CAR_TYPE       0     8161 8161   3  11     0        6
##     CAR_USE       0     8161 8161   7  10     0        2
```

```
##    EDUCATION       0      8161 8161   3  13      0        5
##     HOME_VAL     464      7697 8161   2   8      0     5106
##       INCOME     445      7716 8161   2   8      0     6612
##          JOB     526      7635 8161   6  13      0        8
##      MSTATUS       0      8161 8161   3   4      0        2
##     OLDCLAIM       0      8161 8161   2   7      0     2857
##      PARENT1       0      8161 8161   2   3      0        2
##      RED_CAR       0      8161 8161   2   3      0        2
##      REVOKED       0      8161 8161   2   3      0        2
##          SEX       0      8161 8161   1   3      0        2
##   URBANICITY       0      8161 8161  19  21      0        2
## 
## ── Variable type:numeric ────────────────────────────────────────
────────────────────────────────────────────────
##      variable missing complete    n    mean      sd p0  p25  p50  p75
##           AGE       6     8155 8161   44.79    8.63 16   39   45   51
##       CAR_AGE     510     7651 8161    8.33    5.7  -3    1    8   12
##      CLM_FREQ       0     8161 8161    0.8     1.16  0    0    0    2
##      HOMEKIDS       0     8161 8161    0.72    1.12  0    0    0    1
##         INDEX       0     8161 8161 5151.87 2978.89  1 2559 5133 7745
##      KIDSDRIV       0     8161 8161    0.17    0.51  0    0    0    0
##       MVR_PTS       0     8161 8161    1.7     2.15  0    0    1    3
##    TARGET_AMT       0     8161 8161 1504.32 4704.03  0    0    0 1036
##   TARGET_FLAG       0     8161 8161    0.26    0.44  0    0    0    1
##           TIF       0     8161 8161    5.35    4.15  1    1    4    7
##      TRAVTIME       0     8161 8161   33.49   15.91  5   22   33   44
##           YOJ     454     7707 8161   10.5     4.09  0    9   11   13
##       p100     hist
##       81    ▁▂▇▇▇▃▁▁
##       28    ▁▇▇▇▅▃▁▁
##        5    ▇▂▁▂▁▁▁▁
##        5    ▇▂▁▂▁▁▁▁
##    10302    ▇▇▇▇▇▇▇▇
##        4    ▇▁▁▁▁▁▁▁
##       13    ▇▂▁▁▁▁▁▁
##  107586.14  ▇▁▁▁▁▁▁▁
##        1    ▇▁▁▁▁▁▁▃
##       25    ▇▃▂▁▁▁▁▁
##      142    ▅▇▅▁▁▁▁▁
##       23    ▂▁▂▅▇▃▁▁
```

2. DATA PREPARATION (25 Points) Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value)

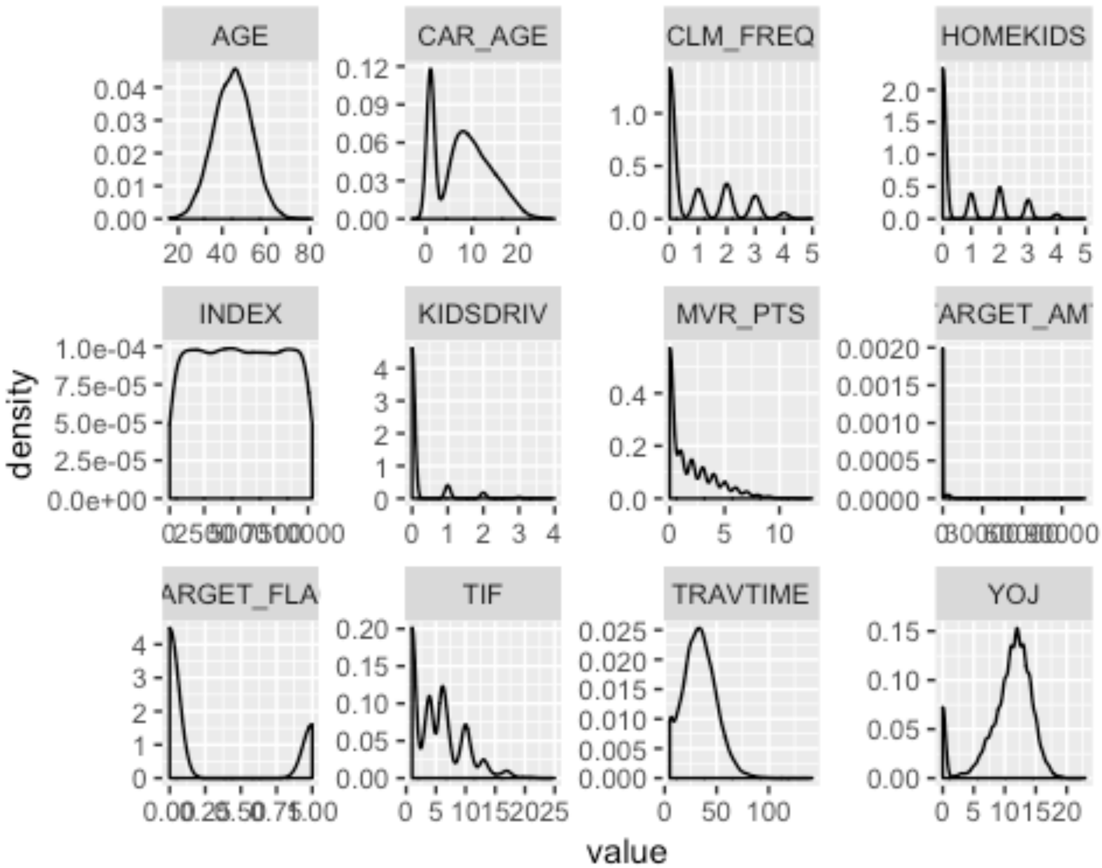
b. Create flags to suggest if a variable was missing

c. Transform data by putting it into buckets

d. Mathematical transforms such as log or square root (or use Box-Cox)

e. Combine variables (such as ratios or adding or multiplying) to create new variables

