# Car Crush


## Anjal Hussan

# Introduction

In this project, I 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.

# Statement of the Problem

The purpose of this report 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.

# Data Exploration

Let’s take a look in the structure of our train data set - excluding the first column **index** which it is not to be used. Evaluation data set structure is similar to the train data set and will go through same

```
## 'data.frame':    8161 obs. of  25 variables:
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/
Urban" ...
```

We can see that the training data has 8161 observations(rows) and 25 variables (columns). Of these 25 columns, many are of factors type but were imported as characters or doubles - there will be properly converted in the preparation section. Also, there may be some ordinal levels within some of the factors.

Below we display a summary of each feature.

```
##   TARGET_FLAG       TARGET_AMT        KIDSDRIV           AGE
##  Min.   :0.0000   Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median :0.0000   Median :     0   Median :0.0000   Median :45.00
##  Mean   :0.2638   Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :1.0000   Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                     NA's   :6
##     HOMEKIDS           YOJ          INCOME            PARENT1
##  Min.   :0.0000   Min.   : 0.0   Length:8161        Length:8161
##  1st Qu.:0.0000   1st Qu.: 9.0   Class :character   Class :character
##  Median :0.0000   Median :11.0   Mode  :character   Mode  :character
##  Mean   :0.7212   Mean   :10.5
##  3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :5.0000   Max.   :23.0
##                   NA's   :454
##    HOME_VAL           MSTATUS              SEX             EDUCATION
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      JOB               TRAVTIME        CAR_USE            BLUEBOOK
##  Length:8161        Min.   :  5.00   Length:8161        Length:8161
##  Class :character   1st Qu.: 22.00   Class :character   Class :character
##  Mode  :character   Median : 33.00   Mode  :character   Mode  :character
##                     Mean   : 33.49
##                     3rd Qu.: 44.00
##                     Max.   :142.00
##
##       TIF           CAR_TYPE           RED_CAR            OLDCLAIM
##  Min.   : 1.000   Length:8161        Length:8161        Length:8161
##  1st Qu.: 1.000   Class :character   Class :character   Class :character
##  Median : 4.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5.351
##  3rd Qu.: 7.000
##  Max.   :25.000
##
##     CLM_FREQ        REVOKED             MVR_PTS         CAR_AGE
##  Min.   :0.0000   Length:8161        Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Class :character   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000   Mode  :character   Median : 1.000   Median : 8.000
##  Mean   :0.7986                      Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000                      3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000                      Max.   :13.000   Max.   :28.000
##                                                       NA's   :510
##   URBANICITY
##  Length:8161
##  Class :character
##  Mode  :character
##
##
##
##
```

We can observe the followings:

**KIDSDRIV**: Max is 4

**AGE**: age is 16 is the youngest and oldest 81. There are 6 NA values

**HOMEKIDS**: Max is 5

**TRAVTIME**: 75% of the population is below 44 but the Max value is 142. It looks like there may be some outliers here.

**TIF**: The majority of people are not long time customers

**CLM_FREQ**: Maximum is over 5 years

**MVR_PTS**: 75% have 3 or less, maximum is 13

**CAR_AGE**: Strange!. The minimum -3 and Max is 28. There are 510 NA values. These negative values will have to be excluded from the analysis.

**INCOME** - **BLUEBOOK** - **HOME_VAL** - **OLDCLAIM** : These are numerical variables that need to be converted accordingly.

# Convertion to numerical

As can be seen below, these four features are now corrected represented.

```
##      INCOME          HOME_VAL         BLUEBOOK        OLDCLAIM    
##  Min.   :     0   Min.   :     0   Min.   : 1500   Min.   :    0  
##  1st Qu.: 28097   1st Qu.:     0   1st Qu.: 9280   1st Qu.:    0  
##  Median : 54028   Median :161160   Median :14440   Median :    0  
##  Mean   : 61898   Mean   :154867   Mean   :15710   Mean   : 4037  
##  3rd Qu.: 85986   3rd Qu.:238724   3rd Qu.:20850   3rd Qu.: 4636  
##  Max.   :367030   Max.   :885282   Max.   :69740   Max.   :57037  
##  NA's   :445      NA's   :464
```

