DATA 621 - Homework 4

Fall 2020 - Business Analytics and Data Mining

Mohamed Thasleem, Kalikul Zaman

11/22//2020

Contents

Introduction

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 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.

The 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:

```r
# load libraries
library(ggpubr)
library(stringr)
library(corrplot)
library(RColorBrewer)
library(mice)
library(kableExtra)
library(car)
library(MASS)
library(caret)
library(pROC)
library(ggplot2)
library(reshape2)
```

```r
library(knitr)
library(tidyverse)
library(psych)
library(ggthemes)
```

1. Data Download

```r
# download data
path <- "https://raw.githubusercontent.com/mohamedthasleem/DATA621/master/HW4"
insurance_train <- read.csv(paste0(path,"/insurance_training_data.csv"))
insurance_test <- read.csv(paste0(path,"/insurance-evaluation-data.csv"))
```

2. Data Exploration

Previewing the data, We will first look at the summary statistics for the data

```r
head(insurance_train)
```

```
##   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
```

```r
glimpse(insurance_train)
```

```
## Rows: 8,161
## Columns: 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      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,...
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", ...
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "...
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Ye...
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", ...
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School"...
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Co...
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, ...
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private...
## $ BLUEBOOK    <chr> "$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    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Spo...
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no...
## $ OLDCLAIM    <chr> "$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     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No",...
## $ 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  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly U...
```

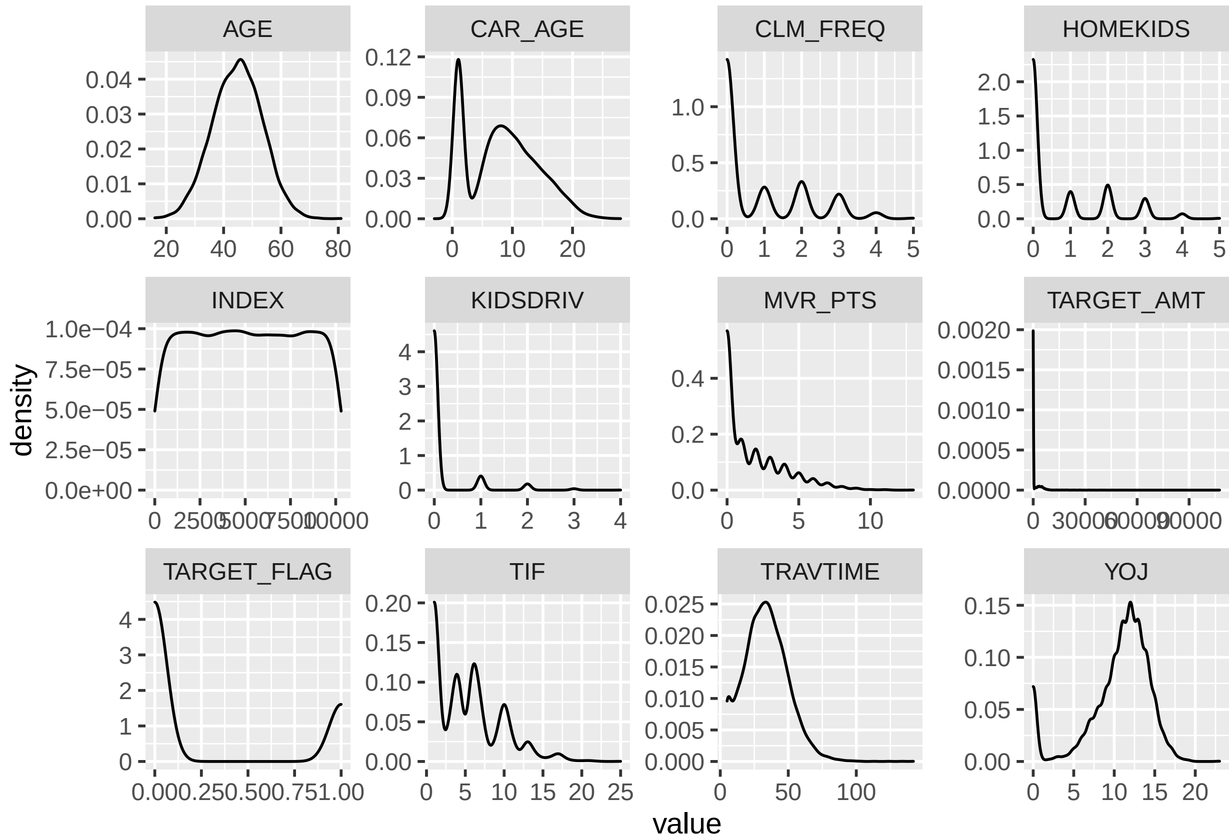
```r
summary(insurance_train)
```

```
##      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
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL            MSTATUS               SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##   EDUCATION             JOB              TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK              TIF          CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM            CLM_FREQ       REVOKED              MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

Density are useful to show how the data is distributed in the dataset. In the histogram plot below, we see several variables have high number of zeros. AGE is the only variable that is normally distributed. Rest of the variables show some skewness. We will perform Box-Cox transformation on these variables.

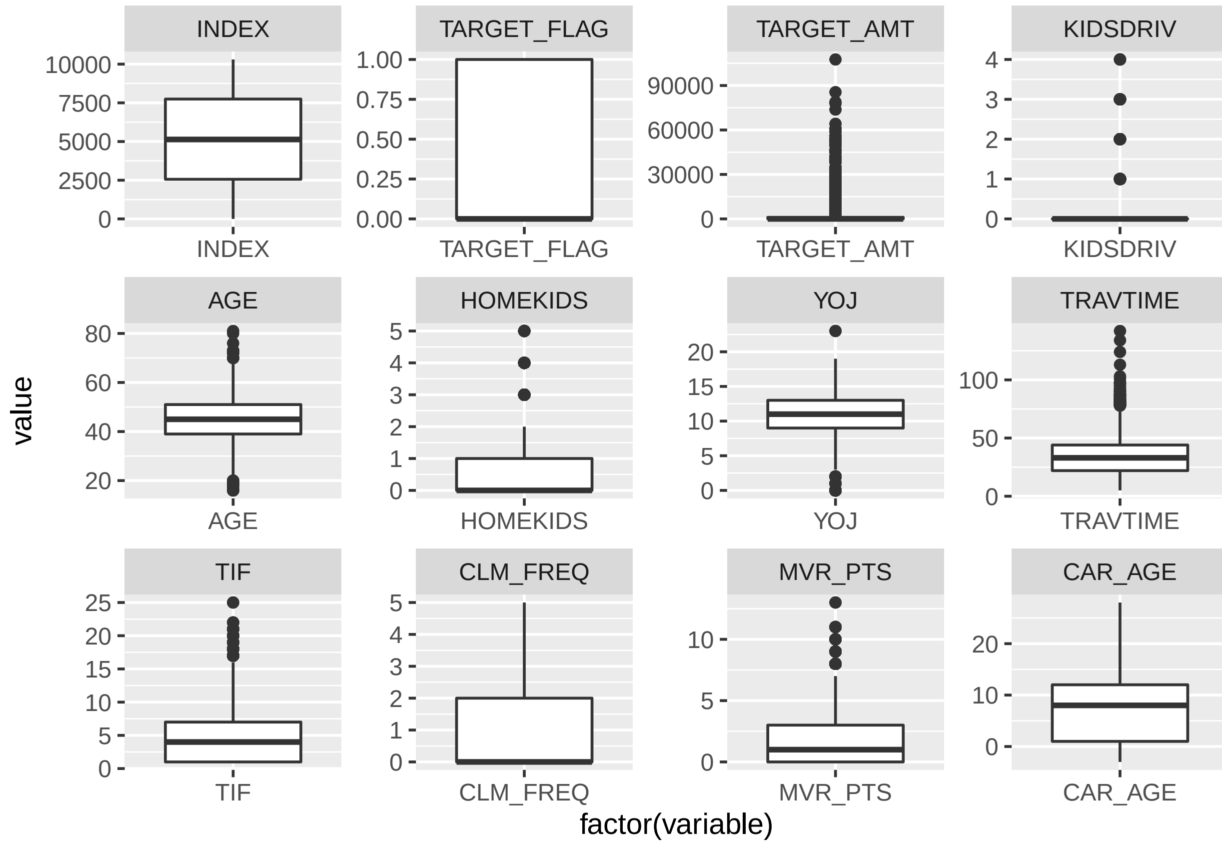
```
ntrain<-select_if(insurance_train, is.numeric)
ntrain %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) + facet_wrap(~ key, scales = "free") + geom_density()
```




