DATA621 Homework 4

Javern Wilson, Joseph Simone, Paul Perez, Jack Russo

4/26/2020

Contents

Overview

In this homework assignment, we 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 "**0**" 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.

Our 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. Only variables given in the project will be used unless new variables are derived from the original variables. Below is a short description of the variables of interest in the data set:

- `INDEX` Identification Variable (do not use)
  - **EFFECT**: None
- `TARGET_FLAG` Was Car in a crash? 1=YES 0=NO
  - **EFFECT**: None
- `TARGET_AMT` If car was in a crash, what was the cost
  - **EFFECT**: None
- `AGE` Age of Driver
  - **EFFECT**: Very young people tend to be risky. Maybe very old people also.
- `BLUEBOOK` Value of Vehicle
  - **EFFECT**: Unknown effect on probability of collision, but probably effect the payout if there is a crash
- `CAR_AGE` Vehicle Age
  - **EFFECT**: Unknown effect on probability of collision, but probably effect the payout if there is a crash

- `CAR_TYPE` Type of Car
  - **EFFECT**: Unknown effect on probability of collision, but probably effect the payout if there is a crash
- `CAR_USE` Vehicle Use
  - **EFFECT**: Commercial vehicles are driven more, so might increase probability of collision
- `CLM_FREQ` # Claims (Past 5 Years)
  - **EFFECT**: The more claims you filed in the past, the more you are likely to file in the future
- `EDUCATION` Max Education Level
  - **EFFECT**: Unknown effect, but in theory more educated people tend to drive more safely
- `HOMEKIDS` # Children at Home
  - **EFFECT**: Unknown effect
- `HOME_VAL` Home Value
  - **EFFECT**: In theory, home owners tend to drive more responsibly
- `INCOME` Income
  - **EFFECT**: In theory, rich people tend to get into fewer crashes
- `JOB` Job Category
  - **EFFECT**: In theory, white collar jobs tend to be safer
- `KIDSDRIV` # Driving Children
  - **EFFECT**: When teenagers drive your car, you are more likely to get into crashes
- `MSTATUS` Marital Status
  - **EFFECT**: In theory, married people drive more safely
- `MVR_PTS` Motor Vehicle Record Points
  - **EFFECT**: If you get lots of traffic tickets, you tend to get into more crashes
- `OLDCLAIM` Total Claims (Past 5 Years)
  - **EFFECT**: If your total payout over the past five years was high, this suggests future payouts will be high
- `PARENT1` Single Parent
  - **EFFECT**: Unknown effect
- `RED_CAR` A Red Car
  - **EFFECT**: Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
- `REVOKED` License Revoked (Past 7 Years)
  - **EFFECT**: If your license was revoked in the past 7 years, you probably are a more risky driver.
- `SEX` Gender
  - **EFFECT**: Urban legend says that women have less crashes then men. Is that true?
- `TIF` Time in Force
  - **EFFECT**: People who have been customers for a long time are usually more safe.
- `TRAVTIME` Distance to Work

  - Long drives to work usually suggest greater risk
- `URBANICITY` Home/Work Area
  - **EFFECT**: Unknown
- `YOJ` Years on Job
  - **EFFECT**: People who stay at a job for a long time are usually more safe

```
library(tidyverse)

library(caret)

library(e1071)

library(pracma)

library(pROC)

library(psych)

library(kableExtra)

library(Hmisc)

library(VIF)

library(FactoMineR)

library(corrplot)

library(purrr)

library(dplyr)

library(MASS)

library(mice)
```

```
insurance_train <- read.csv("https://raw.githubusercontent.com/javernw/DATA621-Business-Analytics-and-Da

insurance_eval <- read.csv("https://raw.githubusercontent.com/javernw/DATA621-Business-Analytics-and-Dat

columns <- colnames(insurance_train)

target <- "TARGET_FLAG"

inputs <- columns[!columns %in% c(target,"INDEX")]
```

Preview

```
insurance_train %>%  tibble(head(10))
```

```
## # A tibble: 8,161 x 2
##    .$INDEX $TARGET_FLAG $TARGET_AMT $KIDSDRIV  $AGE $HOMEKIDS  $YOJ $INCOME
##      <int>        <int>       <dbl>     <int> <int>     <int> <int> <fct>
##  1       1            0           0         0    60         0    11 "$67,3~
##  2       2            0           0         0    43         0    11 "$91,4~
##  3       4            0           0         0    35         1    10 "$16,0~
##  4       5            0           0         0    51         0    14 ""
##  5       6            0           0         0    50         0    NA "$114,~
##  6       7            1        2946         0    34         1    12 "$125,~
##  7       8            0           0         0    54         0    NA "$18,7~
##  8      11            1        4021         1    37         2    NA "$107,~
##  9      12            1        2501         0    34         0    10 "$62,9~
## 10      13            0           0         0    50         0     7 "$106,~
## # ... with 8,151 more rows, and 19 more variables: $PARENT1 <fct>,
## #   $HOME_VAL <fct>, $MSTATUS <fct>, $SEX <fct>, $EDUCATION <fct>, $JOB <fct>,
## #   $TRAVTIME <int>, $CAR_USE <fct>, $BLUEBOOK <fct>, $TIF <int>,
## #   $CAR_TYPE <fct>, $RED_CAR <fct>, $OLDCLAIM <fct>, $CLM_FREQ <int>,
## #   $REVOKED <fct>, $MVR_PTS <int>, $CAR_AGE <int>, $URBANICITY <fct>,
## #   `head(10)` <dbl>
```

DATA EXPLORATION

Structure

```
glimpse(insurance_train)
```

```
## Observations: 8,161
## Variables: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20...
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 402...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53,...
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2...
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0...
## $ INCOME      <fct> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,...
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No...
## $ HOME_VAL    <fct> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "...
## $ MSTATUS     <fct> z_No, z_No, Yes, Yes, Yes, z_No, Yes, Yes, z_No, z_No, ...
## $ SEX         <fct> M, M, z_F, M, z_F, z_F, z_F, M, z_F, M, z_F, z_F, M, M,...
## $ EDUCATION   <fct> PhD, z_High School, z_High School, <High School, PhD, B...
## $ JOB         <fct> Professional, z_Blue Collar, Clerical, z_Blue Collar, D...
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, ...
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commerc...
## $ BLUEBOOK    <fct> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "...
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, ...
## $ CAR_TYPE    <fct> Minivan, Minivan, z_SUV, Minivan, z_SUV, Sports Car, z_...
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes...
## $ OLDCLAIM    <fct> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0",...
```

```
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0...
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, ...
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, ...
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, ...
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/...
```

```
summary(insurance_train)
```

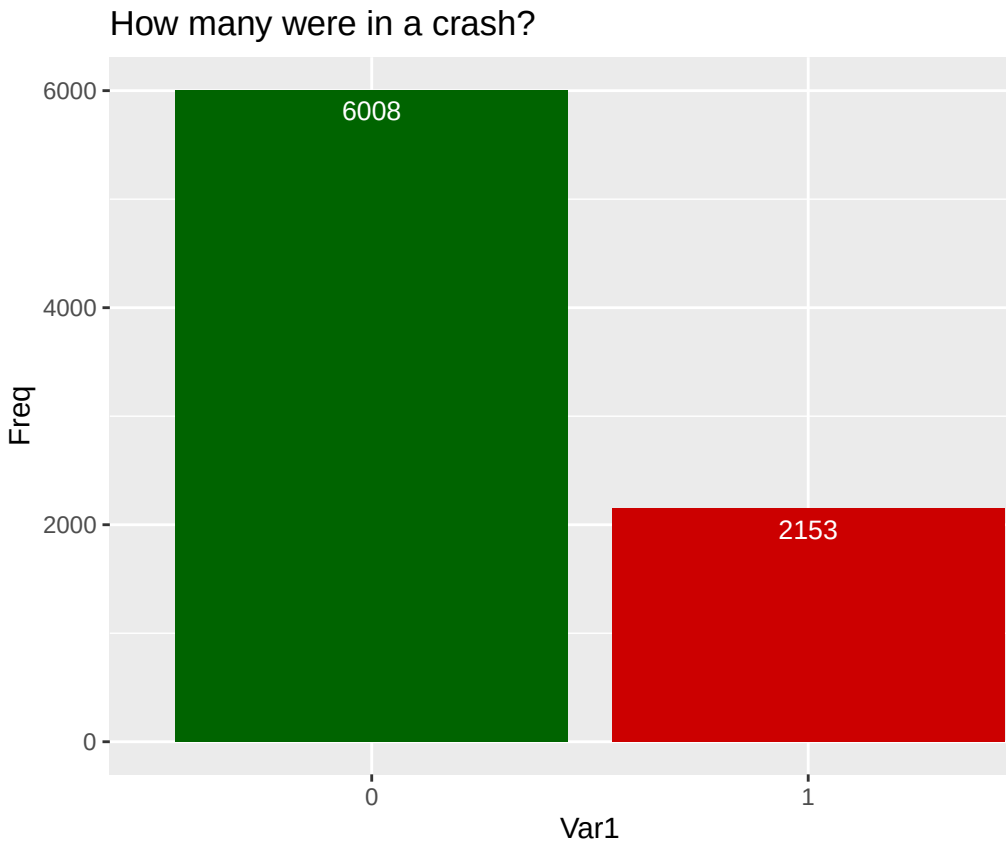
```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ             INCOME     PARENT1
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   $0       : 615   No :7084
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0            : 445   Yes:1077
##  Median :45.00   Median :0.0000   Median :11.0   $26,840 :   4
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   $48,509 :   4
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   $61,790 :   4
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   $107,375:   3
##  NA's   :6                        NA's   :454    (Other) :7086
##      HOME_VAL    MSTATUS      SEX              EDUCATION
##  $0      :2294   Yes :4894   M  :3786   <High School :1203
##          : 464   z_No:3267   z_F:4375   Bachelors    :2242
##  $111,129:   3                          Masters      :1658
##  $115,249:   3                          PhD          : 728
##  $123,109:   3                          z_High School:2330
##  $153,061:   3
##  (Other) :5391
##              JOB          TRAVTIME           CAR_USE        BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   $1,500 : 157
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   $6,000 :  34
##  Professional :1117   Median : 33.00                     $5,800 :  33
##  Manager      : 988   Mean   : 33.49                     $6,200 :  33
##  Lawyer       : 835   3rd Qu.: 44.00                     $6,400 :  31
##  Student      : 712   Max.   :142.00                     $5,900 :  30
##  (Other)      :1413                                      (Other):7843
##       TIF                CAR_TYPE    RED_CAR       OLDCLAIM       CLM_FREQ
##  Min.   : 1.000   Minivan    :2145   no :5783   $0      :5009   Min.   :0.0000
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :   4   1st Qu.:0.0000
##  Median : 4.000   Pickup     :1389              $1,391 :   4   Median :0.0000
##  Mean   : 5.351   Sports Car : 907              $4,263 :   4   Mean   :0.7986
##  3rd Qu.: 7.000   Van        : 750              $1,105 :   3   3rd Qu.:2.0000
##  Max.   :25.000   z_SUV      :2294              $1,332 :   3   Max.   :5.0000
##                                                 (Other):3134
##  REVOKED       MVR_PTS          CAR_AGE                      URBANICITY
##  No :7161   Min.   : 0.000   Min.   :-3.000   Highly Urban/ Urban  :6492
##  Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000   z_Highly Rural/ Rural:1669
##             Median : 1.000   Median : 8.000
##             Mean   : 1.696   Mean   : 8.328
##             3rd Qu.: 3.000   3rd Qu.:12.000
```

```
##              Max.   :13.000   Max.   :28.000
##                                NA's   :510
```

We have some missing values in the training set. Especially variables such as `CAR_AGE`. We will illustrate a graphical view was we move forward.

Targets

```
tf <- table(insurance_train$TARGET_FLAG) %>% data.frame()

ggplot(tf, aes(x = Var1, y = Freq, fill = Var1)) + geom_bar(stat = "identity") + scale_fill_manual(name
```



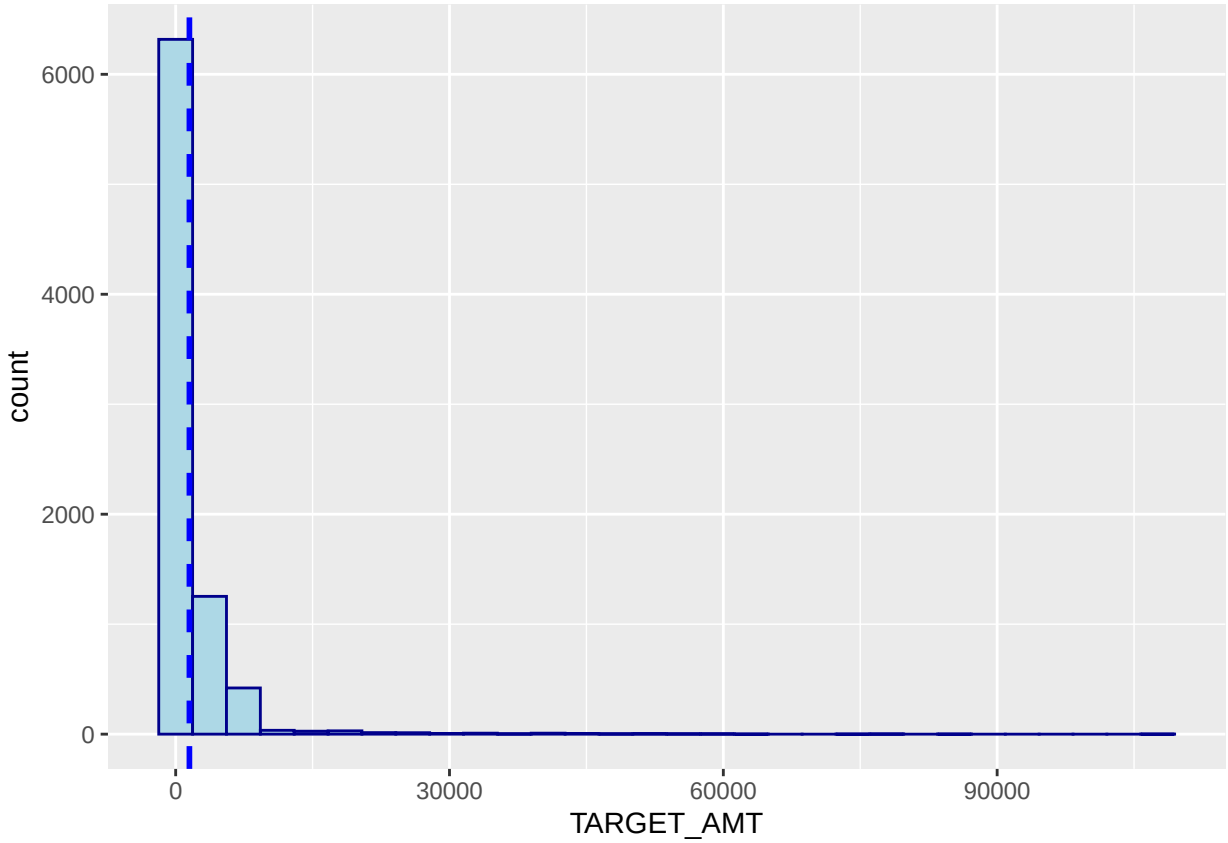
How many were in a car crash?