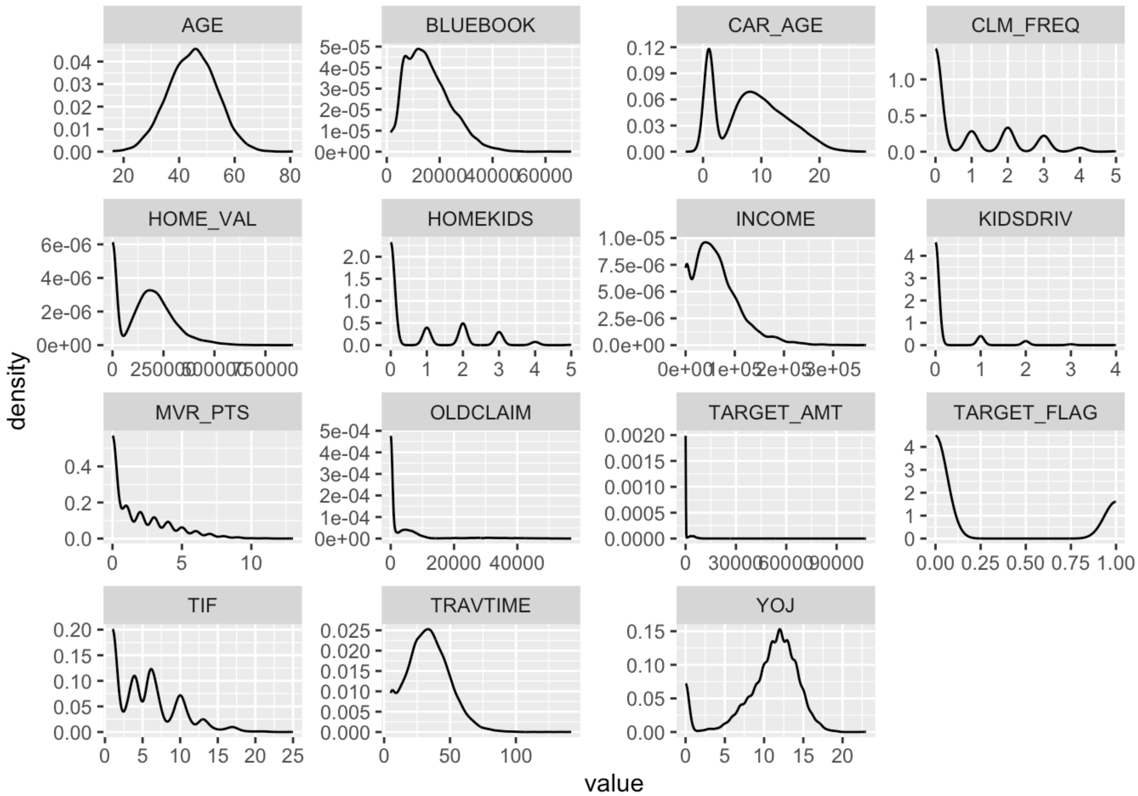
# Missing Values

```
## TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ 
##           0           0           0           6           0         454 
##      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION 
##         445           0         464           0           0           0 
##         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE 
##           0           0           0           0           0           0 
##     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE 
##           0           0           0           0           0         510 
##  URBANICITY 
##           0
```

There are missing values in several variables for a total of 1,879 NA’s or about 1% of the total dataset.