```r
ggplot(melt(insurance_train), aes(x=factor(variable), y=value)) +
  facet_wrap(~variable, scale="free") +
  geom_boxplot()
```

```
## Using INCOME, PARENT1, HOME_VAL, MSTATUS, SEX, EDUCATION, JOB, CAR_USE, BLUEBOOK, CAR_TYPE, RED_CAR,
```

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



The numerical summaries and visualizations associated with the dataset. As with any data, some details to this dataset including the numerous amounts of missing data, as well as skew in the histograms. We will work on the missing value on upcoming sections

3. Data Preparation

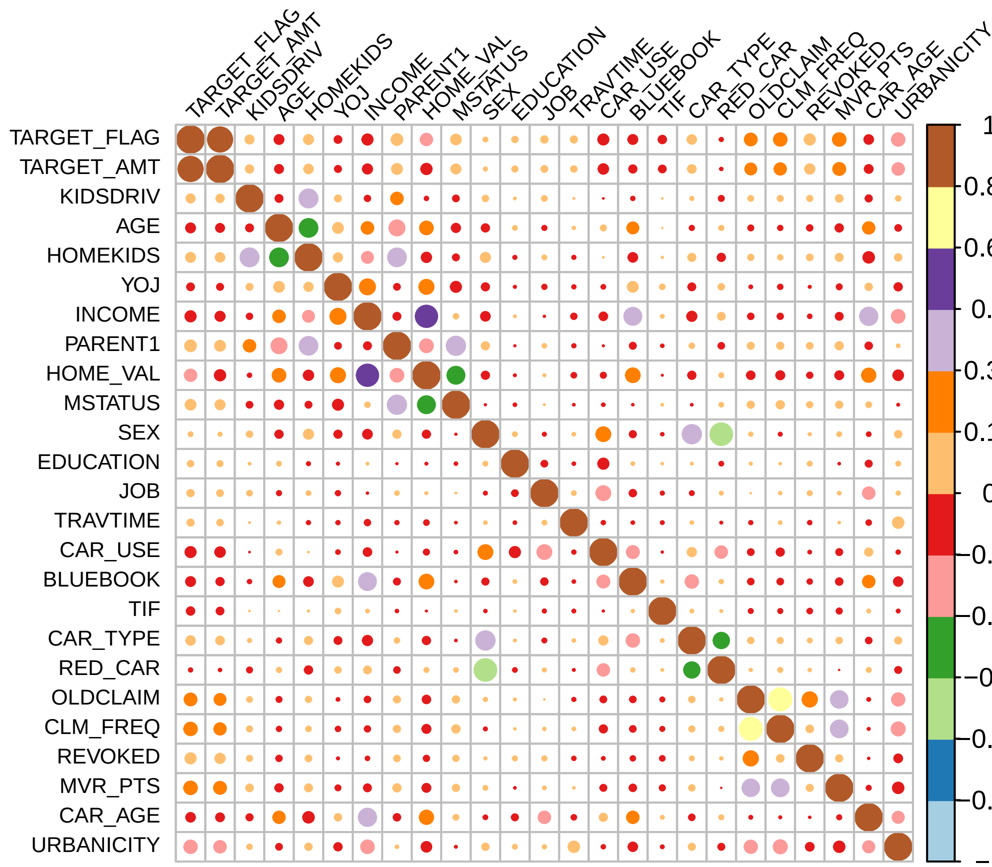
Impute data for Missing value, changing some datatype for data analysis and build correlation plot, VIF values are calculated