We can obviously see that majority of the observations (`73.6`%) in the trainig set were not involved in an accident.

```
ggplot(insurance_train, aes(x=TARGET_AMT)) + geom_histogram(color="darkblue", fill="lightblue") + geom_v
```

What was the cost?

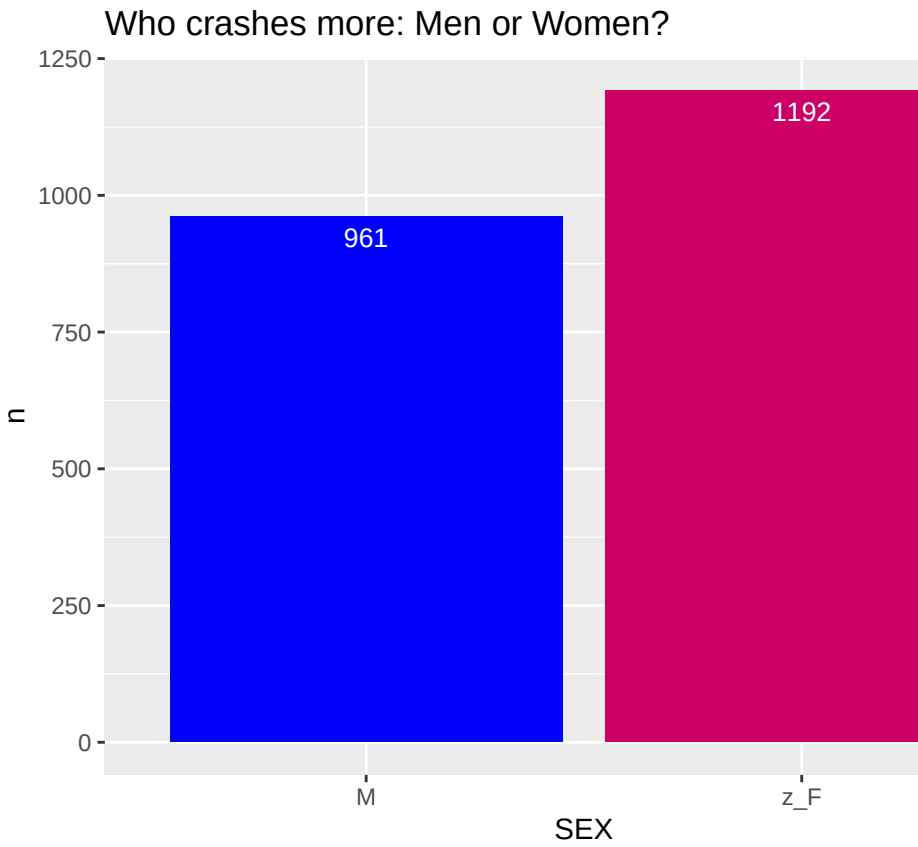
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Accidents vs Gender

```
mvw <- insurance_train %>% dplyr::select(SEX, TARGET_FLAG) %>% count(SEX, TARGET_FLAG) %>% filter(TARGE
```

```
ggplot(mvw, aes(x = SEX, y = n, fill = SEX)) + geom_bar(stat = "identity") + scale_fill_manual(values=c
```

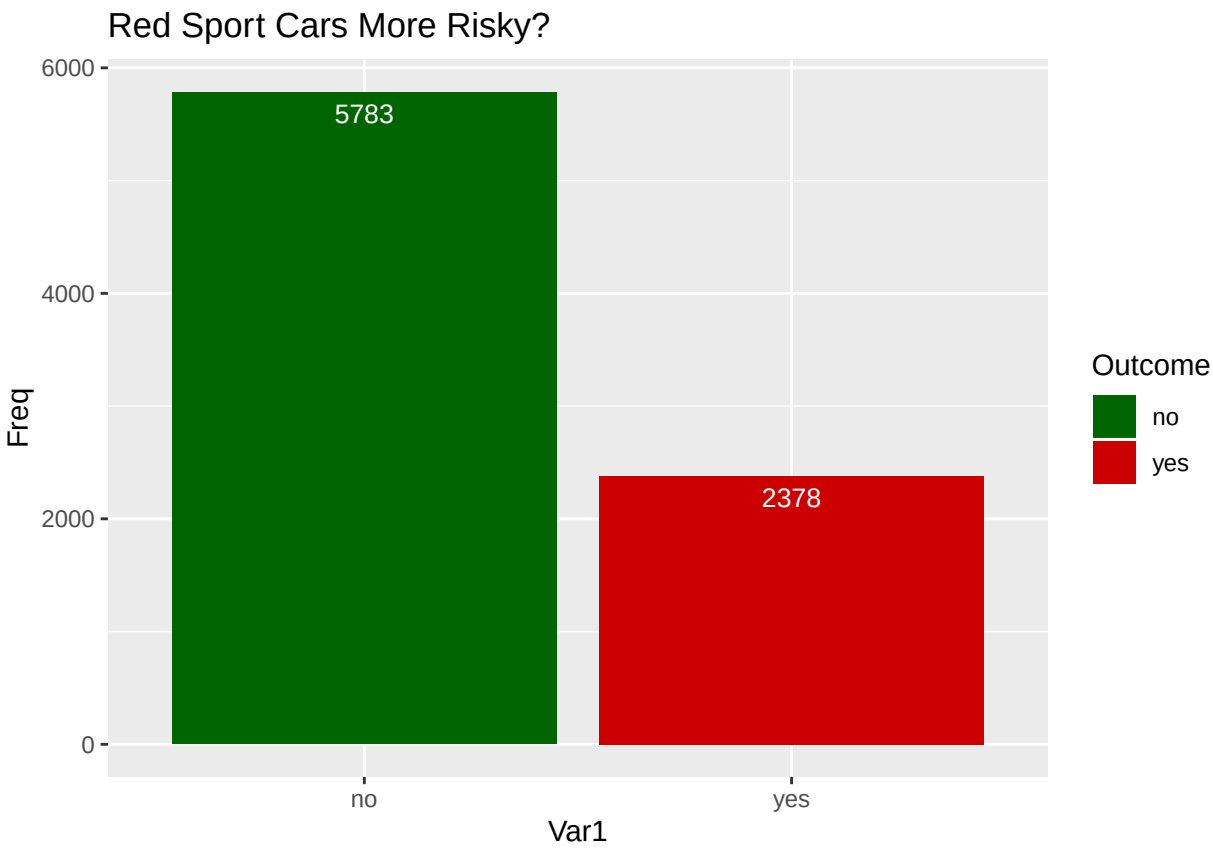


Who crashes more? Men or Women?

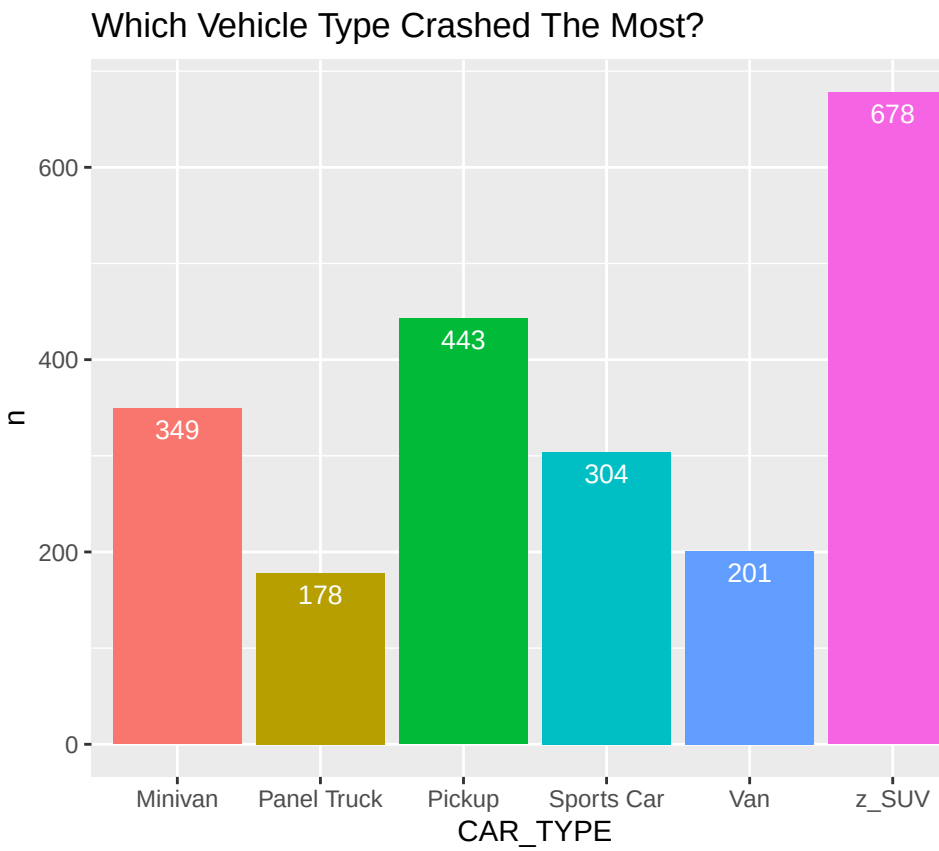
Vehicle VS Accidents

####Are Red Sport Cars more risky?

```
red_cars <- table(insurance_train$RED_CAR) %>% data.frame()

ggplot(red_cars, aes(x = Var1, y = Freq, fill = Var1)) + geom_bar(stat = "identity") + scale_fill_manual
```




```
insurance_train %>% dplyr::select(CAR_TYPE, TARGET_FLAG) %>%
 count(CAR_TYPE, TARGET_FLAG) %>%
 filter(TARGET_FLAG == 1) %>%
 ggplot(aes(x = CAR_TYPE, y = n, fill = CAR_TYPE)) +
 geom_bar(stat = "identity") +
 geom_text(aes(label=n), vjust=1.6, color="white", size=3.5) +
 ggtitle("Which Vehicle Type Crashed The Most?")
```