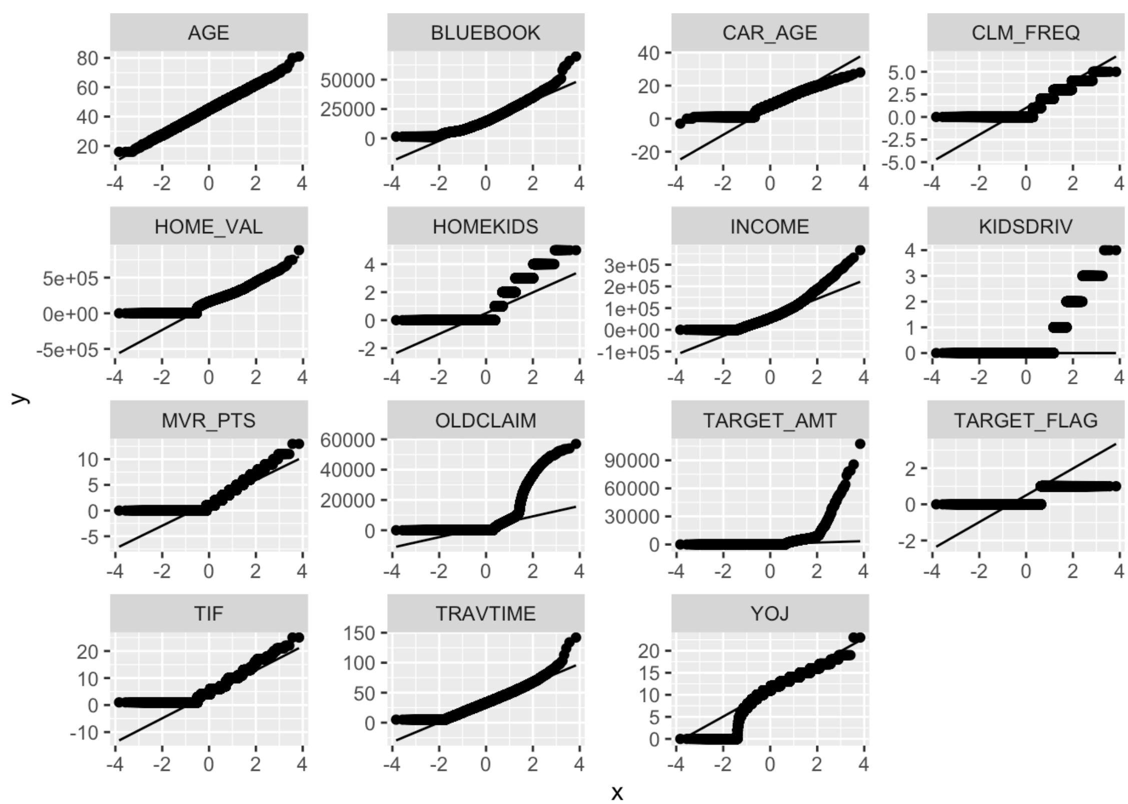
# Univariate Distribution - Histograms

Below the numeric feature distributions are displayed.

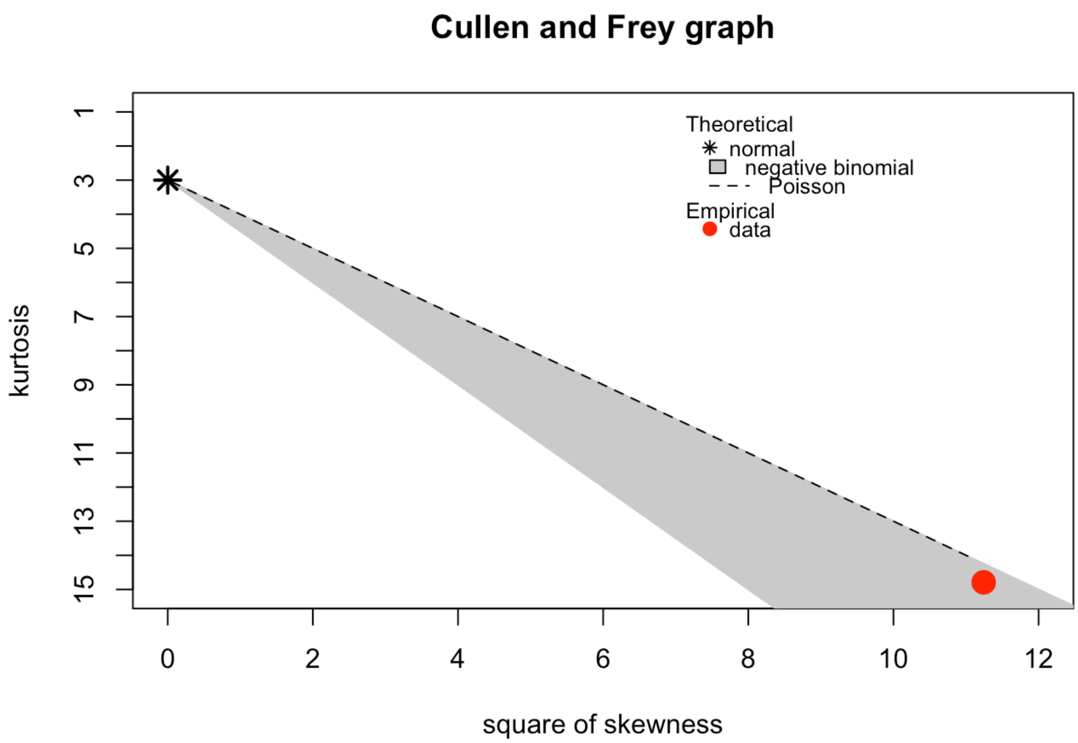



We can see that AGE, BLUEBOOK, CAR_AGE, HOME_VAL, INCOME, TRAVTIME and YOJ resemblance somewhat a normal distribution while CLM_FREQ, HOMEKIDS, KIDSDRIV, MVR_PTS, OLDCLAIM, TARGET_AMT, TIF resemblance either a binomial or Poisson distribution.