```
train$HOME_VAL <- as.numeric(gsub('[$,]', '', train$HOME_VAL))
# trigger a dummy variable if NA is present
train$HOME_VAL_MISSING <- ifelse(is.na(train$HOME_VAL), 1, 0)
# imputing NA to mean
train$HOME_VAL[is.na(train$HOME_VAL)] <- mean(train$HOME_VAL, na.rm=TRUE)

train$INCOME <- as.numeric(gsub('[$,]', '', train$INCOME))
# trigger a dummy variable if NA is present
train$INCOME_MISSING <- ifelse(is.na(train$INCOME), 1, 0)
# imputing NA to mean
train$INCOME[is.na(train$INCOME)] <- mean(train$INCOME, na.rm=TRUE)

# trigger a dummy variable if NA is present
train$CAR_AGE_MISSING <- ifelse(is.na(train$CAR_AGE), 1, 0)
# imputing NA to mean
train$CAR_AGE[is.na(train$CAR_AGE)] <- mean(train$CAR_AGE, na.rm=TRUE)

# trigger a dummy variable if NA is present
train$YOJ_MISSING <- ifelse(is.na(train$YOJ), 1, 0)
# imputing NA to mean
train$YOJ[is.na(train$YOJ)] <- mean(train$YOJ, na.rm=TRUE)

train %>%
  skim()

n_train <- select_if(train, is.numeric)
n_train %>%
  keep(is.numeric) %>%                    #keep only columns with numeric va
lues
  gather() %>%                            #convert to key-value
  ggplot(aes(value)) +                    #plot the values
    facet_wrap(~key, scales="free") +
    geom_density()

## Warning: Removed 970 rows containing non-finite values (stat_density).
```




```
# transform data using log for skewed HOMEKIDS, MVR_PTS, TIF, KIDSDRIVE and C
LM_FREQ

train$HOMEKIDS <- log(train$HOMEKIDS+1)
train$MVR_PTS <- log(train$MVR_PTS+1)
train$TIF <- log(train$TIF+1)
train$KIDSDRIV <- log(train$KIDSDRIV+1)
train$CLM_FREQ <- log(train$CLM_FREQ+1)
```

As we see in the output below, there are a few variables that will need to be transformed. Variables INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM will be transformed back numerical values since they are numerics but were converted to string to include the $ symbol.

```
# Remove index from the dataset
train <- subset(train, select = -c(INDEX))
# Display the structure of the dataset
str(train)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    8161 obs. of  25 variables:
##  $ TARGET_FLAG: num  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : num  0 0 0 0 0 ...
```

```
##  $ AGE        : num  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : num  0 0 0.693 0 0 ...
##  $ YOJ        : num  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" NA ...
##  $ 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 Col
lar" ...
##  $ TRAVTIME   : num  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        : num  2.485 0.693 1.609 2.079 0.693 ...
##  $ 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   : num  1.1 0 1.1 0 1.1 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : num  1.39 0 1.39 0 1.39 ...
##  $ CAR_AGE    : num  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly U
rban/ Urban" "Highly Urban/ Urban" ...
```

```
# Convert Income to numerical value
train$INCOME <- parse_number(train$INCOME)
# Convert home_val to numerical value
train$HOME_VAL <- parse_number(train$HOME_VAL)
# Convert bluebook  to numerical value
train$BLUEBOOK <- parse_number(train$BLUEBOOK)
# Convert oldcaim  to numerical value
train$OLDCLAIM <- parse_number(train$OLDCLAIM)
```

Below we can see the levels of the non-numerical variables. Variable PARENT1, MSTATUS, SEX, CAR_USE ,RED_CAR, REVOKED and URBANICITY have only two levels which makes them good candidates for binary conversion. The remaining variables that have more than two levels will be converted to dummy variables.