```
# change data type
insurance_train_dist <- insurance_train %>%
  dplyr::select(-INDEX) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG),
         KIDSDRIV = as.factor(KIDSDRIV),
         HOMEKIDS = as.factor(HOMEKIDS),
         PARENT1 = as.factor(PARENT1),
         CLM_FREQ = as.factor(CLM_FREQ),
         INCOME = str_replace_all(INCOME, "[\\$,]", ""),
         HOME_VAL = str_replace_all(HOME_VAL, "[\\$,]", ""),
         BLUEBOOK = str_replace_all(BLUEBOOK, "[\\$,]", ""),
         OLDCLAIM = str_replace_all(OLDCLAIM, "[\\$,]", ""),
         OLDCLAIM = as.integer(OLDCLAIM),
         BLUEBOOK = as.integer(BLUEBOOK),
         HOME_VAL = as.integer(HOME_VAL),
         INCOME = as.integer(INCOME))
```

```r
# change data type of some variables for visualization
distribution <- insurance_train_dist %>%
  dplyr::select(c("TARGET_FLAG", "AGE", "YOJ", "INCOME", "HOME_VAL", "TRAVTIME", "BLUEBOOK", "TIF", "OLI
  gather(key, value, -TARGET_FLAG) %>%
  mutate(value = as.integer(value),
         key = as.factor(key),
         TARGET_FLAG = as.factor(TARGET_FLAG))

# change all variable's data type for correlation
insurance_corr <- data.frame(lapply(insurance_train_dist, function(x) as.numeric(as.factor(x))))

# top correlated variables
a <- sort(cor(dplyr::select(insurance_corr, TARGET_FLAG, everything()))[,1], decreasing = T)
b <- sort(cor(dplyr::select(insurance_corr, TARGET_AMT, everything()))[,1], decreasing = T)
kable(cbind(a, b), col.names = c("TARGET_FLAG", "TARGET_AMT")) %>%
  kable_styling(full_width = F) %>%
  add_header_above(c(" ", "Correlation" = 2))
```

	Correlation	
	TARGET_FLAG	TARGET_AMT
TARGET_FLAG	1.0000000	1.0000000
TARGET_AMT	0.8334240	0.8334240
MVR_PTS	0.2191323	0.1970216
CLM_FREQ	0.2161961	0.1741927
OLDCLAIM	0.1947302	0.1611626
PARENT1	0.1576222	0.1359305
REVOKED	0.1519391	0.1263285
MSTATUS	0.1351248	0.1214701
HOMEKIDS	0.1156210	0.1008356
KIDSDRIV	0.1036683	0.0877148
CAR_TYPE	0.1023650	0.0797487
JOB	0.0612262	0.0488313
TRAVTIME	0.0492559	0.0401971
EDUCATION	0.0428730	0.0397864
SEX	0.0210786	0.0088270
RED_CAR	-0.0069473	0.0005877
TIF	-0.0823431	-0.0683183
BLUEBOOK	-0.1092768	-0.0709830
CAR_USE	-0.1426737	-0.1287263
URBANICITY	-0.2242509	-0.1904945

```r
# correlation plot
corrplot(cor(dplyr::select(drop_na(insurance_corr), everything())),
         method = "circle",
         type = "full",
         col = brewer.pal(n = 26, name = "Paired"),
         number.cex = .7, tl.cex = .7,
         tl.col = "black", tl.srt = 45)
```



The correlation table and plot above, we see MVR_PTS, CLM_FREQ, and OLDCLAIM are the most positively correlated variables with our response variables. Whereas, URBANICITY is the most negatively correlated variable. All other are weakly correlated.

```
# check for multicollinearity
insurance_vif <- data.frame(lapply(insurance_imputed, function(x) as.numeric(as.factor(x))))
kable((car::vif(glm(TARGET_FLAG ~. , data = insurance_vif))), col.names = c("VIF Score")) %>%  #remove
  kable_styling(full_width = F)
```

	VIF Score
TARGET_AMT	1.183240
KIDSDRIV	1.322490
AGE	1.409393
HOMEKIDS	2.066177
YOJ	1.216771
INCOME	2.521329
PARENT1	1.845675
HOME_VAL	2.221036
MSTATUS	1.887328
SEX	2.264001
EDUCATION	1.042874
JOB	1.153833
TRAVTIME	1.038871
CAR_USE	1.353390
BLUEBOOK	1.375659
TIF	1.009085
CAR_TYPE	1.409589
RED_CAR	1.808620
OLDCLAIM	2.201159
CLM_FREQ	2.131538
REVOKED	1.148620
MVR_PTS	1.249568
CAR_AGE	1.302538
URBANICITY	1.240198

The multicollinearity check, VIF score is at a conservative level for all variables

4. Build Models

We will be building three different multiple linear regression models and three different binary logistic regression models using the original dataset, the imputed dataset, forward and backward selected variables and a boxcox transformed dataset to see which one yields the best performance.

Model 1 : Multiple Linear Regression

The p-value below shows that the probability of this variables to be irrelevant is very low. R-squared is 0.15, which means this model explains 15% of the data's variation. This is not an good model