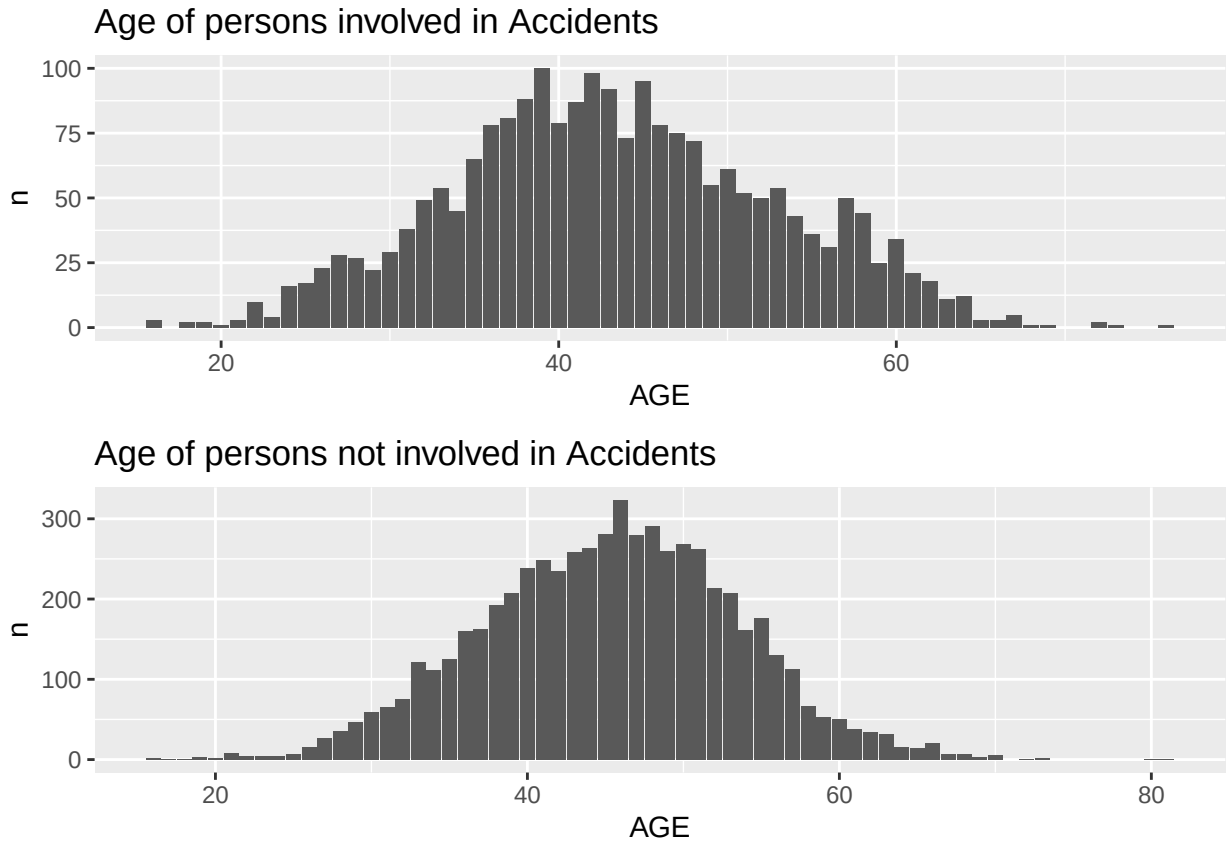
Type of vehicle with most accidents

DISTRIBUTION OF AGE

```
age_crash <- insurance_train %>% dplyr::select(AGE, TARGET_FLAG) %>%
  count(AGE, TARGET_FLAG) %>%
  filter(TARGET_FLAG == 1) %>% ggplot(aes(x = AGE, y = n)) + geom_bar(stat = "identity") + labs(title =


age_no_crash <- insurance_train %>% dplyr::select(AGE, TARGET_FLAG) %>%
  count(AGE, TARGET_FLAG) %>%
  filter(TARGET_FLAG == 0) %>%  ggplot(aes(x = AGE, y = n)) + geom_bar(stat = "identity") + labs(title =



gridExtra::grid.arrange(age_crash, age_no_crash, nrow = 2)
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



Inferences:

- There is a more normal distribution of age range for those involved in accidents than with those who were not.
- In the second plot it shows as people get older, they become more responsible with driving. Frequencies are higher.
- Younger folks become involved in accidents accodring to the first histogram.

Customers

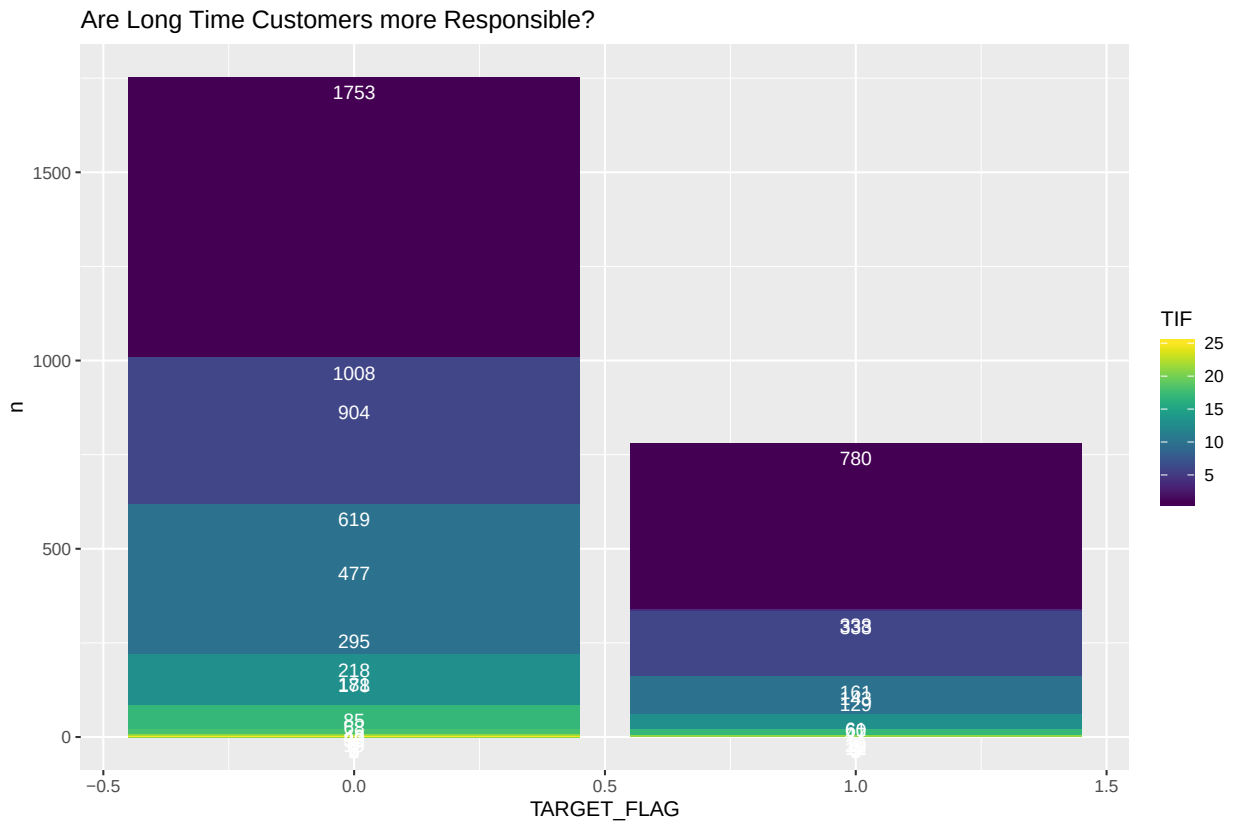
Who are more responsible?

```
library(viridis)
```

```
## Warning: package 'viridis' was built under R version 3.6.3
```

```
## Loading required package: viridisLite
```

```
tif <- insurance_train %>%
  dplyr::select(TARGET_FLAG, TIF) %>%
  count(TARGET_FLAG, TIF) %>%
  ggplot(aes(x = TARGET_FLAG, y = n, fill = TIF)) +
  geom_bar(position = position_dodge(), stat = "identity") +
  scale_fill_viridis(discrete = F) +
  geom_text(aes(label=n), vjust=1.6,
            color="white", size=3.5) +
  ggtitle("Are Long Time Customers more Responsible?")
tif
```



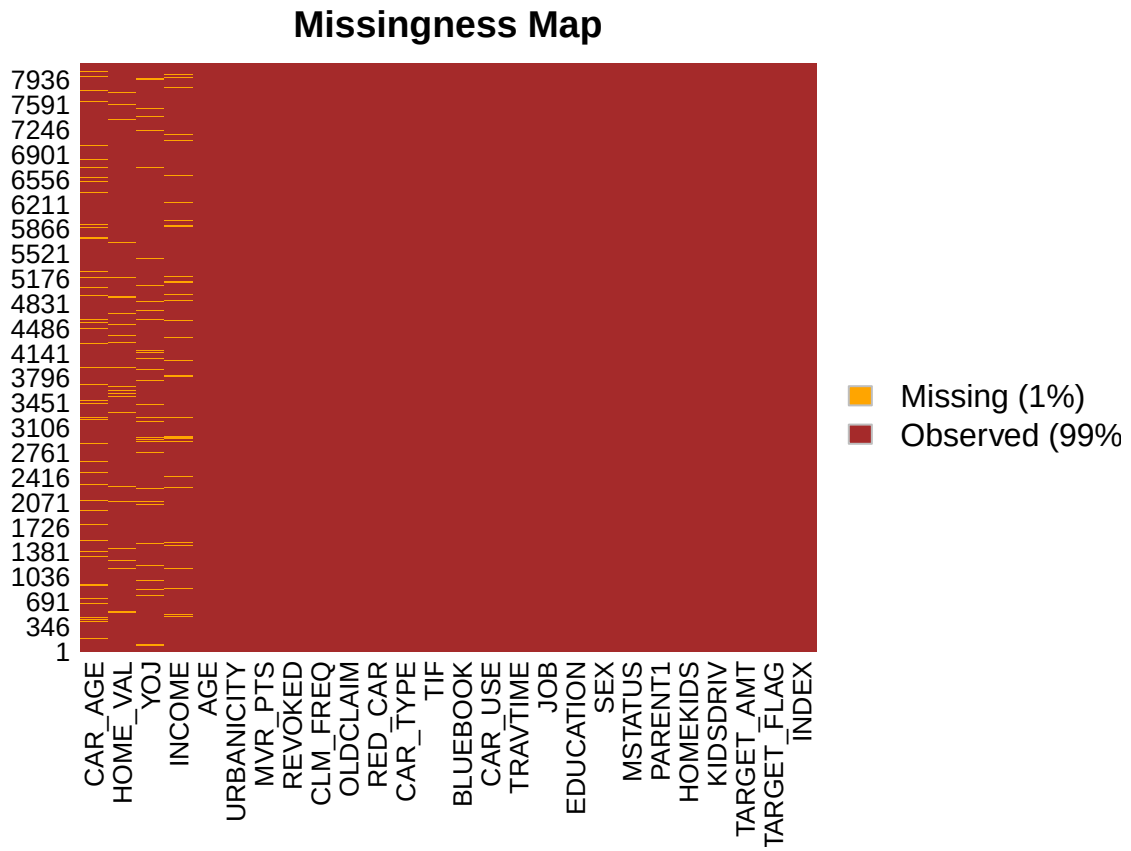
This confirms the asumption made earlier when introducing the variables. Customers who are with the company tend to be more responsible with driving.

MISSING VALUES

Before we check for missing values, some of the variables that should be as numeric and are classified as charactor variables need to be changed. Some charactors will be removed from the affected columns to convert values to numeric nature. This way the missing values visualization will be more accurate.

```
insurance_train$INCOME  <- gsub( "\\$", "", insurance_train$INCOME)

insurance_train$INCOME  <- gsub( "\\,", "", insurance_train$INCOME)

insurance_train$INCOME  <- as.numeric(insurance_train$INCOME)


insurance_train$HOME_VAL  <- gsub( "\\$", "", insurance_train$HOME_VAL)

insurance_train$HOME_VAL  <- gsub( "\\,", "", insurance_train$HOME_VAL)

insurance_train$HOME_VAL  <- as.numeric(insurance_train$HOME_VAL)


insurance_train$BLUEBOOK  <- gsub( "\\$", "", insurance_train$BLUEBOOK)

insurance_train$BLUEBOOK  <- gsub( "\\,", "", insurance_train$BLUEBOOK)

insurance_train$BLUEBOOK  <- as.numeric(insurance_train$BLUEBOOK)


insurance_train$OLDCLAIM  <- gsub( "\\$", "", insurance_train$OLDCLAIM)

insurance_train$OLDCLAIM  <- gsub( "\\,", "", insurance_train$OLDCLAIM)

insurance_train$OLDCLAIM  <- as.numeric(insurance_train$OLDCLAIM)
```

```
Amelia::missmap(insurance_train, col = c("orange", "brown"))
```

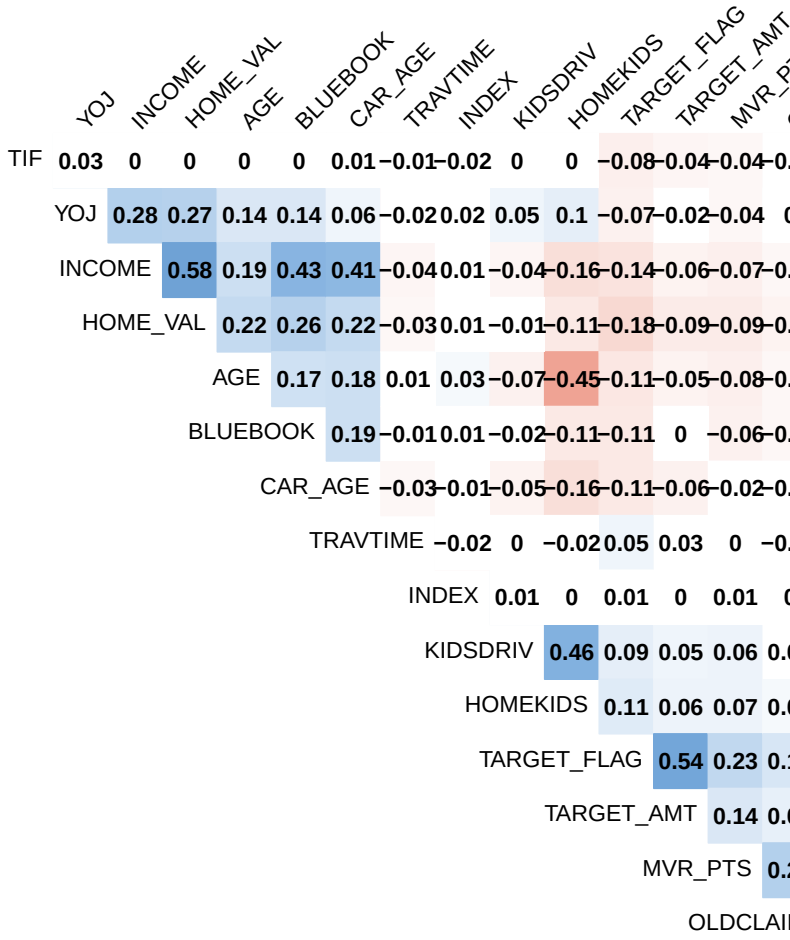


Only four variables have missing values. The missing values ratio is clearly insignificant as the report shows 1% missing values.

Correlation