```
train_factor <- train %>%
  mutate_if(sapply(train, is.character), as.factor)

train_levels <- train_factor %>%
     sapply(levels)
train_levels[sapply(train_levels, is.null)] <- NULL
train_levels

## $PARENT1
## [1] "No"  "Yes"
```

```
##
## $MSTATUS
## [1] "Yes"  "z_No"
##
## $SEX
## [1] "M"   "z_F"
##
## $EDUCATION
## [1] "<High School"  "Bachelors"     "Masters"       "PhD"
## [5] "z_High School"
##
## $JOB
## [1] "Clerical"      "Doctor"        "Home Maker"    "Lawyer"
## [5] "Manager"       "Professional"  "Student"       "z_Blue Collar"
##
## $CAR_USE
## [1] "Commercial" "Private"
##
## $CAR_TYPE
## [1] "Minivan"     "Panel Truck" "Pickup"      "Sports Car"  "Van"
## [6] "z_SUV"
##
## $RED_CAR
## [1] "no"  "yes"
##
## $REVOKED
## [1] "No"  "Yes"
##
## $URBANICITY
## [1] "Highly Urban/ Urban"   "z_Highly Rural/ Rural"
```

```
# Convert variables PARENT1, MSTATUS, SEX, CAR_USE ,RED_CAR, REVOKED and URBA
NICITY to binary values(yes = 1, Commercial = 1, Highly Urban/ Urba = 1 )
train$PARENT1 <- if_else(train$PARENT1 == "Yes", 1, 0)
train$MSTATUS <- if_else(train$MSTATUS == "Yes", 1, 0)
train$SEX <- if_else(train$SEX == "M", 1, 0)
train$CAR_USE <- if_else(train$CAR_USE == "Commercial", 1, 0)
train$RED_CAR <- if_else(train$RED_CAR == "yes", 1, 0)
train$REVOKED <- if_else(train$REVOKED == "Yes", 1, 0)
train$URBANICITY <- if_else(train$URBANICITY == "Highly Urban/ Urba", 1, 0)

# Create dummy variables for EDUCATION, JOB, and, CAR_TYPE
#Education
train$"High School" <- if_else(train$EDUCATION == "<High School", 1, 0)
train$"Bachelors" <- if_else(train$EDUCATION == "Bachelors", 1, 0)
train$"Masters" <- if_else(train$EDUCATION == "Masters", 1, 0)
train$"PhD" <- if_else(train$EDUCATION == "PhD", 1, 0)
train$"z_High School" <- if_else(train$EDUCATION == "z_High School", 1, 0)

#Jobs
```

```r
train$"Clerical" <- if_else(train$JOB == "Clerical", 1, 0)
train$"Doctor" <- if_else(train$JOB == "Doctor", 1, 0)
train$"Home Maker" <- if_else(train$JOB == "Home Maker", 1, 0)
train$"Lawyer" <- if_else(train$JOB == "Lawyer", 1, 0)
train$"Manager" <- if_else(train$JOB == "Manager", 1, 0)
train$"Professional" <- if_else(train$JOB == "Professional", 1, 0)
train$"Student" <- if_else(train$JOB == "Student", 1, 0)
train$"z_Blue Collar" <- if_else(train$JOB == "z_Blue Collar", 1, 0)

# Car type
train$"Minivan" <- if_else(train$CAR_TYPE == "Minivan", 1, 0)
train$"Panel Truck" <- if_else(train$CAR_TYPE == "Panel Truck", 1, 0)
train$"Sports Car" <- if_else(train$CAR_TYPE == "Sports Car", 1, 0)
train$"Van" <- if_else(train$CAR_TYPE == "Van", 1, 0)
train$"z_SUV" <- if_else(train$CAR_TYPE == "z_SUV", 1, 0)

#write.csv(train, "new_train.csv")
#Data after conversion
#str(new_train)
```

3. BUILD MODELS (25 Points) Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Model Building

Model bm1

Model bm1 is our kitchen sink regression. This model basically has all the predictor variables (excluding the index) from our training dataset which includes our dummy variables.