```
# original value model
insurance_corr <- dplyr::select(insurance_corr, -"TARGET_FLAG")
model1 <- lm(TARGET_AMT ~ ., insurance_corr)
summary(model1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = insurance_corr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -898.90 -286.25 -134.43   62.85 1927.07
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.018e+02  7.820e+01   5.138 2.86e-07 ***
## KIDSDRIV     4.205e+01  1.318e+01   3.189  0.00143 **
## AGE         -2.046e-01  8.070e-01  -0.254  0.79986
## HOMEKIDS     1.359e+01  7.532e+00   1.804  0.07121 .
## YOJ         -3.491e-01  1.590e+00  -0.220  0.82625
## INCOME      -2.270e-02  4.803e-03  -4.727 2.33e-06 ***
## PARENT1      7.414e+01  2.336e+01   3.173  0.00151 **
## HOME_VAL    -1.082e-02  5.505e-03  -1.965  0.04946 *
## MSTATUS      7.301e+01  1.685e+01   4.332 1.50e-05 ***
## SEX         -9.673e+00  1.756e+01  -0.551  0.58178
## EDUCATION    6.679e+00  4.132e+00   1.616  0.10606
## JOB         -6.667e-01  2.347e+00  -0.284  0.77637
## TRAVTIME     2.185e+00  3.772e-01   5.793 7.25e-09 ***
## CAR_USE     -1.472e+02  1.392e+01 -10.571  < 2e-16 ***
## BLUEBOOK    -2.513e-02  9.504e-03  -2.644  0.00821 **
## TIF         -7.286e+00  1.416e+00  -5.147 2.73e-07 ***
## CAR_TYPE     1.877e+01  3.517e+00   5.335 9.87e-08 ***
## RED_CAR     -1.768e+01  1.732e+01  -1.021  0.30740
## OLDCLAIM    -4.355e-03  1.039e-02  -0.419  0.67518
## CLM_FREQ     2.315e+01  7.358e+00   3.147  0.00166 **
## REVOKED      1.281e+02  1.915e+01   6.693 2.37e-11 ***
## MVR_PTS      2.597e+01  3.034e+00   8.560  < 2e-16 ***
## CAR_AGE     -3.795e+00  1.182e+00  -3.210  0.00133 **
## URBANICITY  -2.685e+02  1.590e+01 -16.891  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.1 on 6424 degrees of freedom
##   (1713 observations deleted due to missingness)
## Multiple R-squared:  0.1564, Adjusted R-squared:  0.1534
## F-statistic: 51.79 on 23 and 6424 DF,  p-value: < 2.2e-16
```

Model 2 : Multiple Linear Regression (VIF)

Considering the data from VIF, The p-value below shows that the probability of this variables to be irrelevant is very low. R-squared is 0.15, which means this model explains 15% of the data's variation. This is not an good model.

```
# imputed model
insurance_vif <- dplyr::select(insurance_vif, -"TARGET_FLAG")
model2 <- lm(TARGET_AMT ~ ., insurance_vif)
summary(model2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = insurance_vif)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -911.66 -287.41 -134.50   64.31 1927.39
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.447e+02  6.991e+01   4.930 8.38e-07 ***
## KIDSDRIV     5.522e+01  1.173e+01   4.709 2.53e-06 ***
## AGE          7.359e-02  7.196e-01   0.102  0.91855
## HOMEKIDS     1.421e+01  6.724e+00   2.113  0.03461 *
## YOJ         -1.133e+00  1.410e+00  -0.803  0.42179
## INCOME      -2.442e-02  4.074e-03  -5.994 2.13e-09 ***
## PARENT1      6.731e+01  2.095e+01   3.212  0.00132 **
## HOME_VAL    -8.899e-03  4.577e-03  -1.944  0.05189 .
## MSTATUS      8.610e+01  1.461e+01   5.891 3.99e-09 ***
## SEX         -3.071e+00  1.576e+01  -0.195  0.84551
## EDUCATION    7.917e+00  3.691e+00   2.145  0.03199 *
## JOB          4.399e-02  2.092e+00   0.021  0.98323
## TRAVTIME     2.081e+00  3.358e-01   6.198 5.98e-10 ***
## CAR_USE     -1.435e+02  1.248e+01 -11.503  < 2e-16 ***
## BLUEBOOK    -2.503e-02  8.501e-03  -2.944  0.00325 **
## TIF         -7.582e+00  1.263e+00  -6.001 2.04e-09 ***
## CAR_TYPE     1.677e+01  3.150e+00   5.323 1.05e-07 ***
## RED_CAR     -3.530e+00  1.546e+01  -0.228  0.81936
## OLDCLAIM    -7.042e-03  9.190e-03  -0.766  0.44349
## CLM_FREQ     2.142e+01  6.579e+00   3.256  0.00113 **
## REVOKED      1.305e+02  1.701e+01   7.674 1.86e-14 ***
## MVR_PTS      2.595e+01  2.706e+00   9.591  < 2e-16 ***
## CAR_AGE     -3.124e+00  1.041e+00  -3.001  0.00270 **
## URBANICITY  -2.710e+02  1.411e+01 -19.209  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.9 on 8137 degrees of freedom
## Multiple R-squared:  0.1549, Adjusted R-squared:  0.1525
## F-statistic: 64.83 on 23 and 8137 DF,  p-value: < 2.2e-16
```

Model 3 : Multiple Linear Regression (Stepwise Transformed)

We see improved p-value for several variables, The p-value below shows that the probability of this variables to be irrelevant is very low. Lastly, R-squared is 0.15, which means this model explains 15% of the data's variation, seems to be good model