Let's investigate using a qq_plot:




In order to descriptive the distribution, we used function 'descdist' from package 'fitdistrplus'. We display one output for illustrative purposes - on featue **KIDSDRIV**, and results for all other features are shown below:




```
## summary statistics
## ------
## min:  0   max:  4
## median:  0
## mean:  0.1710575
## estimated sd:  0.5115341
## estimated skewness:  3.35307
## estimated kurtosis:  14.79177
```

We can observe the followings:

**AGE**: normal distribution

*BLUEBOOK**: quasi-normal/lognormal - skewed distribution with heavy tails

**CAR_AGE**: quasi-normal/lognormal - skewed distribution with high frequency of <1, including negative.

**CLM_FREQ**: not normal - poisson type

**HOME_VAL**: quasi-normal - skewed distribution with heavy tails

**HOMEKIDS*: Beta distribution

**INCOME**: quasi-normal - skewed distribution with heavy tails

**KIDSDRIV**: Negative binomial / Poisson

**MVR_PTS**: Beta distribution

**TARGET_AMT**: Gamma distribution

**TIF**: Poisson distribution

**TRAVTIME**: quasi-normal - skewed distribution with heavy tails

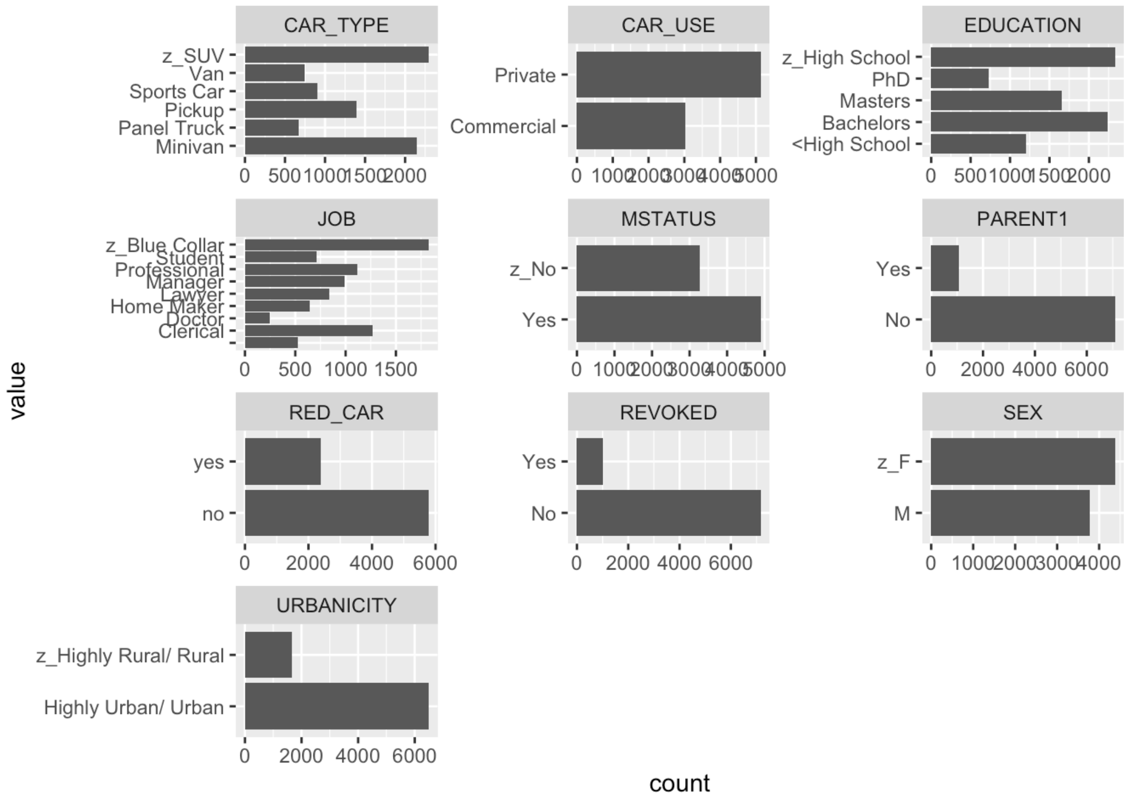
**YOJ**: normal distribution with heavy tail