```r
bm1 <- glm(TARGET_FLAG ~. - TARGET_AMT, data = train)
(bm1sum <- summary(bm1))
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, data = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8807  -0.2831  -0.1333   0.3355   1.1270
##
## Coefficients: (19 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.719e-01  4.549e-02   5.977 2.41e-09 ***
## KIDSDRIV                6.763e-02  2.067e-02   3.272 0.001073 **
## AGE                    -9.096e-04  7.352e-04  -1.237 0.216094
## HOMEKIDS                8.098e-03  1.466e-02   0.552 0.580638
## YOJ                    -1.545e-03  1.511e-03  -1.022 0.306656
## INCOME                 -3.163e-07  2.090e-07  -1.513 0.130230
## PARENT1                 8.556e-02  2.129e-02   4.019 5.91e-05 ***
## HOME_VAL               -1.801e-07  6.535e-08  -2.755 0.005881 **
## MSTATUS                -5.019e-02  1.543e-02  -3.252 0.001151 **
## SEX                     2.802e-02  1.864e-02   1.503 0.132930
## EDUCATIONBachelors     -5.527e-02  2.071e-02  -2.669 0.007633 **
## EDUCATIONMasters       -4.482e-02  3.104e-02  -1.444 0.148737
## EDUCATIONPhD            1.411e-02  3.896e-02   0.362 0.717223
## EDUCATIONz_High School -3.754e-03  1.709e-02  -0.220 0.826127
## JOBDoctor              -5.402e-02  4.512e-02  -1.197 0.231239
## JOBHome Maker          -4.102e-02  2.558e-02  -1.604 0.108866
## JOBLawyer               1.425e-02  3.055e-02   0.467 0.640794
## JOBManager             -7.570e-02  2.339e-02  -3.236 0.001219 **
## JOBProfessional        -5.460e-03  2.158e-02  -0.253 0.800287
## JOBStudent             -3.834e-02  2.418e-02  -1.586 0.112861
## JOBz_Blue Collar        3.662e-03  1.912e-02   0.192 0.848124
## TRAVTIME                1.153e-03  3.267e-04   3.530 0.000419 ***
## CAR_USE                 1.193e-01  1.673e-02   7.130 1.12e-12 ***
## BLUEBOOK               -2.658e-06  8.784e-07  -3.026 0.002491 **
## TIF                    -4.468e-02  7.364e-03  -6.068 1.38e-09 ***
## CAR_TYPEPanel Truck     7.522e-02  3.006e-02   2.503 0.012347 *
## CAR_TYPEPickup          6.824e-02  1.718e-02   3.971 7.23e-05 ***
## CAR_TYPESports Car      1.406e-01  2.163e-02   6.500 8.67e-11 ***
## CAR_TYPEVan             6.433e-02  2.211e-02   2.909 0.003635 **
## CAR_TYPEz_SUV           9.979e-02  1.780e-02   5.605 2.17e-08 ***
## RED_CAR                -2.936e-02  1.564e-02  -1.877 0.060521 .
## OLDCLAIM               -3.369e-06  7.893e-07  -4.268 2.00e-05 ***
## CLM_FREQ                1.452e-01  1.213e-02  11.972  < 2e-16 ***
## REVOKED                 1.795e-01  1.789e-02  10.037  < 2e-16 ***
## MVR_PTS                 6.261e-02  7.794e-03   8.033 1.13e-15 ***
## CAR_AGE                -7.367e-04  1.322e-03  -0.557 0.577394
## URBANICITY                     NA         NA      NA       NA
## `High School`                  NA         NA      NA       NA
## Bachelors                      NA         NA      NA       NA
## Masters                        NA         NA      NA       NA
## PhD                            NA         NA      NA       NA
## `z_High School`                NA         NA      NA       NA
## Clerical                       NA         NA      NA       NA
## Doctor                         NA         NA      NA       NA
```

```
## `Home Maker`                   NA         NA       NA        NA
## Lawyer                         NA         NA       NA        NA
## Manager                        NA         NA       NA        NA
## Professional                   NA         NA       NA        NA
## Student                        NA         NA       NA        NA
## `z_Blue Collar`                NA         NA       NA        NA
## Minivan                        NA         NA       NA        NA
## `Panel Truck`                  NA         NA       NA        NA
## `Sports Car`                   NA         NA       NA        NA
## Van                            NA         NA       NA        NA
## z_SUV                          NA         NA       NA        NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1609716)
##
##     Null deviance: 1177.45  on 6044  degrees of freedom
## Residual deviance:  967.28  on 6009  degrees of freedom
##   (2116 observations deleted due to missingness)
## AIC: 6151.5
##
## Number of Fisher Scoring iterations: 2
```

Model bm2

Model bm2 reviews factors which are considered risky and makes the assumption that riskier factors are the cause of accidents.

```
bm2 <- glm(TARGET_FLAG ~ RED_CAR + (AGE < 30) + (MVR_PTS > 3) + (REVOKED == 1
), family = binomial(link = "logit"), train)
(bm2sum <- summary(bm2))
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ RED_CAR + (AGE < 30) + (MVR_PTS >
##     3) + (REVOKED == 1), family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5220  -0.7279  -0.7279   1.2407   1.7341
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.19305    0.03270 -36.483   <2e-16 ***
## RED_CAR          -0.05912    0.05661  -1.044    0.296
## AGE < 30TRUE      1.04550    0.11620   8.998   <2e-16 ***
## MVR_PTS > 3TRUE        NA         NA      NA       NA
## REVOKED == 1TRUE  0.92884    0.06991  13.287   <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: 9404.0  on 8154  degrees of freedom
## Residual deviance: 9153.8  on 8151  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 9161.8
##
## Number of Fisher Scoring iterations: 4
```

Model bm3

Model bm3 reviews factors which are considered less risky to determine if there is any alignment with the target variable.