```r
# stepwise transformed model
model3 <- stepAIC(model2, direction = "both", trace = FALSE)
summary(model3)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY, data = insurance_vif)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -917.76 -288.17 -135.02   63.27 1931.63
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.333e+02  4.960e+01   6.719 1.95e-11 ***
## KIDSDRIV     5.573e+01  1.156e+01   4.822 1.45e-06 ***
## HOMEKIDS     1.327e+01  6.166e+00   2.153 0.031374 *
## INCOME      -2.519e-02  3.925e-03  -6.418 1.46e-10 ***
## PARENT1      6.750e+01  2.084e+01   3.239 0.001204 **
## HOME_VAL    -8.860e-03  4.557e-03  -1.944 0.051898 .
## MSTATUS      8.719e+01  1.454e+01   5.995 2.12e-09 ***
## EDUCATION    7.975e+00  3.656e+00   2.182 0.029164 *
## TRAVTIME     2.084e+00  3.355e-01   6.212 5.47e-10 ***
## CAR_USE     -1.440e+02  1.136e+01 -12.673  < 2e-16 ***
## BLUEBOOK    -2.519e-02  8.274e-03  -3.045 0.002338 **
## TIF         -7.593e+00  1.262e+00  -6.015 1.88e-09 ***
## CAR_TYPE     1.673e+01  2.741e+00   6.104 1.08e-09 ***
## CLM_FREQ     1.820e+01  5.060e+00   3.597 0.000324 ***
## REVOKED      1.263e+02  1.607e+01   7.858 4.41e-15 ***
## MVR_PTS      2.568e+01  2.675e+00   9.601  < 2e-16 ***
## CAR_AGE     -3.054e+00  1.017e+00  -3.002 0.002687 **
## URBANICITY  -2.703e+02  1.408e+01 -19.199  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.7 on 8143 degrees of freedom
## Multiple R-squared:  0.1547, Adjusted R-squared:  0.153
## F-statistic: 87.68 on 17 and 8143 DF,  p-value: < 2.2e-16
```

Model 4: Multiple Linear Regression (Box Cox)

The p-value below shows that the probability of this variables to be irrelevant is very low. Lastly, R-squared is 0.22, which means this model explains 22% of the data's variation. Overall, this looks best model.

```r
# boxcox transformation model
insurance_boxcox <- preProcess(insurance_vif, c("BoxCox"))
in_bc_transformed <- predict(insurance_boxcox, insurance_vif)
model4 <- lm(TARGET_AMT ~ ., in_bc_transformed)
summary(model4)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = in_bc_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7274 -0.5441 -0.2230  0.5821  2.3657
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0890160  0.1095045   9.945  < 2e-16 ***
## KIDSDRIV     0.4292298  0.0744740   5.763 8.54e-09 ***
```

```
## AGE          -0.0006901  0.0011825  -0.584 0.559512
## HOMEKIDS      0.1135830  0.0602623   1.885 0.059491 .
## YOJ           0.0002315  0.0006369   0.363 0.716255
## INCOME       -0.0020014  0.0002492  -8.031 1.10e-15 ***
## PARENT1       0.1191388  0.0354780   3.358 0.000788 ***
## HOME_VAL     -0.0039288  0.0012432  -3.160 0.001582 **
## MSTATUS       0.1451800  0.0245088   5.924 3.28e-09 ***
## SEX           0.0044098  0.0249194   0.177 0.859543
## EDUCATION     0.0170122  0.0089206   1.907 0.056546 .
## JOB           0.0010812  0.0033090   0.327 0.743861
## TRAVTIME      0.0112922  0.0013998   8.067 8.23e-16 ***
## CAR_USE      -0.2703080  0.0197428 -13.691  < 2e-16 ***
## BLUEBOOK     -0.0013051  0.0002075  -6.291 3.32e-10 ***
## TIF          -0.0529241  0.0068487  -7.728 1.23e-14 ***
## CAR_TYPE      0.0526202  0.0077014   6.833 8.94e-12 ***
## RED_CAR      -0.0092348  0.0245359  -0.376 0.706645
## OLDCLAIM     -0.0059499  0.0105003  -0.567 0.570973
## CLM_FREQ      0.3649700  0.1403018   2.601 0.009303 **
## REVOKED       0.2552282  0.0260147   9.811  < 2e-16 ***
## MVR_PTS       0.1327377  0.0184202   7.206 6.27e-13 ***
## CAR_AGE      -0.0244435  0.0048421  -5.048 4.56e-07 ***
## URBANICITY   -0.5166161  0.0224956 -22.965  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7489 on 8137 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.2128
## F-statistic: 96.92 on 23 and 8137 DF,  p-value: < 2.2e-16
```

Model 1: Binary Logistic Regression

This model shows many variables with significant p-value. We will observe with following model whether AIC score improves.

```r
# original value model
logit_data <- data.frame(lapply(insurance_imputed, function(x) as.numeric(as.factor(x)))) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG)) %>%
  dplyr::select(-"TARGET_AMT")

model5 <- glm(TARGET_FLAG ~ ., family = "binomial", logit_data)
summary(model5)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = logit_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5102  -0.7264  -0.4161   0.6507   3.1100
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.215e-01  3.796e-01   1.110 0.266863
```