```
num_pred <- dplyr::select_if(insurance_train, is.numeric)


np_corr <- cor(num_pred, use = "na.or.complete")

p_matrix <- rcorr(as.matrix(num_pred))


col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))

corrplot(np_corr, method="color", col=col(200),

         type="upper", order="hclust",

         addCoef.col = "black",
```

```
        tl.col="black", tl.srt=45,

        p.mat = p_matrix$P, sig.level = 0.01, insig = "blank",

        diag=FALSE

        )
```



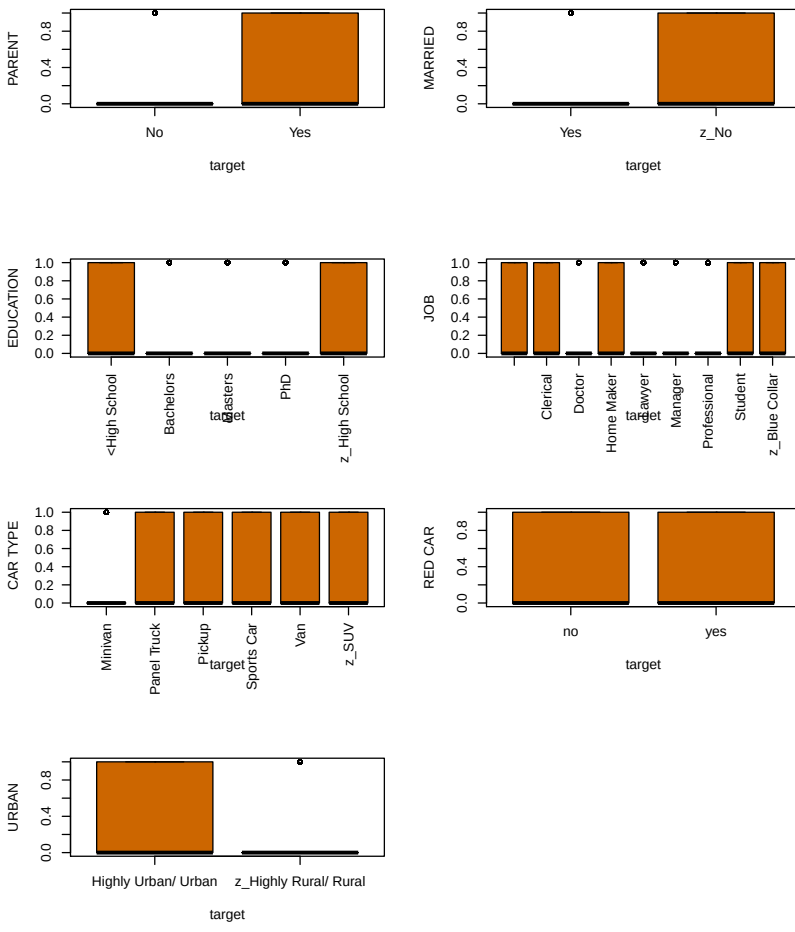
Numeric Predictors VS Both Targets

In this corrplot, only significant relationships are highlighted, that is, with a significance below `0.01`.

From the results of the corrplot, for instance, the target variables have a significant, moderat3ely but positive relationship scored at `0.54`. There also some moderate correlation between the variables `INCOME` and `HOME_VAL` with `0.58`. We can watch out for this was we progress on.

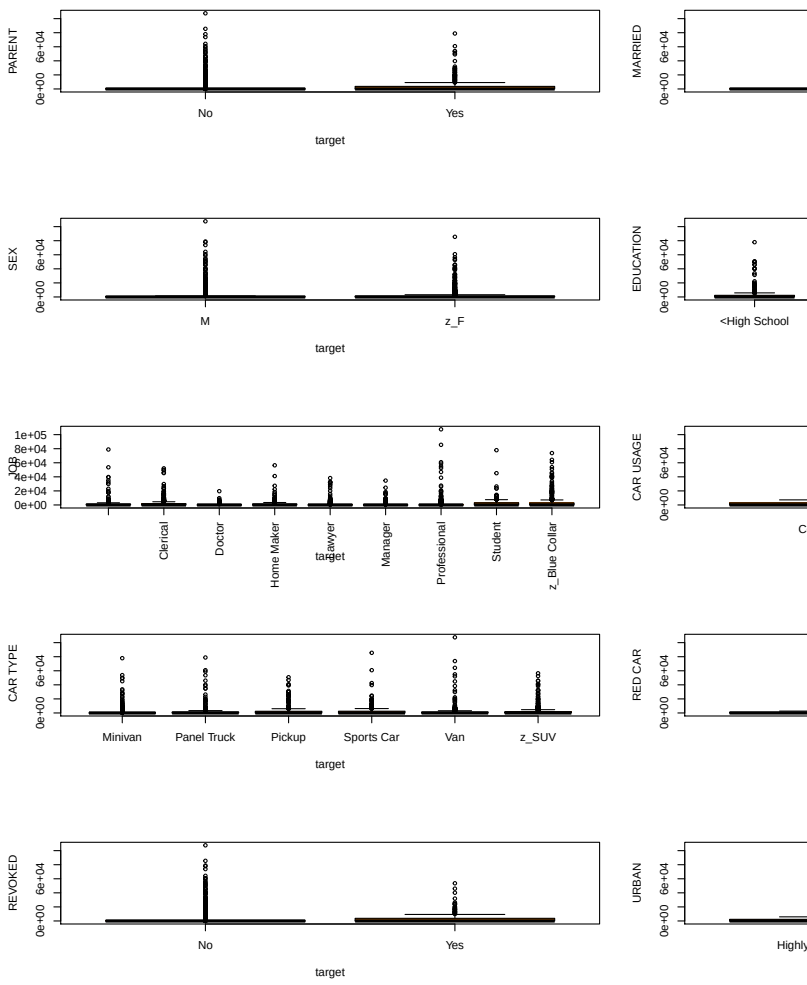
```
char_pred <- dplyr::select_if(insurance_train, is.factor)

par(mfrow = c(4,3))

boxplot(TARGET_FLAG~PARENT1, ylab="PARENT", xlab= "target", col="#CC6600", data = insurance_train)
```

```r
boxplot(TARGET_FLAG~MSTATUS, ylab="MARRIED", xlab= "target", col="#CC6600", data = insurance_train)

boxplot(TARGET_FLAG~SEX, ylab="SEX", xlab= "target", col="#CC6600", data = insurance_train)

boxplot(TARGET_FLAG~EDUCATION, ylab="EDUCATION", xlab= "target", col="#CC6600", data = insurance_train,

boxplot(TARGET_FLAG~JOB, ylab="JOB", xlab= "target", col="#CC6600", data = insurance_train, las=2)

boxplot(TARGET_FLAG~CAR_USE, ylab="CAR USAGE", xlab= "target", col="#CC6600", data = insurance_train)

boxplot(TARGET_FLAG~CAR_TYPE, ylab=" CAR TYPE", xlab= "target", col="#CC6600", data = insurance_train, l

boxplot(TARGET_FLAG~RED_CAR, ylab="RED CAR", xlab= "target", col="#CC6600", data = insurance_train)

boxplot(TARGET_FLAG~REVOKED, ylab="REVOKED", xlab= "target", col="#CC6600", data = insurance_train)

boxplot(TARGET_FLAG~URBANICITY, ylab="URBAN", xlab= "target", col="#CC6600", data = insurance_train)
```



Non_numeric Predictors VS TARGET_FLAG

```r
par(mfrow = c(5,2))
boxplot(TARGET_AMT~PARENT1, ylab="PARENT", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~MSTATUS, ylab="MARRIED", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~SEX, ylab="SEX", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~EDUCATION, ylab="EDUCATION", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~JOB, ylab="JOB", xlab= "target", col="#CC6600", data = insurance_train, las=2)
boxplot(TARGET_AMT~CAR_USE, ylab="CAR USAGE", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~CAR_TYPE, ylab=" CAR TYPE", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~RED_CAR, ylab="RED CAR", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~REVOKED, ylab="REVOKED", xlab= "target", col="#CC6600", data = insurance_train)
boxplot(TARGET_AMT~URBANICITY, ylab="URBAN", xlab= "target", col="#CC6600", data = insurance_train)
```



Non_numeric Predictors VS TARGET_AMT

DATA PREPARATION

Handling Missing Values

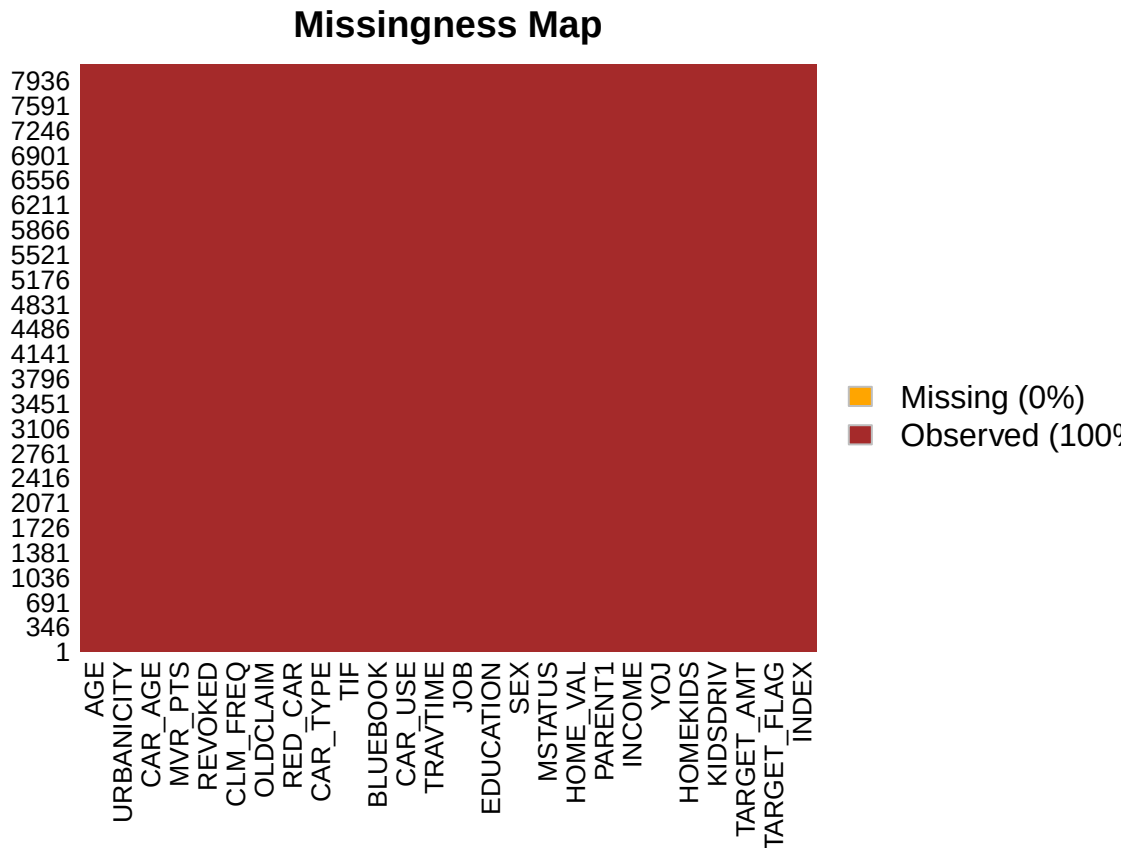
In this case, instead of just removing observations with missing values, we will impute the mean for each predictor variable except target variables.

```
insurance_train$CAR_AGE <- replace(insurance_train$CAR_AGE, -3, 0)

insurance_train$CAR_AGE[is.na(insurance_train$CAR_AGE)] <- mean(insurance_train$CAR_AGE, na.rm=TRUE)

insurance_train$HOME_VAL[is.na(insurance_train$HOME_VAL)] <- mean(insurance_train$HOME_VAL, na.rm=TRUE)

insurance_train$YOJ[is.na(insurance_train$YOJ)] <- mean(insurance_train$YOJ, na.rm=TRUE)

insurance_train$INCOME[is.na(insurance_train$INCOME)] <- mean(insurance_train$INCOME, na.rm=TRUE)
```

```
#insurance_train$CAR_AGE <- replace_na_mean(insurance_train$CAR_AGE)

#insurance_train$HOME_VAL <- replace_na_mean(insurance_train$HOME_VAL)

#insurance_train$YOJ <- replace_na_mean(insurance_train$YOJ)

#insurance_train$INCOME <- replace_na_mean(insurance_train$INCOME)
```

Missingmapness for original dataset

```
Amelia::missmap(insurance_train, col = c("orange", "brown"))
```



The variable `INDEX` from original dataset will also be removed as is it not relevant to prediction of the target variables.

```
insurance_train <- insurance_train[,-1]
```

Data Transformations

After exploration of the data, we thought certain conversions were in order:

- Convert INCOME to numeric
- Convert PARENT1 to binary (1/0)

- Convert HOME_VAL to binary
- Convert MSTATUS to binary (1/0)
- Convert SEX to Flag (IS_MALE)
- Convert CAR_USE to binary (1/0)
- Convert BLUEBOOK to numeric
- Parse CAR_TYPE into: CAR_PANEL_TRUCK,CAR_PICKUP,CAR_SPORTS_CAR,CAR_VAN,CAR_SUV
- Convert RED_CAR to binary (1/0)
- Convert OLDCLAIM to numeric
- Convert REVOKED to binary (1/0)
- Convert URBANICITY to binary (1/0)

All binary variables will have suffixed of "_BIN"

```
string_split<- function(y, z){
  temp <- as.numeric(gsub("[\\$,]","", y))
  if (!is.na(temp) && temp == 0 && z) { NA } else {temp}
}
```

```
transformation <- function(value){
  column_outputs<- c("TARGET_FLAG","TARGET_AMT","AGE", "YOJ", "CAR_AGE","KIDSDRIV","HOMEKIDS","TRAVTIME

  # Convert INCOME to numeric, replace 0 for NA
  value['INCOME'] <- string_split(value['INCOME'],TRUE)
  value['INCOME'] <- replace(value['INCOME'], is.na(value['INCOME']), 0)
  column_outputs <- c(column_outputs,'INCOME')

  # Convert PARENT1 to flag (1/0)
  value['PARENT1_BIN'] <- if (value['PARENT1']=="Yes") {1} else {0}
  column_outputs <- c(column_outputs,'PARENT1_BIN')

  # Convert HOME_VAL to binary(1/0)
```

```
value['HOME_VAL_BIN'] <- if (is.na(string_split(value['HOME_VAL'],TRUE))) {1} else {0}

column_outputs <- c(column_outputs,'HOME_VAL_BIN')


# Convert MSTATUS to binary  IS_SINGLE (1/0)

value['MSTATUS_BIN'] <- if (value['MSTATUS']=="z_No") {1} else {0}

column_outputs <- c(column_outputs,'MSTATUS_BIN')


# Convert SEX to binary (IS_MALE)

value['IS_MALE_BIN'] <- if (value['SEX']=="M") {1} else {0}

column_outputs <- c(column_outputs,'IS_MALE_BIN')


# Convert CAR_USE to binary (1/0)

value['IS_COMMERCIAL_BIN'] <- if (value['CAR_USE']=="Commercial") {1} else {0}

column_outputs <- c(column_outputs,'IS_COMMERCIAL_BIN')



# Convert BLUEBOOK to numeric

value['BLUEBOOK'] <- string_split(value['BLUEBOOK'],FALSE)

column_outputs <- c(column_outputs,'BLUEBOOK')