```
bm3 <- glm(TARGET_FLAG ~ (RED_CAR == 0) + (CLM_FREQ < 1) + (SEX == 0)  + (AGE
> 30 & AGE < 60) + (MVR_PTS < 2) + (REVOKED == 0) + (YOJ > 10), family = bino
mial(link = "logit"), train)
(bm3sum <- summary(bm3))

##
## Call:
## glm(formula = TARGET_FLAG ~ (RED_CAR == 0) + (CLM_FREQ < 1) +
##     (SEX == 0) + (AGE > 30 & AGE < 60) + (MVR_PTS < 2) + (REVOKED ==
##     0) + (YOJ > 10), family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1773  -0.6899  -0.6175   0.9075   1.9296
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.13987    0.16623  12.873  < 2e-16 ***
## RED_CAR == 0TRUE         0.03823    0.08079   0.473    0.636
## CLM_FREQ < 1TRUE        -0.78998    0.05797 -13.627  < 2e-16 ***
## SEX == 0TRUE             0.09412    0.07318   1.286    0.198
## AGE > 30 & AGE < 60TRUE -0.58770    0.08565  -6.862 6.80e-12 ***
## MVR_PTS < 2TRUE         -1.35175    0.12752 -10.600  < 2e-16 ***
## REVOKED == 0TRUE        -0.85712    0.07425 -11.544  < 2e-16 ***
## YOJ > 10TRUE            -0.24619    0.05506  -4.472 7.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8873.4  on 7700  degrees of freedom
## Residual deviance: 8273.8  on 7693  degrees of freedom
##   (460 observations deleted due to missingness)
## AIC: 8289.8
```

```
##
## Number of Fisher Scoring iterations: 4
```

Model bm4

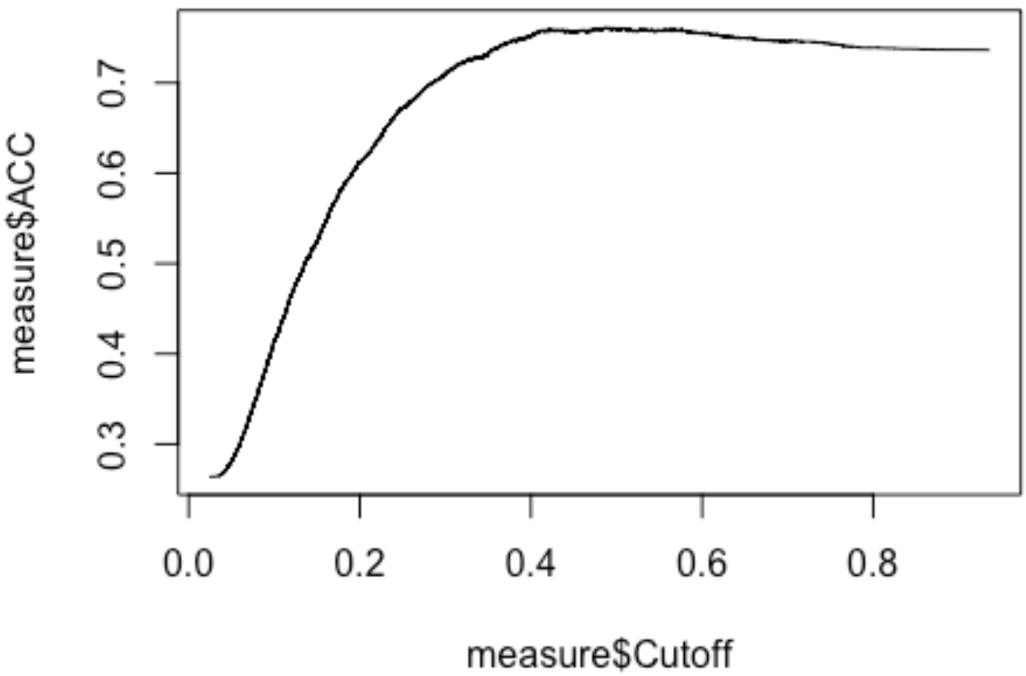
Model bm4 removes all factors which theoretical effects are unknown as target impact are known to have an impact on target variable.