```
## KIDSDRIV     3.745e-01  6.035e-02   6.206 5.43e-10 ***
## AGE         -2.522e-03  3.907e-03  -0.646 0.518570
## HOMEKIDS     5.909e-02  3.651e-02   1.619 0.105549
## YOJ         -8.343e-03  7.603e-03  -1.097 0.272500
## INCOME      -1.445e-04  2.188e-05  -6.604 4.01e-11 ***
## PARENT1      3.650e-01  1.083e-01   3.369 0.000754 ***
## HOME_VAL    -7.799e-05  2.495e-05  -3.126 0.001772 **
## MSTATUS      5.224e-01  8.010e-02   6.522 6.94e-11 ***
## SEX          1.923e-02  8.797e-02   0.219 0.827006
## EDUCATION    3.421e-02  1.984e-02   1.724 0.084685 .
## JOB         -7.718e-03  1.130e-02  -0.683 0.494613
## TRAVTIME     1.536e-02  1.874e-03   8.198 2.45e-16 ***
## CAR_USE     -9.289e-01  6.835e-02 -13.591  < 2e-16 ***
## BLUEBOOK    -2.791e-04  4.696e-05  -5.944 2.79e-09 ***
## TIF         -5.460e-02  7.274e-03  -7.507 6.06e-14 ***
## CAR_TYPE     1.181e-01  1.788e-02   6.605 3.98e-11 ***
## RED_CAR     -2.648e-02  8.542e-02  -0.310 0.756603
## OLDCLAIM    -4.396e-05  4.495e-05  -0.978 0.328158
## CLM_FREQ     1.708e-01  3.206e-02   5.329 9.86e-08 ***
## REVOKED      7.655e-01  8.447e-02   9.062  < 2e-16 ***
## MVR_PTS      1.158e-01  1.358e-02   8.527  < 2e-16 ***
## CAR_AGE     -2.350e-02  5.779e-03  -4.067 4.76e-05 ***
## URBANICITY  -2.313e+00  1.127e-01 -20.530  < 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: 7423.6  on 8137  degrees of freedom
## AIC: 7471.6
##
## Number of Fisher Scoring iterations: 5
```

Model 2: Binary Logistic Regression (Stepwise)

This model's variables selection is better with better p-value. However AIC score has not improved.

```r
# stepwise transformed model
model6 <- stepAIC(model5, direction = "both", trace = FALSE)
summary(model6)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY, family = "binomial", data = logit_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5182  -0.7256  -0.4175   0.6533   3.0820
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.716e-01  2.659e-01   0.645 0.518786
## KIDSDRIV     3.699e-01  5.934e-02   6.234 4.55e-10 ***
## HOMEKIDS     6.353e-02  3.352e-02   1.896 0.058011 .
## INCOME      -1.518e-04  2.097e-05  -7.238 4.54e-13 ***
## PARENT1      3.759e-01  1.076e-01   3.494 0.000477 ***
## HOME_VAL    -8.008e-05  2.487e-05  -3.220 0.001284 **
## MSTATUS      5.310e-01  7.975e-02   6.658 2.78e-11 ***
## EDUCATION    3.638e-02  1.967e-02   1.850 0.064384 .
## TRAVTIME     1.533e-02  1.872e-03   8.190 2.60e-16 ***
## CAR_USE     -9.109e-01  6.201e-02 -14.690  < 2e-16 ***
## BLUEBOOK    -2.763e-04  4.588e-05  -6.024 1.71e-09 ***
## TIF         -5.450e-02  7.265e-03  -7.502 6.28e-14 ***
## CAR_TYPE     1.226e-01  1.546e-02   7.931 2.17e-15 ***
## CLM_FREQ     1.510e-01  2.519e-02   5.996 2.03e-09 ***
## REVOKED      7.364e-01  7.930e-02   9.286  < 2e-16 ***
## MVR_PTS      1.146e-01  1.341e-02   8.543  < 2e-16 ***
## CAR_AGE     -2.250e-02  5.625e-03  -4.001 6.31e-05 ***
## URBANICITY  -2.302e+00  1.123e-01 -20.500  < 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: 7427.2  on 8143  degrees of freedom
## AIC: 7463.2
##
## Number of Fisher Scoring iterations: 5
```

Model 3: Binary Logistic Regression (Box Cox)

This model too shows many variables with significant p-value. and the AIC score so far

```r
# boxcox transformation model
insurance_boxcox1 <- preProcess(logit_data, c("BoxCox"))
in_bc_transformed1 <- predict(insurance_boxcox1, logit_data)
model7 <- glm(TARGET_FLAG ~ ., family = "binomial", in_bc_transformed1)
summary(model7)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = in_bc_transformed1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3247  -0.7292  -0.4175   0.6740   3.1428
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.4577822  0.3769399   3.867  0.00011 ***
## KIDSDRIV     1.4365999  0.2440971   5.885 3.97e-09 ***
```

```
## AGE         -0.0017805  0.0040504  -0.440  0.66025
## HOMEKIDS     0.4486481  0.2130188   2.106  0.03519 *
## YOJ          0.0013092  0.0022019   0.595  0.55211
## INCOME      -0.0067510  0.0008460  -7.980 1.46e-15 ***
## PARENT1      0.2580577  0.1181181   2.185  0.02891 *
## HOME_VAL    -0.0129153  0.0041388  -3.121  0.00181 **
## MSTATUS      0.5579178  0.0856969   6.510 7.50e-11 ***
## SEX         -0.0041099  0.0876479  -0.047  0.96260
## EDUCATION    0.0453297  0.0302872   1.497  0.13448
## JOB         -0.0043591  0.0112373  -0.388  0.69808
## TRAVTIME     0.0422292  0.0049851   8.471  < 2e-16 ***
## CAR_USE     -0.9168905  0.0681340 -13.457  < 2e-16 ***
## BLUEBOOK    -0.0047031  0.0007108  -6.617 3.67e-11 ***
## TIF         -0.1818143  0.0238004  -7.639 2.19e-14 ***
## CAR_TYPE     0.1993687  0.0278749   7.152 8.54e-13 ***
## RED_CAR     -0.0292156  0.0854130  -0.342  0.73231
## OLDCLAIM    -0.0191629  0.0316683  -0.605  0.54510
## CLM_FREQ     1.1302182  0.4227955   2.673  0.00751 **
## REVOKED      0.7464926  0.0810774   9.207  < 2e-16 ***
## MVR_PTS      0.4123012  0.0621576   6.633 3.29e-11 ***
## CAR_AGE     -0.0844190  0.0166790  -5.061 4.16e-07 ***
## URBANICITY  -2.2923436  0.1131763 -20.255  < 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: 7414.5  on 8137  degrees of freedom
## AIC: 7462.5
##
## Number of Fisher Scoring iterations: 5
```

5. Select Models

Multiple Linear Regression Metrics