# Convert OLDCLAIM to numeric

value['OLDCLAIM'] <- string_split(value['OLDCLAIM'],TRUE)

value['OLDCLAIM'] <- replace(value['OLDCLAIM'], is.na(value['OLDCLAIM']), 0)

column_outputs <- c(column_outputs,'OLDCLAIM')


# Breakout CAR_TYPE into:

value['CAR_PANEL_TRUCK_BIN'] <- if (value['CAR_TYPE']=="Panel Truck") {1} else {0}
```

```
  value['CAR_PICKUP_BIN'] <- if (value['CAR_TYPE']=="Pickup") {1} else {0}
  value['CAR_SPORTS_CAR_BIN'] <- if (value['CAR_TYPE']=="Sports Car") {1} else {0}
  value['CAR_VAN_BIN'] <- if (value['CAR_TYPE']=="Van") {1} else {0}
  value['CAR_SUV_BIN'] <- if (value['CAR_TYPE']=="z_SUV") {1} else {0}
  column_outputs <- c(column_outputs,'CAR_PANEL_TRUCK_BIN','CAR_PICKUP_BIN','CAR_SPORTS_CAR_BIN','CAR_V

  # Convert RED_CAR to binary(1/0)

  value['RED_CAR_BIN'] <- if (value['RED_CAR']=="yes") {1} else {0}
  column_outputs <- c(column_outputs,'RED_CAR_BIN')

  # Convert REVOKED to bianry (1/0)
  value['REVOKED_BIN'] <- if (value['REVOKED']=="Yes") {1} else {0}
  column_outputs <- c(column_outputs,'REVOKED_BIN')

  # Convert URBANICITY to bunary (1/0)
  value['IS_URBAN_BIN'] <- if (value['URBANICITY']=="Highly Urban/ Urban") {1} else {0}
  column_outputs <- c(column_outputs,'IS_URBAN_BIN')



  final <- as.numeric(value[column_outputs])
  names(final) <- column_outputs
  final
}
# form dataframe by function
transform_insurance_train<-data.frame(t(rbind(apply(insurance_train,1,transformation))))
transform_insurance_eval<-data.frame(t(rbind(apply(insurance_eval,1,transformation))))
```

```
columns <- colnames(transform_insurance_train)

target_bin <- c("TARGET_FLAG")

target_lm <- c("TARGET_AMT")

target <- c(target_bin,target_lm)

inputs_bin <- columns[grep("_BIN",columns)]

inputs_num <- columns[!columns %in% c(target,"INDEX",inputs_bin)]

inputs<- c(inputs_bin,inputs_num)
```

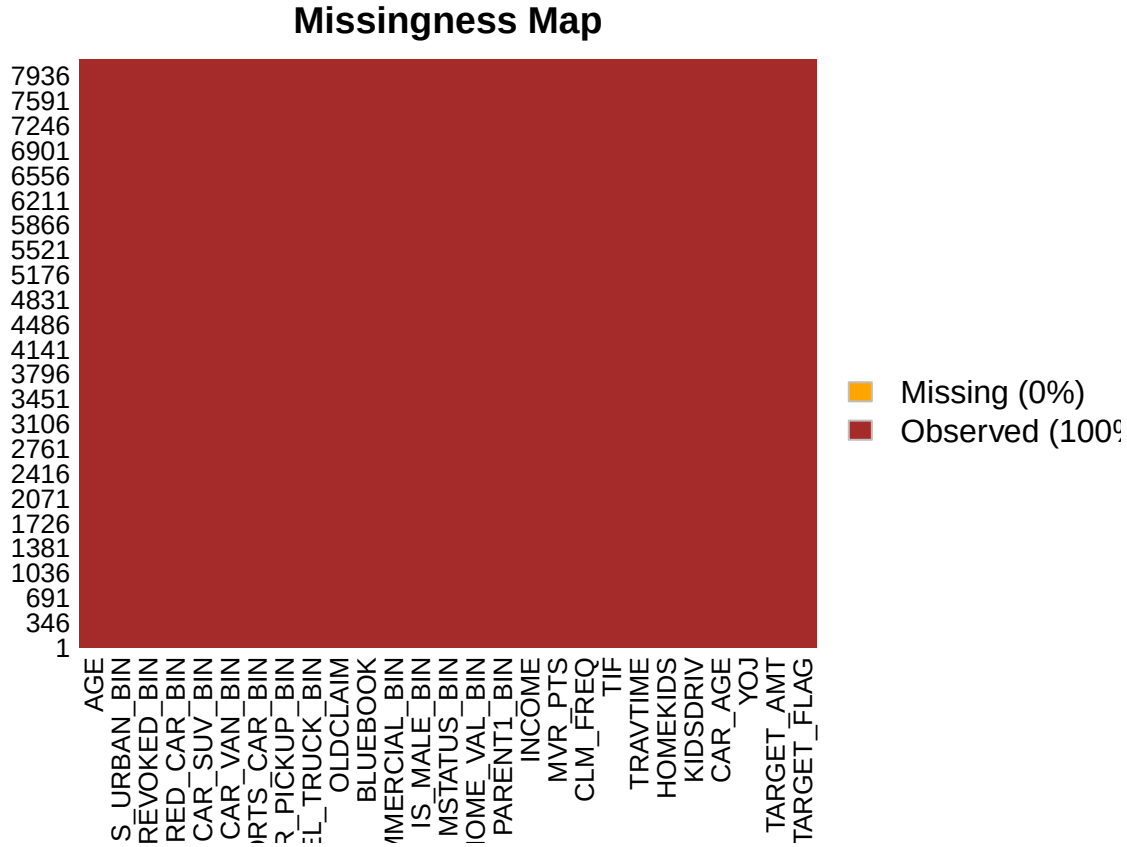
```
#temp <- mice(insurance_train[,-c(1,2)] ,m=5,maxit=50,meth='pmm',seed=500, printFlag = F)

#temp <- complete(temp)

#temp$TARGET_FLAG <- insurance_train$TARGET_FLAG

#temp$TARGET_AMT <- insurance_train$TARGET_AMT

#insurance_train <- temp
```

```
#boxcox_trans <- function(column) {

#  new_column <- column ^ boxcoxfit(column[column > 0])$lambda

#  return(new_column)

#}
```

```
#transform_insurance_train$INCOME_BC <- boxcox_trans(transform_insurance_train$INCOME)

#transform_insurance_train$BLUEBOOK_BC <- boxcox_trans(transform_insurance_train$BLUEBOOK)

#transform_insurance_train$OLDCLAIM_BC <- boxcox_trans(transform_insurance_train$OLDCLAIM)
```

Missingness Map for the Transformed Dataset

```
Amelia::missmap(transform_insurance_train, col = c("orange", "brown"))
```



BUILD MODELS

LINEAR REGRESSION

Model 1 Raw Data

```
lm_mod1 <- lm(TARGET_AMT ~., data = insurance_train[,-1])

summary(lm_mod1)
```

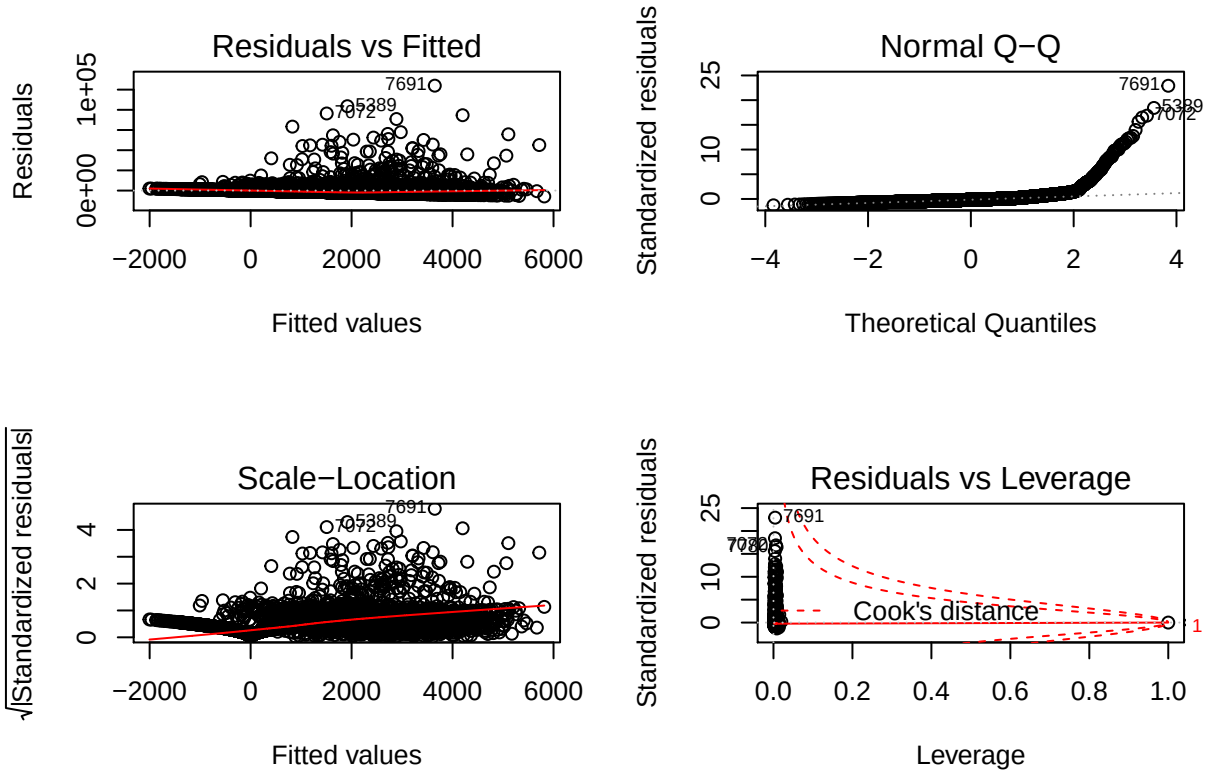
```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = insurance_train[, -1])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5816  -1691   -765    343 103938
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    9.890e+02  5.560e+02   1.779  0.07533 .
## KIDSDRIV                       3.165e+02  1.133e+02   2.794  0.00522 **
## AGE                            5.021e+00  7.075e+00   0.710  0.47791
## HOMEKIDS                       7.729e+01  6.545e+01   1.181  0.23765
```

```
## YOJ                                -4.434e+00  1.512e+01  -0.293  0.76930
## INCOME                             -4.480e-03  1.806e-03  -2.481  0.01311 *
## PARENT1Yes                          5.742e+02  2.023e+02   2.838  0.00456 **
## HOME_VAL                           -5.250e-04  5.909e-04  -0.889  0.37425
## MSTATUSz_No                         5.709e+02  1.449e+02   3.940 8.22e-05 ***
## SEXz_F                             -3.788e+02  1.840e+02  -2.059  0.03954 *
## EDUCATIONBachelors                 -3.958e+02  1.941e+02  -2.039  0.04145 *
## EDUCATIONMasters                   -2.435e+02  2.717e+02  -0.896  0.37025
## EDUCATIONPhD                        2.928e+01  3.342e+02   0.088  0.93019
## EDUCATIONz_High School             -1.173e+02  1.716e+02  -0.684  0.49427
## JOBClerical                         5.282e+02  3.417e+02   1.546  0.12214
## JOBDoctor                          -5.043e+02  4.090e+02  -1.233  0.21758
## JOBHome Maker                       3.478e+02  3.649e+02   0.953  0.34054
## JOBLawyer                           2.313e+02  2.958e+02   0.782  0.43428
## JOBManager                         -4.763e+02  2.886e+02  -1.650  0.09892 .
## JOBProfessional                     4.598e+02  3.090e+02   1.488  0.13673
## JOBStudent                          2.769e+02  3.743e+02   0.740  0.45949
## JOBz_Blue Collar                    5.069e+02  3.221e+02   1.574  0.11558
## TRAVTIME                            1.192e+01  3.225e+00   3.697  0.00022 ***
## CAR_USEPrivate                     -7.837e+02  1.646e+02  -4.760 1.97e-06 ***
## BLUEBOOK                            1.441e-02  8.630e-03   1.669  0.09506 .
## TIF                                -4.834e+01  1.219e+01  -3.965 7.42e-05 ***
## CAR_TYPEPanel Truck                 2.612e+02  2.784e+02   0.938  0.34822
## CAR_TYPEPickup                      3.763e+02  1.708e+02   2.203  0.02765 *
## CAR_TYPESports Car                  1.028e+03  2.179e+02   4.718 2.43e-06 ***
## CAR_TYPEVan                         5.154e+02  2.135e+02   2.414  0.01580 *
## CAR_TYPEz_SUV                       7.575e+02  1.794e+02   4.221 2.46e-05 ***
## RED_CARyes                         -5.546e+01  1.495e+02  -0.371  0.71062
## OLDCLAIM                           -1.047e-02  7.452e-03  -1.405  0.16011
## CLM_FREQ                            1.419e+02  5.513e+01   2.575  0.01005 *
## REVOKEDYes                          5.493e+02  1.737e+02   3.162  0.00157 **
## MVR_PTS                             1.743e+02  2.596e+01   6.714 2.02e-11 ***
## CAR_AGE                            -1.392e+02  4.558e+02  -0.305  0.76003
## URBANICITYz_Highly Rural/ Rural -1.662e+03  1.395e+02 -11.912  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4547 on 8117 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.07038,    Adjusted R-squared:  0.06615
## F-statistic: 16.61 on 37 and 8117 DF,  p-value: < 2.2e-16
```

A lot of the variables remain insignificant, however there is room for improvement. Let's run the same model with only significant varibales.