```
bm4 <- glm(formula = TARGET_FLAG ~ KIDSDRIV +  MSTATUS + SEX + EDUCATION + TR
AVTIME + CAR_USE + TIF + RED_CAR + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS, f
amily = "binomial", data = train)
(bm4sum <- summary(bm4))

##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + MSTATUS + SEX + EDUCATION +
##     TRAVTIME + CAR_USE + TIF + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0361  -0.7568  -0.5207   0.8272   2.5183
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.014e+00  1.221e-01  -8.305  < 2e-16 ***
## KIDSDRIV                6.725e-01  8.679e-02   7.749 9.27e-15 ***
## MSTATUS                -6.481e-01  5.529e-02 -11.721  < 2e-16 ***
## SEX                    -2.730e-01  7.642e-02  -3.573 0.000353 ***
## EDUCATIONBachelors     -6.727e-01  8.720e-02  -7.714 1.21e-14 ***
## EDUCATIONMasters       -6.948e-01  9.408e-02  -7.386 1.52e-13 ***
## EDUCATIONPhD           -9.725e-01  1.250e-01  -7.782 7.13e-15 ***
## EDUCATIONz_High School -1.088e-01  8.383e-02  -1.297 0.194488
## TRAVTIME                6.898e-03  1.705e-03   4.045 5.24e-05 ***
## CAR_USE                 6.692e-01  5.987e-02  11.177  < 2e-16 ***
## TIF                    -2.786e-01  3.890e-02  -7.162 7.94e-13 ***
## RED_CAR                 1.686e-02  8.144e-02   0.207 0.835969
## OLDCLAIM               -1.965e-05  3.865e-06  -5.083 3.71e-07 ***
## CLM_FREQ                8.240e-01  5.949e-02  13.851  < 2e-16 ***
## REVOKED                 1.069e+00  8.643e-02  12.363  < 2e-16 ***
## MVR_PTS                 3.671e-01  3.926e-02   9.351  < 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: 9418.0  on 8160  degrees of freedom
## Residual deviance: 8166.7  on 8145  degrees of freedom
## AIC: 8198.7
```

```
##
## Number of Fisher Scoring iterations: 4
```

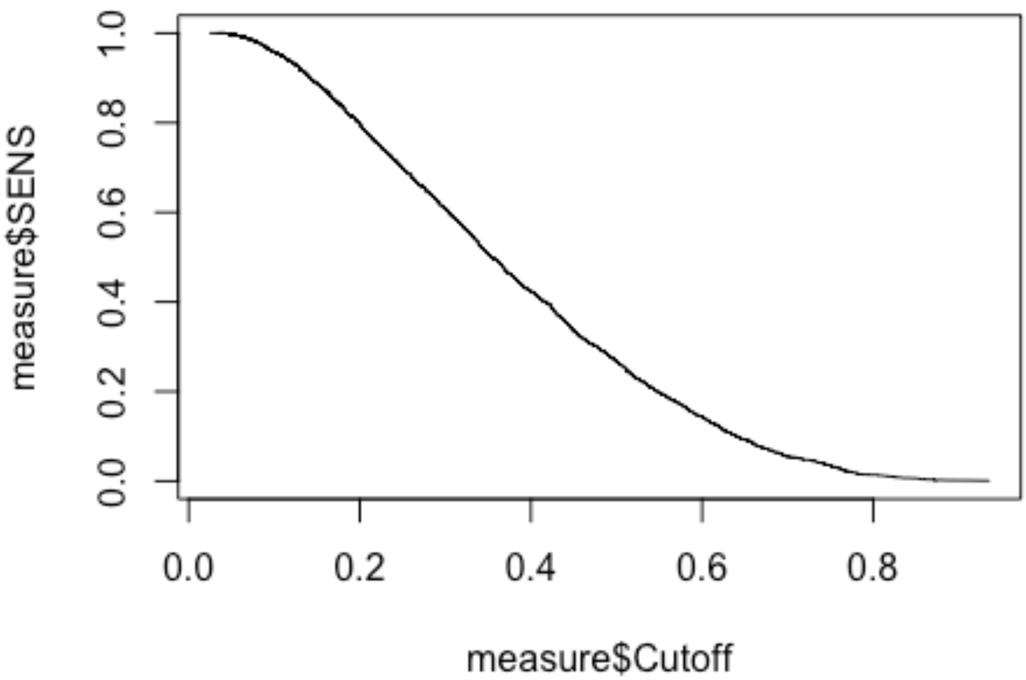
4. SELECT MODELS (25 Points) Decide on the criteria for selecting the best multiple linear regression model and the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models. For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. For the binary logistic regression model, will you use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.