```r
# predict
predict <- predict(model5, insurance_eval_imputed, interval = "prediction")
eval <- table(as.integer(predict > .5))
print(paste(eval[1], "car crash has not happened", "and", eval[2], "car crash has happened"))
```

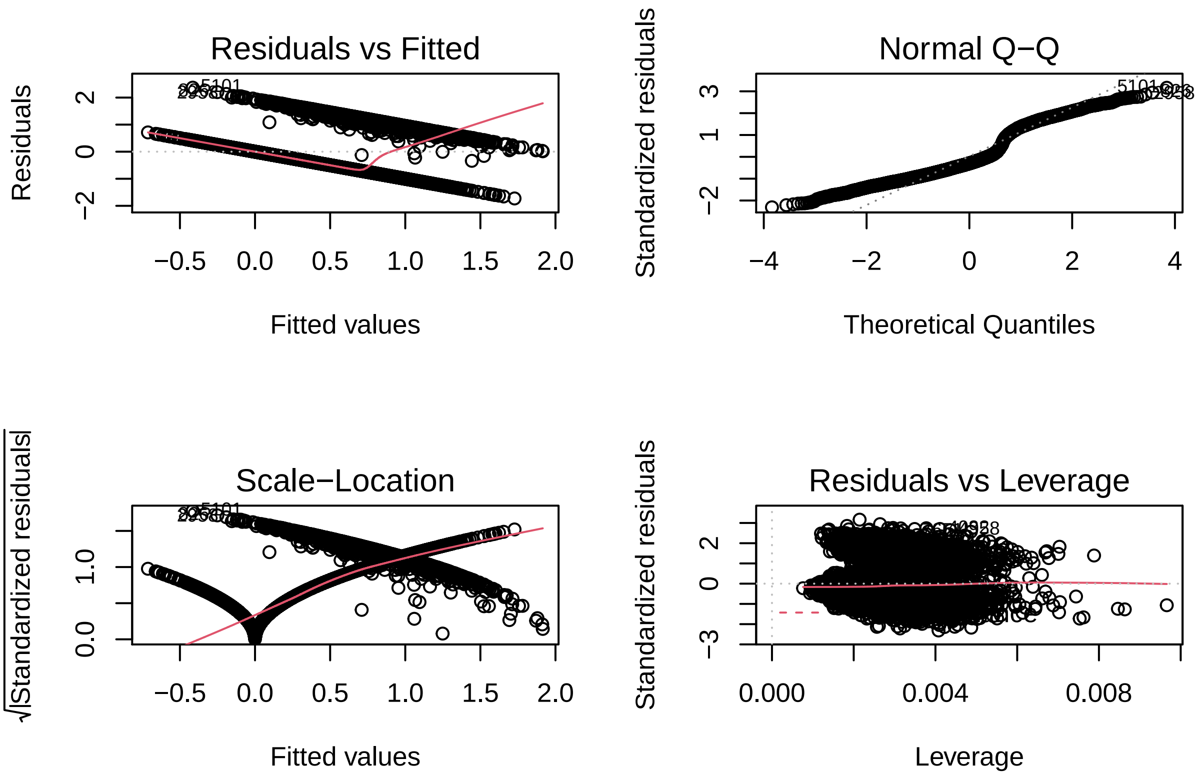
```
## [1] "1783 car crash has not happened and 358 car crash has happened"
```

```r
# comparing all binary logistic models using various measures
a1 <- mean((summary(model1))$residuals^2)
a2 <- mean((summary(model2))$residuals^2)
a3 <- mean((summary(model3))$residuals^2)
a4 <- mean((summary(model4))$residuals^2)
a5 <- rbind(a1, a2, a3, a4)

b1 <- summary(model2)$r.squared
```

```r
b2 <- summary(model3)$r.squared
b3 <- summary(model1)$r.squared
b4 <- summary(model4)$r.squared
b5 <- rbind(b1, b2, b3, b4)

c1 <- summary(model1)$fstatistic
c2 <- summary(model2)$fstatistic
c3 <- summary(model3)$fstatistic
c4 <- summary(model4)$fstatistic
c5 <- rbind(c1, c2, c3, c4)

mlr_metrics <- data.frame(cbind(a5, b5, c5), row.names = c("Model 1", "Model 2", "Model 3", "Model 4"))
colnames(mlr_metrics) <- c("MSE", "R-Squared", "value", "numdf", "dendf")
kable(mlr_metrics) %>%
  kable_styling(full_width = T) %>%
  add_header_above(c(" ", " " = 2, "F-Statistic" = 3))
```

			F-Statistic		
	MSE	R-Squared	value	numdf	dendf
Model 1	2.201851e+05	0.1548631	51.79085	23	6424
Model 2	2.220132e+05	0.1547258	64.82720	23	8137
Model 3	2.220493e+05	0.1564228	87.68001	17	8143
Model 4	5.591757e-01	0.2150415	96.91969	23	8137