```
par(mfrow=c(2,2))

plot(lm_mod1)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Model 2

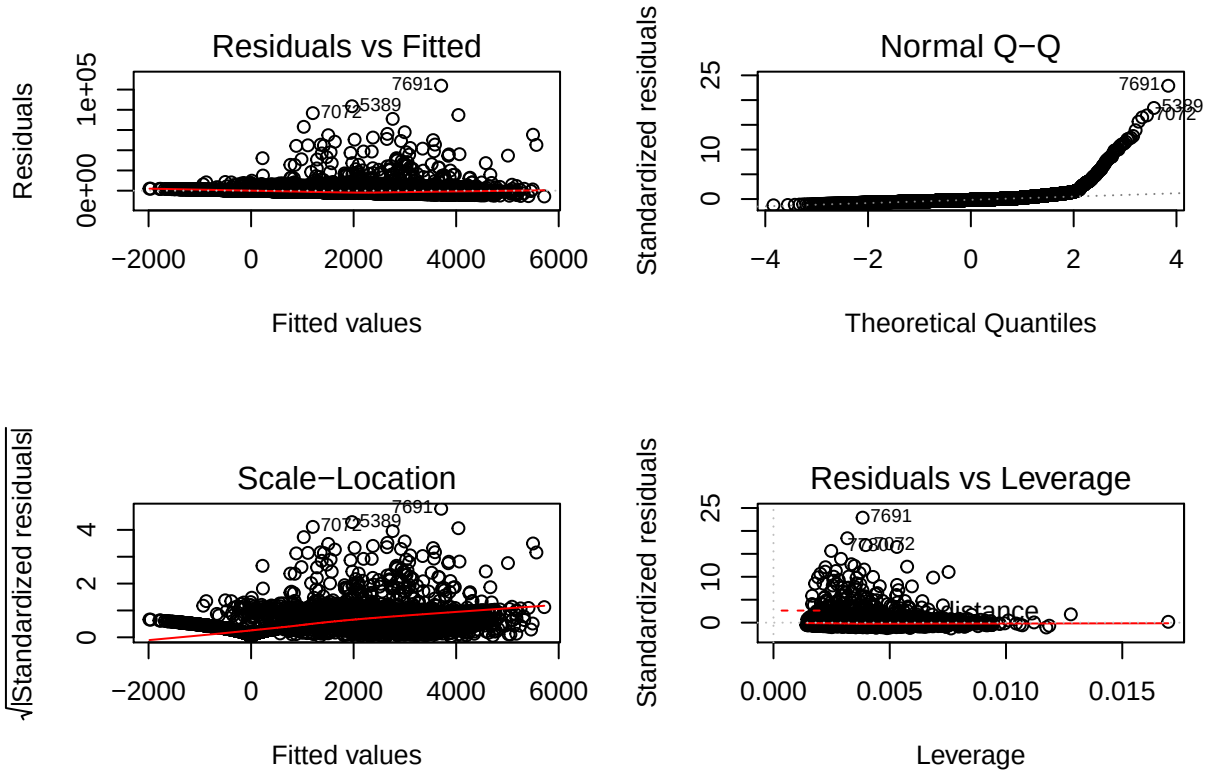
```
#lm_mod2 <- lm(TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1YES + MSTATUSz_No + SEXz_F + EDUCATIONBachelors +


amt_data <- insurance_train[,-1]

amt_data <- na.omit(amt_data) # missing values in character catergorical columns removed

lm_mod2 <- lm(TARGET_AMT ~., data = amt_data)

lm_mod2 <-  stepAIC(lm_mod2, trace = F)

#lm_inter2 <- lm(TARGET_AMT ~ 1, data = amt_data)

summary(lm_mod2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS +
##     SEX + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY, data = amt_data)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5719  -1686   -769    331 103878
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      9.264e+02  3.589e+02   2.582 0.009854 **
## KIDSDRIV                         3.778e+02  1.022e+02   3.697 0.000219 ***
## INCOME                          -5.276e-03  1.551e-03  -3.402 0.000673 ***
## PARENT1Yes                       6.471e+02  1.770e+02   3.656 0.000258 ***
## MSTATUSz_No                      5.952e+02  1.194e+02   4.985 6.32e-07 ***
## SEXz_F                          -3.383e+02  1.609e+02  -2.102 0.035564 *
## JOBClerical                      5.230e+02  2.849e+02   1.836 0.066466 .
## JOBDoctor                       -3.575e+02  3.770e+02  -0.948 0.342940
## JOBHome Maker                    3.039e+02  3.324e+02   0.914 0.360618
## JOBLawyer                        1.165e+02  2.872e+02   0.406 0.685070
## JOBManager                      -6.204e+02  2.665e+02  -2.329 0.019907 *
## JOBProfessional                  2.510e+02  2.645e+02   0.949 0.342619
## JOBStudent                       3.391e+02  3.185e+02   1.065 0.287024
## JOBz_Blue Collar                 4.786e+02  2.572e+02   1.860 0.062856 .
## TRAVTIME                         1.173e+01  3.222e+00   3.641 0.000273 ***
## CAR_USEPrivate                  -7.125e+02  1.568e+02  -4.545 5.58e-06 ***
## BLUEBOOK                         1.454e-02  8.528e-03   1.705 0.088181 .
## TIF                             -4.782e+01  1.218e+01  -3.925 8.75e-05 ***
## CAR_TYPEPanel Truck              3.175e+02  2.751e+02   1.154 0.248413
## CAR_TYPEPickup                   4.143e+02  1.694e+02   2.445 0.014486 *
## CAR_TYPESports Car               1.044e+03  2.164e+02   4.825 1.42e-06 ***
## CAR_TYPEVan                      5.437e+02  2.122e+02   2.562 0.010427 *
## CAR_TYPEz_SUV                    7.641e+02  1.785e+02   4.280 1.89e-05 ***
## OLDCLAIM                        -1.060e-02  7.439e-03  -1.425 0.154248
## CLM_FREQ                         1.443e+02  5.507e+01   2.620 0.008811 **
## REVOKEDYes                       5.579e+02  1.736e+02   3.215 0.001312 **
## MVR_PTS                          1.751e+02  2.592e+01   6.757 1.51e-11 ***
## URBANICITYz_Highly Rural/ Rural -1.654e+03  1.394e+02 -11.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4547 on 8127 degrees of freedom
## Multiple R-squared:  0.06927,    Adjusted R-squared:  0.06618
## F-statistic:  22.4 on 27 and 8127 DF,  p-value: < 2.2e-16
```

```
#summary(step(lm_inter2, direccion='both', scope = formula(lm_mod2), trace = F))
```

```
par(mfrow=c(2,2))

plot(lm_mod2)
```



LOGISITIC REGRESSION

Model 1 Raw Data

```
flg_data <- transform_insurance_train[,-c(2)]

log_mod1 <- glm(TARGET_FLAG ~., family = binomial, data = flg_data)

summary(log_mod1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = flg_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4659  -0.7334  -0.4176   0.6449   3.0844
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.872e+00  2.555e-01 -15.156  < 2e-16 ***
## AGE                 -6.746e-03  3.887e-03  -1.736 0.082635 .
## YOJ                 -2.536e-03  7.770e-03  -0.326 0.744094
## CAR_AGE             -1.034e+00  1.970e+01  -0.052 0.958143
## KIDSDRIV             3.776e-01  6.054e-02   6.238 4.44e-10 ***
```

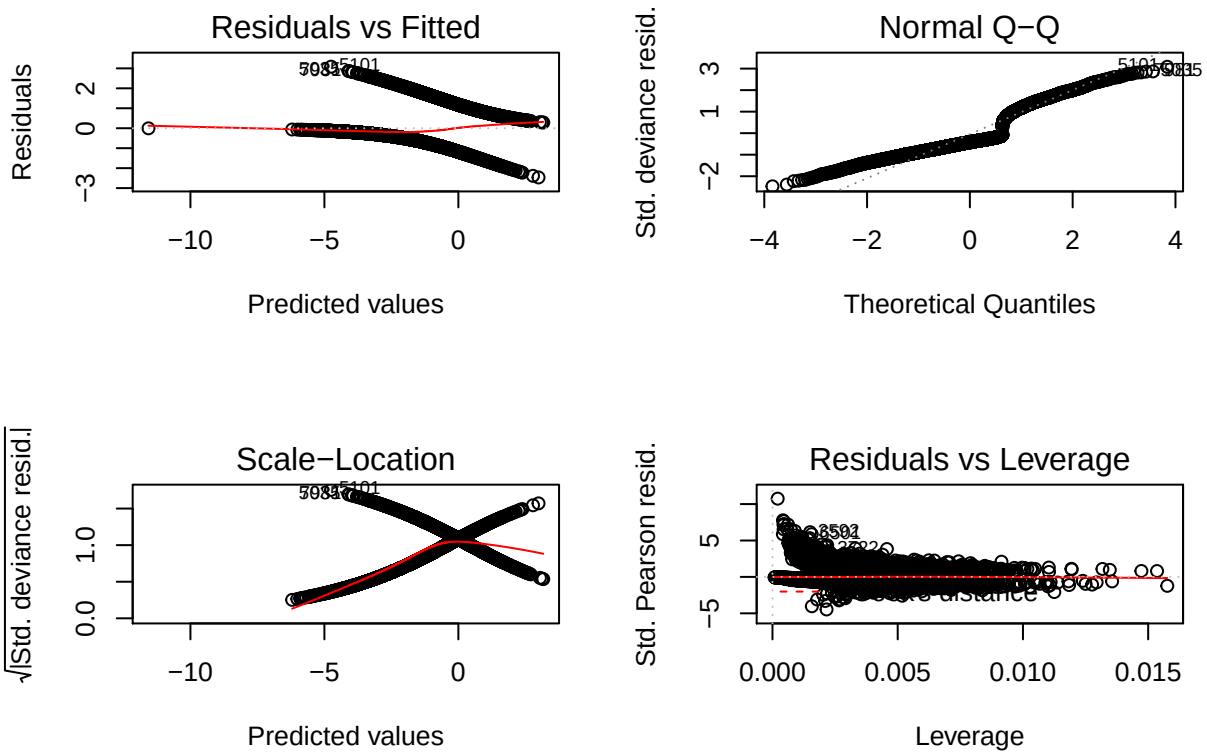
```
## HOMEKIDS             6.286e-02  3.657e-02   1.719 0.085624 .
## TRAVTIME             1.493e-02  1.859e-03   8.031 9.68e-16 ***
## TIF                 -5.452e-02  7.276e-03  -7.493 6.72e-14 ***
## CLM_FREQ             1.949e-01  2.820e-02   6.912 4.79e-12 ***
## MVR_PTS              1.153e-01  1.350e-02   8.542  < 2e-16 ***
## INCOME              -8.584e-06  8.044e-07 -10.672  < 2e-16 ***
## PARENT1_BIN          3.161e-01  1.084e-01   2.917 0.003538 **
## HOME_VAL_BIN         2.947e-01  7.616e-02   3.870 0.000109 ***
## MSTATUS_BIN          4.700e-01  8.296e-02   5.665 1.47e-08 ***
## IS_MALE_BIN          8.240e-02  1.100e-01   0.749 0.453666
## IS_COMMERCIAL_BIN    8.871e-01  6.936e-02  12.790  < 2e-16 ***
## BLUEBOOK            -2.271e-05  5.210e-06  -4.360 1.30e-05 ***
## OLDCLAIM            -1.365e-05  3.864e-06  -3.532 0.000413 ***
## CAR_PANEL_TRUCK_BIN  4.504e-01  1.506e-01   2.991 0.002778 **
## CAR_PICKUP_BIN       5.014e-01  9.721e-02   5.158 2.50e-07 ***
## CAR_SPORTS_CAR_BIN   9.773e-01  1.282e-01   7.623 2.49e-14 ***
## CAR_VAN_BIN          5.456e-01  1.228e-01   4.442 8.90e-06 ***
## CAR_SUV_BIN          7.392e-01  1.097e-01   6.736 1.63e-11 ***
## RED_CAR_BIN         -4.806e-02  8.568e-02  -0.561 0.574853
## REVOKED_BIN          8.837e-01  9.012e-02   9.806  < 2e-16 ***
## IS_URBAN_BIN         2.268e+00  1.125e-01  20.166  < 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: 7415.7  on 8129  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 7467.7
##
## Number of Fisher Scoring iterations: 10
```

```
par(mfrow=c(2,2))

plot(log_mod1)
```

```
## Warning: not plotting observations with leverage one:
##   3
```



```
## Warning: not plotting observations with leverage one:
##   3
```



The p-values here are really high. We can try to improve this classifcation model.