```
besttrainmodel <- bm4

#Accuracy
class<-besttrainmodel$y
score<-besttrainmodel$fitted.values
measure<-measureit(score=score,class=class,measure=c("ACC", "SENS", "FSCR", "
SPEC", "PREC"))
plot(measure$ACC~measure$Cutoff, type= "l")
```

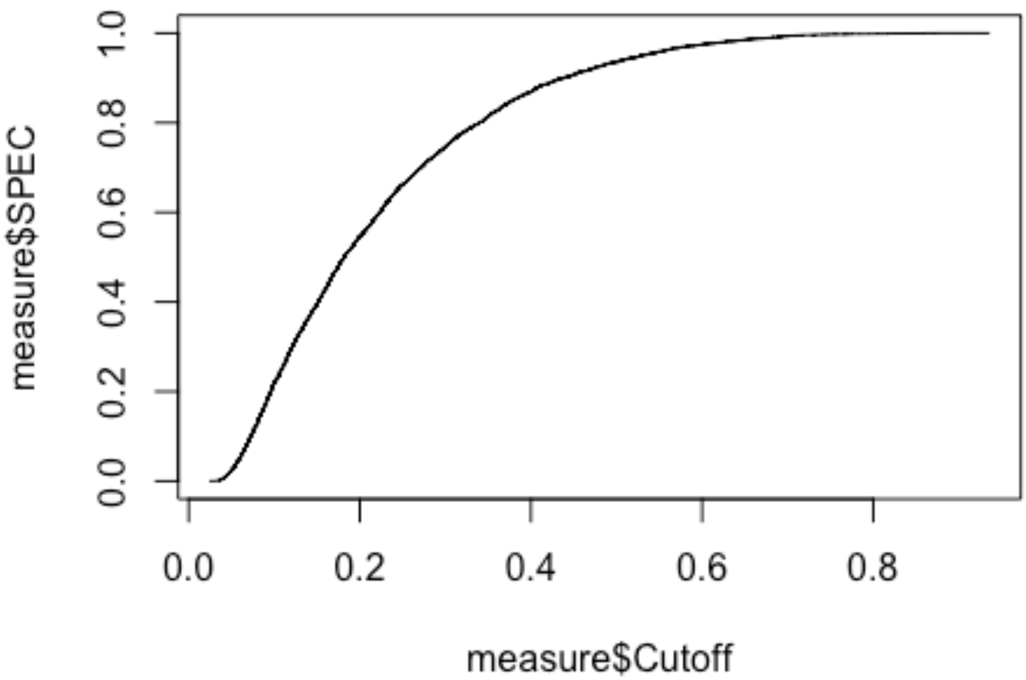


Sensitivity (also called the true positive rate, the recall, or probability of detection) measures the proportion of actual positives that are correctly identified.

```
#Sensitivity
plot(measure$SENS~measure$Cutoff, type= "l")
```




```r
#Specificity
plot(measure$SPEC~measure$Cutoff, type= "l")
```



Calculating the exponentials:

```
#calculating the exponentials
exp(besttrainmodel$coefficients)

##          (Intercept)             KIDSDRIV              MSTATUS
##            0.3628748            1.9591283            0.5230441
##                  SEX    EDUCATIONBachelors      EDUCATIONMasters
##            0.7610587            0.5103147            0.4991651
##         EDUCATIONPhD EDUCATIONz_High School             TRAVTIME
##            0.3781373            0.8969392            1.0069216
##              CAR_USE                  TIF              RED_CAR
##            1.9525921            0.7568076            1.0170059
##             OLDCLAIM             CLM_FREQ              REVOKED
##            0.9999804            2.2795856            2.9112727
##              MVR_PTS
##            1.4435541
```

Calculating the distribution:

```
##          (Intercept)             KIDSDRIV              MSTATUS
##        -1.665074e-01         1.104631e-01        -1.064536e-01
##                  SEX    EDUCATIONBachelors      EDUCATIONMasters
##        -4.484966e-02        -1.105006e-01        -1.141292e-01
##         EDUCATIONPhD EDUCATIONz_High School             TRAVTIME
```

```
##         -1.597401e-01          -1.786584e-02           1.133015e-03
##               CAR_USE                    TIF                RED_CAR
##          1.099142e-01          -4.576975e-02           2.769864e-03
##              OLDCLAIM               CLM_FREQ                REVOKED
##         -3.227031e-06           1.353472e-01           1.755240e-01
##               MVR_PTS
##          6.030029e-02
```

```r
eva$HOME_VAL <- as.numeric(gsub('[$,]', '', eva$HOME_VAL))
# trigger a dummy variable if NA is present
eva$HOME_VAL_MISSING <- ifelse(is.na(eva$HOME_VAL), 1, 0)
# imputing NA to mean
eva$HOME_VAL[is.na(eva$HOME_VAL)] <- mean(eva$HOME_VAL, na.rm=TRUE)

eva$INCOME <- as.numeric(gsub('[$,]', '', eva$INCOME))
# trigger a dummy variable if NA is present
eva$INCOME_MISSING <- ifelse(is.na(eva$INCOME), 1, 0)
# imputing NA to mean
eva$INCOME[is.na(eva$INCOME)] <- mean(eva$INCOME, na.rm=TRUE)

# trigger a dummy variable if NA is present
eva$CAR_AGE_MISSING <- ifelse(is.na(eva$CAR_AGE), 1, 0)
# imputing NA to mean
eva$CAR_AGE[is.na(eva$CAR_AGE)] <- mean(eva$CAR_AGE, na.rm=TRUE)