```r
# residual plot
par(mfrow=c(2,2))
plot(model4)
```




```r
# prediction
prediction <- predict(model4, insurance_eval_imputed, interval = "prediction")
```

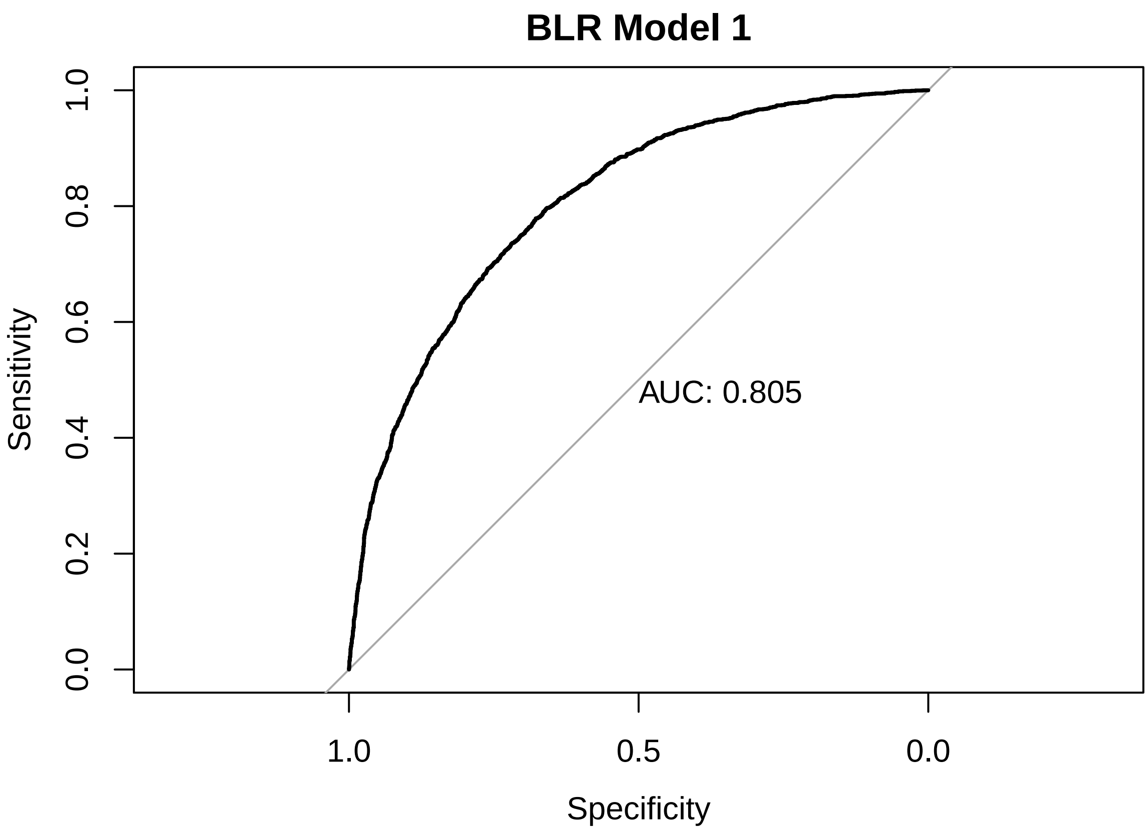
The variance of residuals are not uniform which indicates our explanatory variable is not an complete picture of data, also not normally distributed, this is not good model selection.

Binary Logistic Regression Metrics

```r
# comparing all binary logistic models using various measures
c1 <- confusionMatrix(as.factor(as.integer(fitted(model5) > .5)), as.factor(model5$y), positive = "1")
c2 <- confusionMatrix(as.factor(as.integer(fitted(model6) > .5)), as.factor(model6$y), positive = "1")
c3 <- confusionMatrix(as.factor(as.integer(fitted(model7) > .5)), as.factor(model7$y), positive = "1")

roc1 <- roc(logit_data$TARGET_FLAG,  predict(model5, logit_data, interval = "prediction"))
roc2 <- roc(logit_data$TARGET_FLAG,  predict(model6, logit_data, interval = "prediction"))
roc3 <- roc(logit_data$TARGET_FLAG,  predict(model7, logit_data, interval = "prediction"))

metrics1 <- c(c1$overall[1], "Class. Error Rate" = 1 - as.numeric(c1$overall[1]), c1$byClass[c(1, 2, 5,
metrics2 <- c(c2$overall[1], "Class. Error Rate" = 1 - as.numeric(c2$overall[1]), c2$byClass[c(1, 2, 5,
metrics3 <- c(c3$overall[1], "Class. Error Rate" = 1 - as.numeric(c3$overall[1]), c3$byClass[c(1, 2, 5,

kable(cbind(metrics1, metrics2, metrics3), col.names = c("BLR Model 1", "BLR Model 2", "BLR Model 3"))
  kable_styling(full_width = T)
```

	BLR Model 1	BLR Model 2	BLR Model 3
Accuracy	0.7866683	0.7876486	0.7858106
Class. Error Rate	0.2133317	0.2123514	0.2141894
Sensitivity	0.3975848	0.3952624	0.3934046
Specificity	0.9260985	0.9282623	0.9264314
Precision	0.6584615	0.6638066	0.6570985
F1	0.4958008	0.4954876	0.4921557
AUC	0.8050443	0.8048232	0.5779827

```
# plotting roc curve of model 3
plot(roc(logit_data$TARGET_FLAG,  predict(model5, logit_data, interval = "prediction")), print.auc = TRU
```



Upon all three models' accuracy, classification error rate, precision, sensitivity, specificity, F1 score, AUC, and confusion matrix. Even though all models yield similar metrics value, BLR model 1 has the highest AUC value. We will pick Model 1 on BLR with imputed values for our prediction