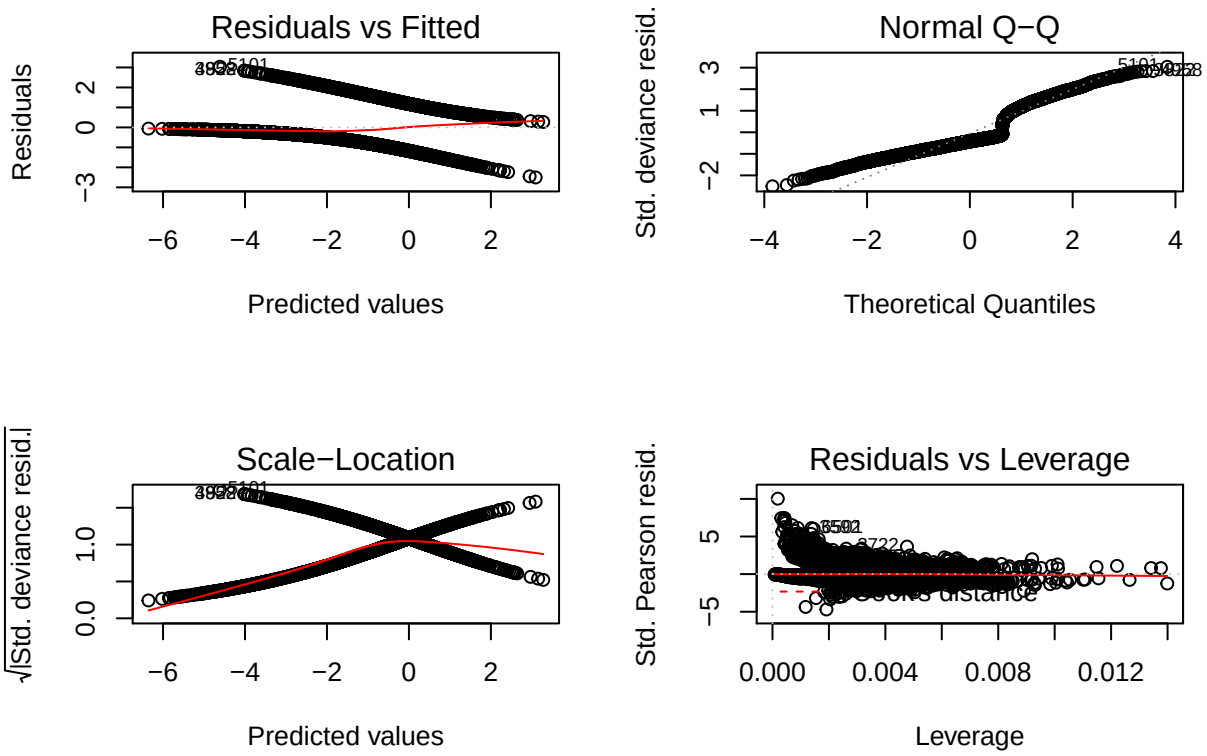
Model 2 Manually update model by only keeping the signifcant predictors in the first logistic model.

```
log_mod2 <- glm(TARGET_FLAG ~. -AGE-YOJ-CAR_AGE-HOMEKIDS-IS_MALE_BIN-RED_CAR_BIN , family = binomial, da

summary(log_mod2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - AGE - YOJ - CAR_AGE - HOMEKIDS -
##     IS_MALE_BIN - RED_CAR_BIN, family = binomial, data = flg_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5047  -0.7361  -0.4196   0.6377   3.0400
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.070e+00  1.694e-01 -24.025  < 2e-16 ***
## KIDSDRIV             4.203e-01  5.451e-02   7.709 1.27e-14 ***
## TRAVTIME             1.481e-02  1.856e-03   7.976 1.51e-15 ***
## TIF                 -5.404e-02  7.268e-03  -7.435 1.04e-13 ***
## CLM_FREQ             1.940e-01  2.816e-02   6.889 5.60e-12 ***
## MVR_PTS              1.168e-01  1.348e-02   8.666  < 2e-16 ***
```

```
## INCOME              -8.875e-06  7.739e-07 -11.468  < 2e-16 ***
## PARENT1_BIN          4.713e-01  9.307e-02   5.064 4.11e-07 ***
## HOME_VAL_BIN         3.126e-01  7.552e-02   4.140 3.47e-05 ***
## MSTATUS_BIN          4.139e-01  7.901e-02   5.238 1.62e-07 ***
## IS_COMMERCIAL_BIN    8.957e-01  6.927e-02  12.930  < 2e-16 ***
## BLUEBOOK            -2.527e-05  4.677e-06  -5.403 6.54e-08 ***
## OLDCLAIM            -1.376e-05  3.857e-06  -3.568 0.000360 ***
## CAR_PANEL_TRUCK_BIN  4.847e-01  1.399e-01   3.464 0.000533 ***
## CAR_PICKUP_BIN       4.938e-01  9.707e-02   5.087 3.63e-07 ***
## CAR_SPORTS_CAR_BIN   9.291e-01  1.053e-01   8.826  < 2e-16 ***
## CAR_VAN_BIN          5.633e-01  1.186e-01   4.749 2.05e-06 ***
## CAR_SUV_BIN          7.014e-01  8.435e-02   8.315  < 2e-16 ***
## REVOKED_BIN          8.921e-01  8.998e-02   9.915  < 2e-16 ***
## IS_URBAN_BIN         2.263e+00  1.124e-01  20.137  < 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  on 8154  degrees of freedom
## Residual deviance: 7426  on 8135  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 7466
##
## Number of Fisher Scoring iterations: 5
```

```r
par(mfrow=c(2,2))

plot(log_mod2)
```



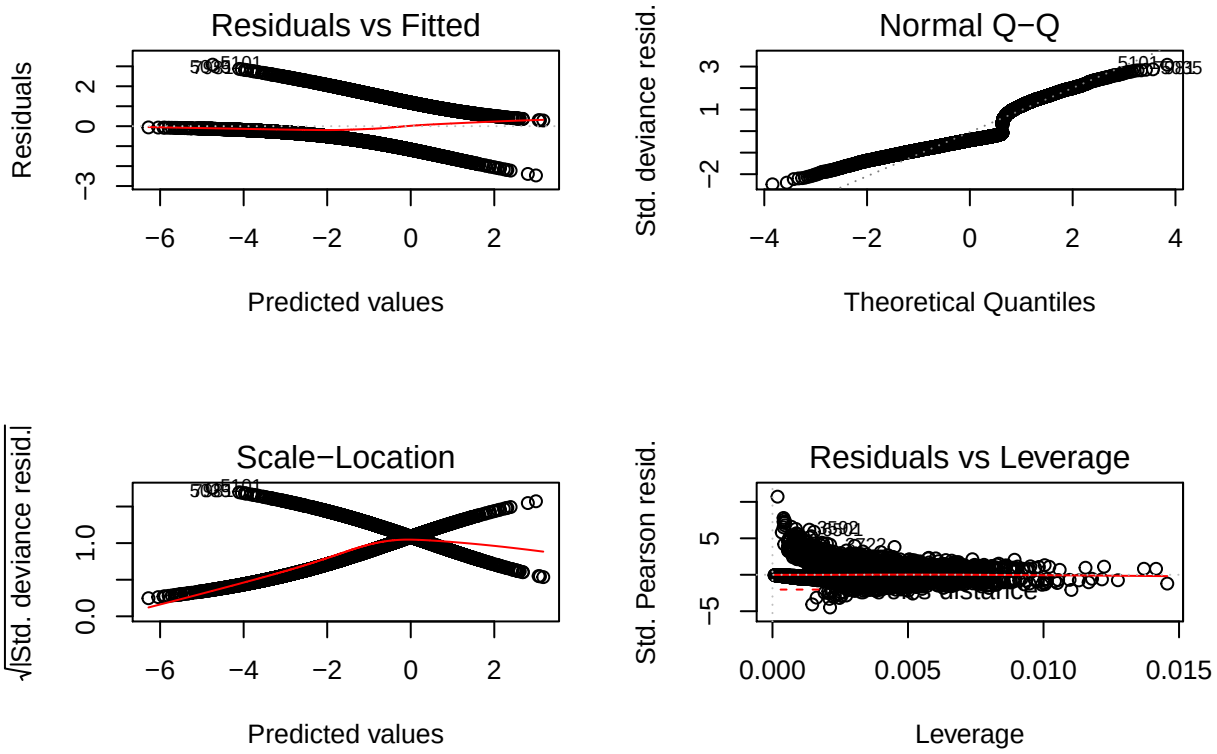
Model 3 Using the step regression algorthim to pick an optimal logistic model for the data.

```
log_mod3 <- stepAIC(log_mod1, trace = F)

summary(log_mod3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + KIDSDRIV + HOMEKIDS + TRAVTIME +
##     TIF + CLM_FREQ + MVR_PTS + INCOME + PARENT1_BIN + HOME_VAL_BIN +
##     MSTATUS_BIN + IS_COMMERCIAL_BIN + BLUEBOOK + OLDCLAIM + CAR_PANEL_TRUCK_BIN +
##     CAR_PICKUP_BIN + CAR_SPORTS_CAR_BIN + CAR_VAN_BIN + CAR_SUV_BIN +
##     REVOKED_BIN + IS_URBAN_BIN, family = binomial, data = flg_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4685  -0.7347  -0.4172   0.6450   3.0825
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.850e+00  2.411e-01 -15.969  < 2e-16 ***
## AGE                 -6.600e-03  3.822e-03  -1.727 0.084230 .
## KIDSDRIV             3.784e-01  6.045e-02   6.261 3.83e-10 ***
## HOMEKIDS             6.028e-02  3.599e-02   1.675 0.093978 .
## TRAVTIME             1.495e-02  1.858e-03   8.048 8.43e-16 ***
```

```
## TIF                -5.452e-02  7.274e-03  -7.495 6.62e-14 ***
## CLM_FREQ            1.952e-01  2.819e-02   6.926 4.34e-12 ***
## MVR_PTS             1.154e-01  1.350e-02   8.546  < 2e-16 ***
## INCOME             -8.633e-06  7.772e-07 -11.107  < 2e-16 ***
## PARENT1_BIN         3.171e-01  1.083e-01   2.926 0.003431 **
## HOME_VAL_BIN        2.967e-01  7.563e-02   3.923 8.73e-05 ***
## MSTATUS_BIN         4.709e-01  8.291e-02   5.680 1.34e-08 ***
## IS_COMMERCIAL_BIN   8.879e-01  6.934e-02  12.805  < 2e-16 ***
## BLUEBOOK           -2.390e-05  4.709e-06  -5.075 3.88e-07 ***
## OLDCLAIM           -1.380e-05  3.859e-06  -3.576 0.000349 ***
## CAR_PANEL_TRUCK_BIN  4.813e-01  1.401e-01   3.435 0.000592 ***
## CAR_PICKUP_BIN      5.003e-01  9.714e-02   5.151 2.60e-07 ***
## CAR_SPORTS_CAR_BIN  9.410e-01  1.056e-01   8.907  < 2e-16 ***
## CAR_VAN_BIN         5.640e-01  1.187e-01   4.750 2.04e-06 ***
## CAR_SUV_BIN         7.033e-01  8.450e-02   8.324  < 2e-16 ***
## REVOKED_BIN         8.861e-01  9.006e-02   9.839  < 2e-16 ***
## IS_URBAN_BIN        2.267e+00  1.123e-01  20.187  < 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: 7416.9  on 8133  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 7460.9
##
## Number of Fisher Scoring iterations: 5
```

```r
par(mfrow=c(2,2))

plot(log_mod3)
```



SELECT MODELS

Selected Linear Model Evaluation

```
anova(lm_mod1, lm_mod2, test = "Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY
## Model 2: TARGET_AMT ~ KIDSDRIV + INCOME + PARENT1 + MSTATUS + SEX + JOB +
##     TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + URBANICITY
##   Res.Df        RSS  Df  Sum of Sq Pr(>Chi)
## 1   8117 1.6781e+11
## 2   8127 1.6801e+11 -10 -200297050   0.4682
```

Linear **Model** 2 is clearly an improvement over Linear Model 1 and will be the optimal model to fit our data under linear regression.

Selected Logistic Model Metrics

ANOVA

```
anova(log_mod1, log_mod2, log_mod3)
```

```
## Analysis of Deviance Table
##
## Model 1: TARGET_FLAG ~ AGE + YOJ + CAR_AGE + KIDSDRIV + HOMEKIDS + TRAVTIME +
##     TIF + CLM_FREQ + MVR_PTS + INCOME + PARENT1_BIN + HOME_VAL_BIN +
##     MSTATUS_BIN + IS_MALE_BIN + IS_COMMERCIAL_BIN + BLUEBOOK +
##     OLDCLAIM + CAR_PANEL_TRUCK_BIN + CAR_PICKUP_BIN + CAR_SPORTS_CAR_BIN +
##     CAR_VAN_BIN + CAR_SUV_BIN + RED_CAR_BIN + REVOKED_BIN + IS_URBAN_BIN
## Model 2: TARGET_FLAG ~ (AGE + YOJ + CAR_AGE + KIDSDRIV + HOMEKIDS + TRAVTIME +
##     TIF + CLM_FREQ + MVR_PTS + INCOME + PARENT1_BIN + HOME_VAL_BIN +
##     MSTATUS_BIN + IS_MALE_BIN + IS_COMMERCIAL_BIN + BLUEBOOK +
##     OLDCLAIM + CAR_PANEL_TRUCK_BIN + CAR_PICKUP_BIN + CAR_SPORTS_CAR_BIN +
##     CAR_VAN_BIN + CAR_SUV_BIN + RED_CAR_BIN + REVOKED_BIN + IS_URBAN_BIN) -
##     AGE - YOJ - CAR_AGE - HOMEKIDS - IS_MALE_BIN - RED_CAR_BIN
## Model 3: TARGET_FLAG ~ AGE + KIDSDRIV + HOMEKIDS + TRAVTIME + TIF + CLM_FREQ +
##     MVR_PTS + INCOME + PARENT1_BIN + HOME_VAL_BIN + MSTATUS_BIN +
##     IS_COMMERCIAL_BIN + BLUEBOOK + OLDCLAIM + CAR_PANEL_TRUCK_BIN +
##     CAR_PICKUP_BIN + CAR_SPORTS_CAR_BIN + CAR_VAN_BIN + CAR_SUV_BIN +
##     REVOKED_BIN + IS_URBAN_BIN
##   Resid. Df Resid. Dev Df Deviance
## 1      8129     7415.7
## 2      8135     7426.0 -6 -10.3795
## 3      8133     7416.9  2   9.1508
```

Due to **Model 3** having the lowest AIC score, this will be the optimal model used for classification in our data.

```
probabilities <- predict(log_mod3, flg_data, type = "response")

predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

flg_data$pred.class <- predicted.classes

table("Predictions" = flg_data$pred.class, "Actual" = flg_data$TARGET_FLAG)
```

METRICS for selected classification model

```
##            Actual
## Predictions    0    1
##           0 5566 1295
##           1  441  853
```

ACCURACY Accuracy can be defined as the fraction of predicitons our model got right. Also known as the error rate, the accuracy rate makes no distinction about the type of error being made.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

```
cl_accuracy <- function(df){

  cm <- table("Predictions" = df$pred.class, "Actual" = df$TARGET_FLAG)



  TP <- cm[2,2]

  TN <- cm[1,1]

  FP <- cm[2,1]

  FN <- cm[1,2]



  return((TP + TN)/(TP + FP + TN + FN))

}
```

CLASSIFICATION ERROR RATE The Classification Error Rate calculates the number of incorrect predictions out of the total number of predictions in the dataset.

$$\text{Classification Error Rate} = \frac{FP + FN}{TP + FP + TN + FN}$$

```
cl_cer <- function(df){

  cm <- table("Predictions" = df$pred.class, "Actual" = df$TARGET_FLAG)



  TP <- cm[2,2]

  TN <- cm[1,1]

  FP <- cm[2,1]

  FN <- cm[1,2]



  return((FP + FN)/(TP + FP + TN + FN))

}
```