# trigger a dummy variable if NA is present
eva$YOJ_MISSING <- ifelse(is.na(eva$YOJ), 1, 0)
# imputing NA to mean
eva$YOJ[is.na(eva$YOJ)] <- mean(eva$YOJ, na.rm=TRUE)

summary(eva)
```

```
##      INDEX        TARGET_FLAG     TARGET_AMT        KIDSDRIV
##  Min.   :    3   Mode:logical   Mode:logical   Min.   :0.0000
##  1st Qu.: 2632   NA's:2141      NA's:2141      1st Qu.:0.0000
##  Median : 5224                                 Median :0.0000
##  Mean   : 5150                                 Mean   :0.1625
##  3rd Qu.: 7669                                 3rd Qu.:0.0000
##  Max.   :10300                                 Max.   :3.0000
##
##       AGE           HOMEKIDS           YOJ           INCOME
##  Min.   :17.00   Min.   :0.0000   Min.   : 0.00   Length:2141
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.00   Class :character
##  Median :45.00   Median :0.0000   Median :11.00   Mode  :character
##  Mean   :45.02   Mean   :0.7174   Mean   :10.38
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.00
##  Max.   :73.00   Max.   :5.0000   Max.   :19.00
##  NA's   :1                        NA's   :94
##    PARENT1            HOME_VAL           MSTATUS
##  Length:2141        Length:2141        Length:2141
```

```
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                           
##                                                           
##                                                           
##                                                           
##      SEX              EDUCATION              JOB              TRAVTIME     
##  Length:2141        Length:2141        Length:2141        Min.   :  5.00  
##  Class :character   Class :character   Class :character   1st Qu.: 22.00  
##  Mode  :character   Mode  :character   Mode  :character   Median : 33.00  
##                                                           Mean   : 33.15  
##                                                           3rd Qu.: 43.00  
##                                                           Max.   :105.00  
##                                                                           
##    CAR_USE            BLUEBOOK              TIF           CAR_TYPE        
##  Length:2141        Length:2141        Min.   : 1.000   Length:2141       
##  Class :character   Class :character   1st Qu.: 1.000   Class :character  
##  Mode  :character   Mode  :character   Median : 4.000   Mode  :character  
##                                        Mean   : 5.245                     
##                                        3rd Qu.: 7.000                     
##                                        Max.   :25.000                     
##                                                                           
##    RED_CAR            OLDCLAIM            CLM_FREQ       REVOKED         
##  Length:2141        Length:2141        Min.   :0.000   Length:2141       
##  Class :character   Class :character   1st Qu.:0.000   Class :character  
##  Mode  :character   Mode  :character   Median :0.000   Mode  :character  
##                                        Mean   :0.809                     
##                                        3rd Qu.:2.000                     
##                                        Max.   :5.000                     
##                                                                           
##     MVR_PTS          CAR_AGE        URBANICITY       
##  Min.   : 0.000   Min.   : 0.000   Length:2141       
##  1st Qu.: 0.000   1st Qu.: 1.000   Class :character  
##  Median : 1.000   Median : 8.000   Mode  :character  
##  Mean   : 1.766   Mean   : 8.183                     
##  3rd Qu.: 3.000   3rd Qu.:12.000                     
##  Max.   :12.000   Max.   :26.000                     
##                   NA's   :129

# transform data using log for skewed HOMEKIDS, MVR_PTS, TIF, KIDSDRIVE and C
LM_FREQ

eva$HOMEKIDS <- log(eva$HOMEKIDS+1)
eva$MVR_PTS <- log(eva$MVR_PTS+1)
eva$TIF <- log(eva$TIF+1)
eva$KIDSDRIV <- log(eva$KIDSDRIV+1)
eva$CLM_FREQ <- log(eva$CLM_FREQ+1)

# Convert variables PARENT1, MSTATUS, SEX, CAR_USE ,RED_CAR, REVOKED and URBA
NICITY to binary values(yes = 1, Commercial = 1, Highly Urban/ Urba = 1 )
```

```r
eva$PARENT1 <- if_else(eva$PARENT1 == "Yes", 1, 0)
eva$MSTATUS <- if_else(eva$MSTATUS == "Yes", 1, 0)
eva$SEX <- if_else(eva$SEX == "M", 1, 0)
eva$CAR_USE <- if_else(eva$CAR_USE == "Commercial", 1, 0)
eva$RED_CAR <- if_else(eva$RED_CAR == "yes", 1, 0)
eva$REVOKED <- if_else(eva$REVOKED == "Yes", 1, 0)
eva$URBANICITY <- if_else(eva$URBANICITY == "Highly Urban/ Urba", 1, 0)
```

Summary of the predicted values for the train data:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02429 0.12478 0.21964 0.26382 0.36475 0.93606
```