PRECISION This is the positive value or the fraction of the positive predictions that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

```
cl_precision <- function(df){

  cm <- table("Predictions" = df$pred.class, "Actual" = df$TARGET_FLAG)



  TP <- cm[2,2]

  TN <- cm[1,1]

  FP <- cm[2,1]

  FN <- cm[1,2]



  return(TP/(TP + FP))

}
```

SENSITIVITY The sensitivity is sometimes considered the true positive rate since it measures the accuracy in the event population.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

```
cl_sensitivity <- function(df){

  cm <- table("Predictions" = df$pred.class, "Actual" = df$TARGET_FLAG)



  TP <- cm[2,2]

  TN <- cm[1,1]

  FP <- cm[2,1]

  FN <- cm[1,2]



  return((TP)/(TP + FN))

}
```

SPECIFICITY This is the true negatitive rate or the proportion of negatives that are correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

```
cl_specificity<- function(df){

  cm <- table("Predictions" = df$pred.class, "Actual" = df$TARGET_FLAG)



  TP <- cm[2,2]

  TN <- cm[1,1]

  FP <- cm[2,1]

  FN <- cm[1,2]



  return((TN)/(TN + FP))

}
```

F1 SCORE OF PREDICTIONS The F1 Score of Predictions measures the test's accuracy, on a scale of 0 to 1 where a value of 1 is the most accurate and the value of 0 is the least accurate.

$$\text{F1 Score} = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

```
cl_f1score <- function(df){

  cm <- table("Predictions" = df$pred.class, "Actual" = df$TARGET_FLAG)



  TP <- cm[2,2]

  TN <- cm[1,1]

  FP <- cm[2,1]

  FN <- cm[1,2]



  f1score <- (2 * cl_precision(df) * cl_sensitivity(df)) / (cl_precision(df) + cl_sensitivity(df))

  return(f1score)

}
```

```
f1_score_function <- function(cl_precision, cl_sensitivity){
  f1_score <- (2*cl_precision*cl_sensitivity)/(cl_precision+cl_sensitivity)
  return (f1_score)
}

(f1_score_function(0, .5))
```

F1 SCORE BOUNDS

```
## [1] 0
```

```
(f1_score_function(1, 1))
```

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

```
p <- runif(100, min = 0, max = 1)
s <- runif(100, min = 0, max = 1)
f <- (2*p*s)/(p+s)
summary(f)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.007589 0.253108 0.387737 0.438079 0.669418 0.992182
```

```
Metric <- c('Accuracy','Classification Error Rate', 'Precision', 'Sensitivity','Specificity', 'F1 Score
Score <- round(c(cl_accuracy(flg_data), cl_cer(flg_data), cl_precision (flg_data), cl_sensitivity(flg_da
df_1 <- as.data.frame(cbind(Metric, Score))
kable(df_1)
```

Results from Selected Classification model

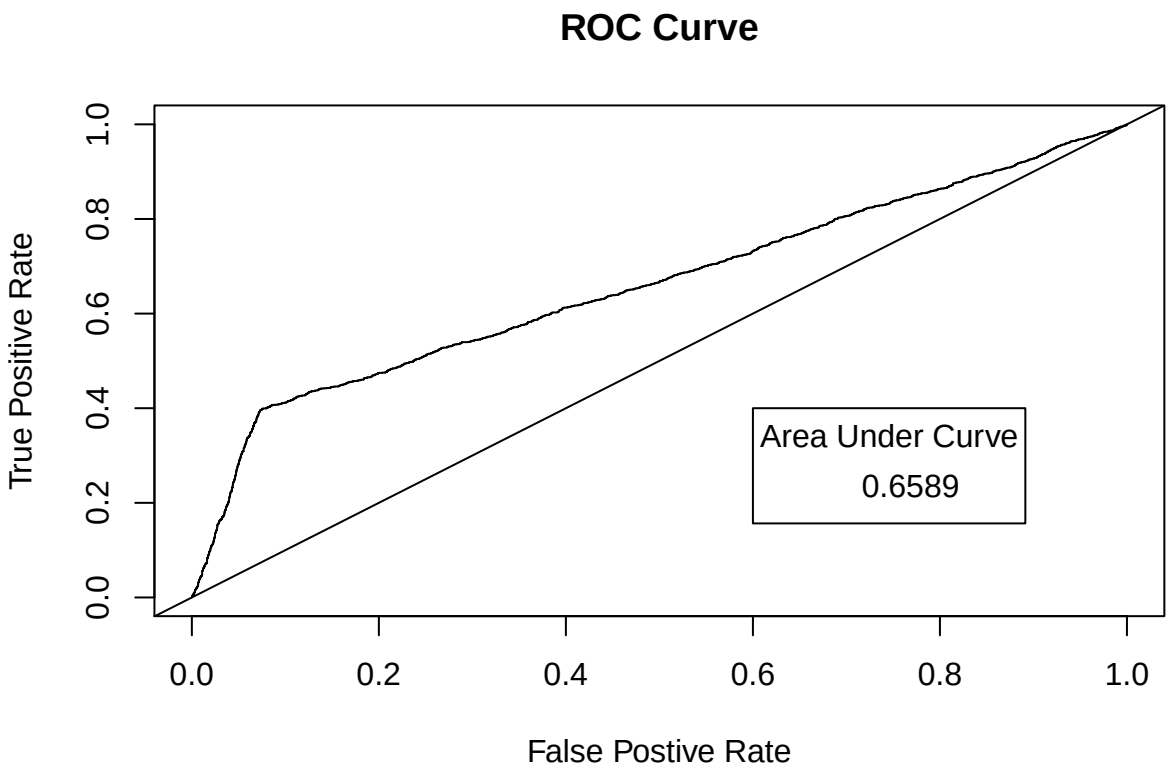
Metric	Score
Accuracy	0.7871
Classification Error Rate	0.2129
Precision	0.6592
Sensitivity	0.3971
Specificity	0.9266
F1 Score	0.4956

ROC CURVE Shows how the true positive rate against the false positive rate at various threshold settings. The AUC (Area Under Curve) tells how much model is capable of distinguishing between classes. Higher the AUC is better, that is, how well the model is at predicting 0s as 0s and 1s as 1s.

Creating an ROC Function

```
ROC <- function(x, y){

  x <- x[order(y, decreasing = TRUE)]

 t_p_r <- cumsum(x) / sum(x)

 f_p_r <- cumsum(!x) / sum(!x)

  xy <- data.frame(t_p_r,f_p_r, x)



 f_p_r_df <- c(diff(xy$f_p_r), 0)

 t_p_r_df <- c(diff(xy$t_p_r), 0)

  A_U_C <- round(sum(xy$t_p_r *f_p_r_df) + sum(t_p_r_df *f_p_r_df)/2, 4)



  plot(xy$f_p_r, xy$t_p_r, type = "l",

       main = "ROC Curve",

       xlab = "False Postive Rate",

       ylab = "True Positive Rate")

  abline(a = 0, b = 1)

  legend(.6, .4, A_U_C, title = "Area Under Curve")

}
```

```
ROC1 <- ROC(flg_data$TARGET_FLAG, flg_data$pred.class)
```




```
ROC1
```

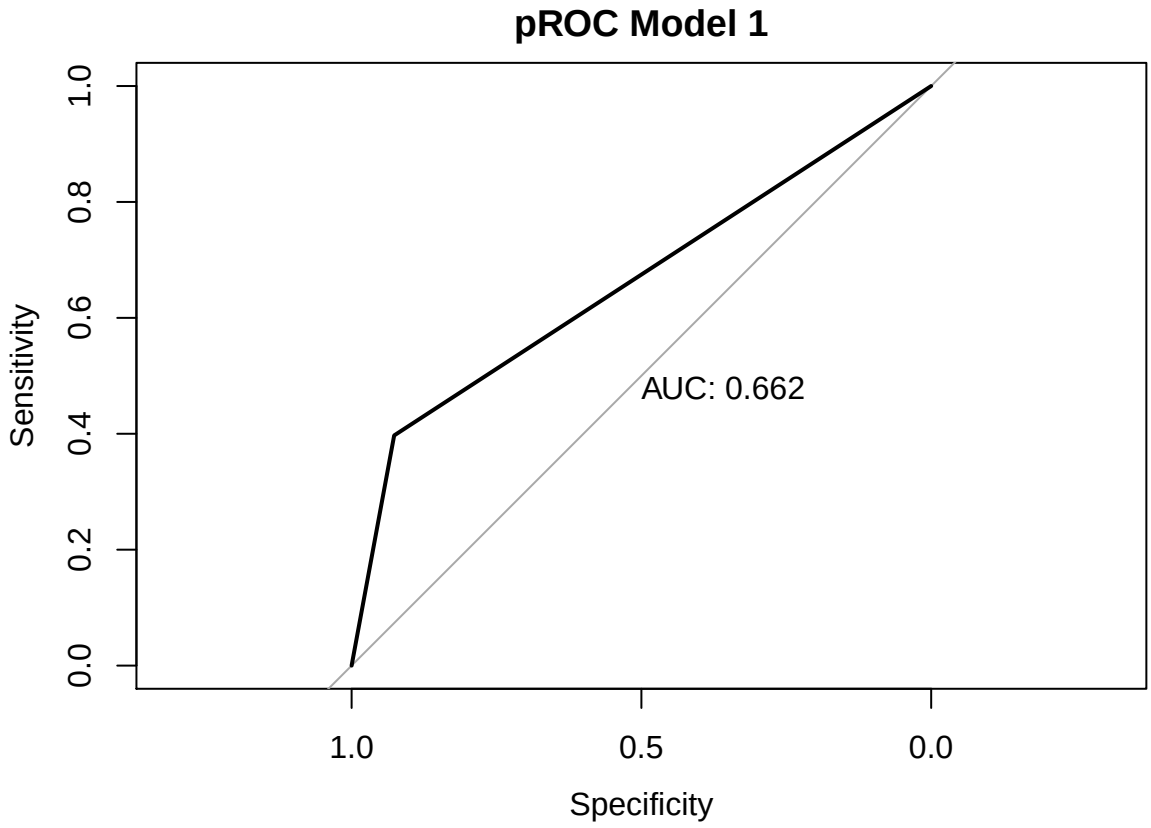
```
## $rect
## $rect$w
## [1] 0.2913194
##
## $rect$h
## [1] 0.2434959
##
## $rect$left
## [1] 0.6
##
## $rect$top
## [1] 0.4
##
##
## $text
## $text$x
## [1] 0.7164354
##
## $text$y
## [1] 0.2376694
```

```
roc.mod1 <- roc(flg_data$TARGET_FLAG, flg_data$pred.class)
```

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

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

```
plot(roc.mod1, print.auc = TRUE , main = "pROC Model 1")
```



Based on the AUC the classification model performed at a satisfactory level with a score of `0.662`.

Predictions

```
insurance_eval$INCOME  <- gsub( "\\$", "", insurance_eval$INCOME)

insurance_eval$INCOME  <- gsub( "\\,", "", insurance_eval$INCOME)

insurance_eval$INCOME  <- as.numeric(insurance_eval$INCOME)


insurance_eval$HOME_VAL  <- gsub( "\\$", "", insurance_eval$HOME_VAL)

insurance_eval$HOME_VAL  <- gsub( "\\,", "", insurance_eval$HOME_VAL)

insurance_eval$HOME_VAL  <- as.numeric(insurance_eval$HOME_VAL)
```

```
insurance_eval$BLUEBOOK  <- gsub( "\\$", "", insurance_eval$BLUEBOOK)

insurance_eval$BLUEBOOK  <- gsub( "\\,", "", insurance_eval$BLUEBOOK)

insurance_eval$BLUEBOOK  <- as.numeric(insurance_eval$BLUEBOOK)


insurance_eval$OLDCLAIM  <- gsub( "\\$", "", insurance_eval$OLDCLAIM)

insurance_eval$OLDCLAIM  <- gsub( "\\,", "", insurance_eval$OLDCLAIM)

insurance_eval$OLDCLAIM  <- as.numeric(insurance_eval$OLDCLAIM)


eval_amt <- insurance_eval[,-c(1,2)]
```

```
eval_amt <- predict(lm_mod2, newdata = eval_amt, interval="prediction")

insurance_eval$TARGET_AMT <- eval_amt[,1]
```

Linear Model

```
prob <- predict(log_mod3, transform_insurance_eval[,-1], type='response')

transform_insurance_eval$TARGET_FLAG <- ifelse(prob >= 0.50, 1, 0)
```

Logisitic Model

Final Test Data Result Full Test Set Here

```
insurance_eval$TARGET_FLAG <- transform_insurance_eval$TARGET_FLAG

insurance_eval %>% head(10) %>% as.tibble()
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
## # A tibble: 10 x 26
##    INDEX TARGET_FLAG TARGET_AMT KIDSDRIV   AGE HOMEKIDS   YOJ INCOME PARENT1
##    <int>       <dbl>      <dbl>    <int> <int>    <int> <int>  <dbl> <fct>
##  1     3           0      1380.        0    48        0    11  52881 No
##  2     9           0      1640.        1    40        1    11  50815 Yes
##  3    10           0      1149.        0    44        2    12  43486 Yes
##  4    18           0      1901.        0    35        2    NA  21204 Yes
```

```
##  5    21           0       945.        0    59        0    12  87460 No
##  6    30           0        NA         0    46        0    14     NA No
##  7    31           0      2215.        0    60        0    12  37940 No
##  8    37           0      2898.        0    54        0    12  33212 No
##  9    39           0       985.        2    36        2    12 130540 Yes
## 10    47           0      1363.        0    50        0     8 167469 No
## # ... with 17 more variables: HOME_VAL <dbl>, MSTATUS <fct>, SEX <fct>,
## #   EDUCATION <fct>, JOB <fct>, TRAVTIME <int>, CAR_USE <fct>, BLUEBOOK <dbl>,
## #   TIF <int>, CAR_TYPE <fct>, RED_CAR <fct>, OLDCLAIM <dbl>, CLM_FREQ <int>,
## #   REVOKED <fct>, MVR_PTS <int>, CAR_AGE <int>, URBANICITY <fct>
```

```
write.csv(insurance_eval, "insurance_predictions.csv", row.names = F)
```

Source code found on GITHUB