**OLDCLAIM**: Poisson distribution

**CLM_FREQ**: Beta distribution

Some of the variables present a lot of zeros which could be explained as lack of data, and as such should be excluded, for example in HOMEVAL, while in others they are a rightful part of the distribution, and should be considered in the analysis, such as in INCOME, CLM_FREQ, HOMEKIDS, KIDSDRV, MVR_PTS, OLDCLAIM, etc.

For the categorical features, we will displayed their distribution using bar charts.

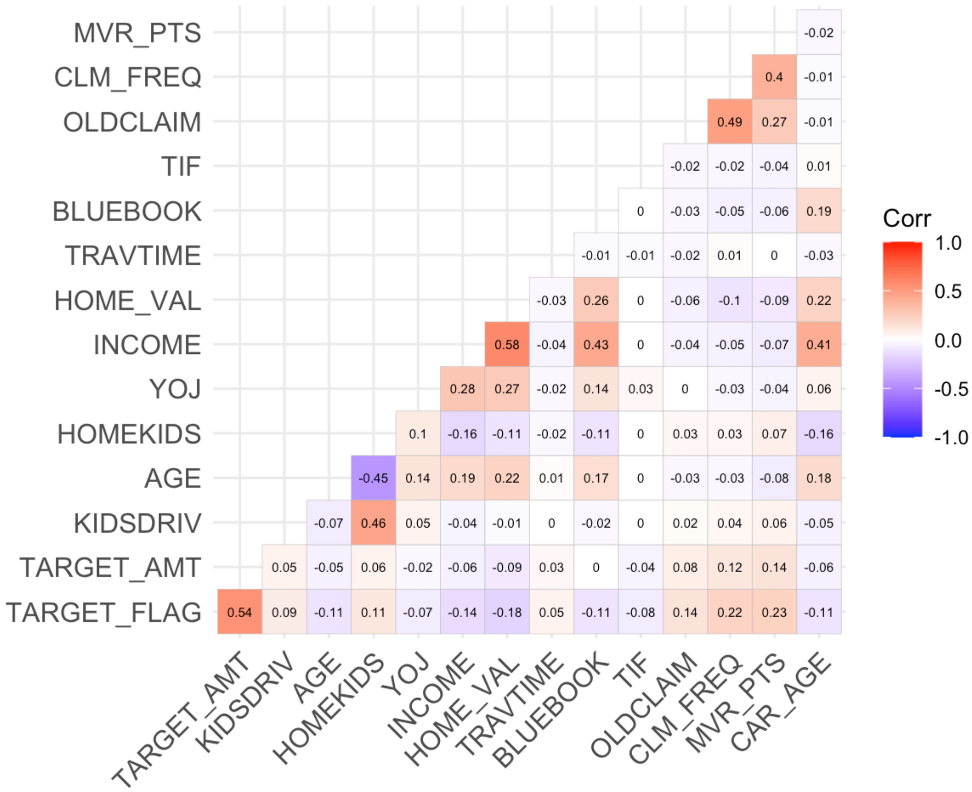


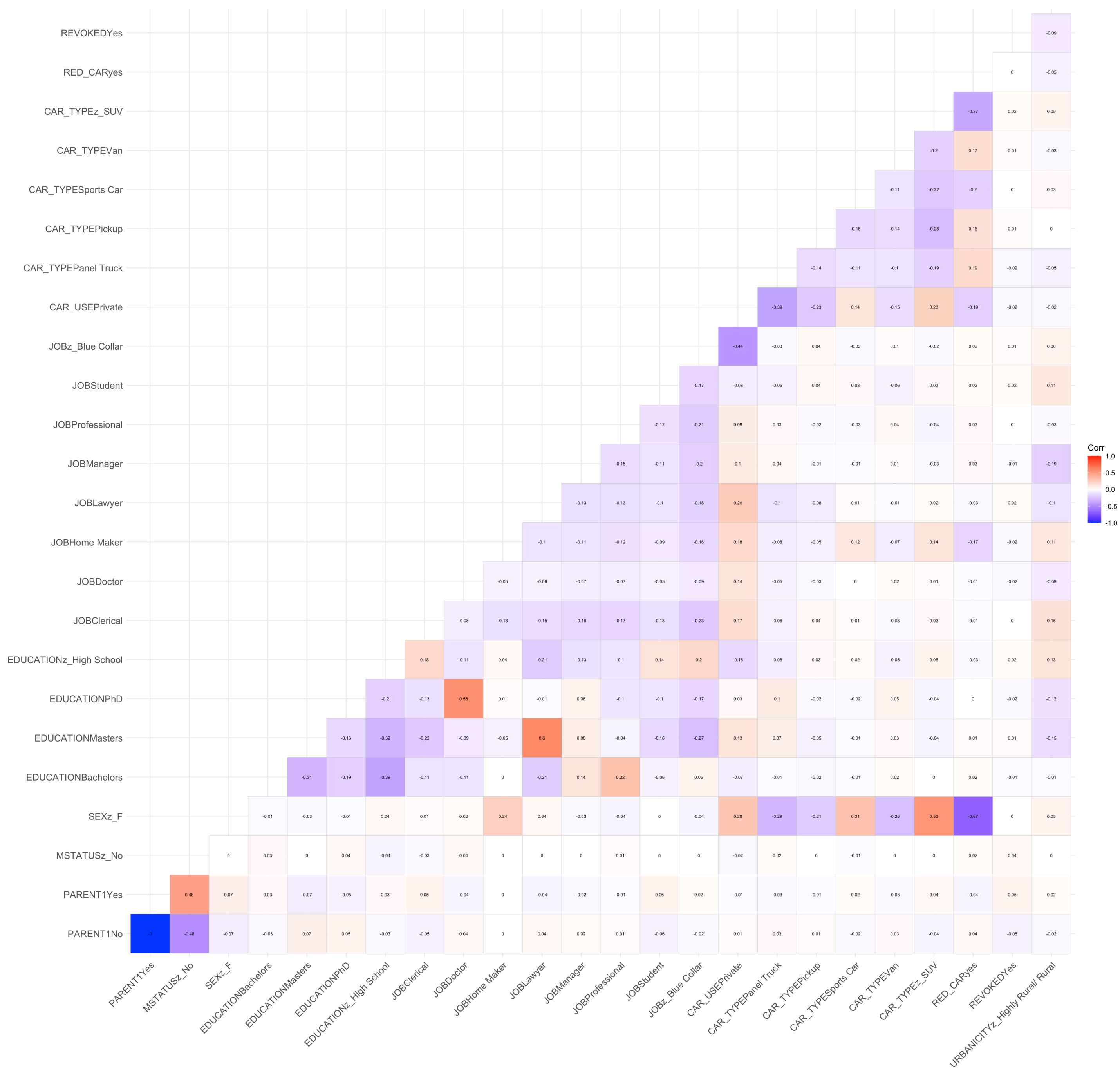

Some of the features have several sub-categories, like **CAR_TYPE**, **EDUCATION**, and **JOB**, while the other features are binary in nature. Interaction between these sub-categories and the continous variables to be taken into consideration while building models.

# Correlation matrix

Considering the number of variables and sub-categories within the discrete features, the correlation matrix visualization is challenging. We will then show two matrices one with numeric only and other with discrete variable. Analysis are based on the whole dataset, though.

We also ran ‘pairs’ a function that produces a matrix of scatterplots - not displayed here due to size.


MVR_PTS
-0.02
CLM_FREQ
0.4 -0.01
OLDCLAIM
0.49 0.27 -0.01
TIF
-0.02 -0.02 -0.04 0.01
BLUEBOOK
0 -0.03 -0.05 -0.06 0.19
TRAVTIME
-0.01 -0.01 -0.02 0.01 0 -0.03
HOME_VAL
-0.03 0.26 0 -0.06 -0.1 -0.09 0.22
INCOME
0.58 -0.04 0.43 0 -0.04 -0.05 -0.07 0.41
YOJ
0.28 0.27 -0.02 0.14 0.03 0 -0.03 -0.04 0.06
HOMEKIDS
0.1 -0.16 -0.11 -0.02 -0.11 0 0.03 0.03 0.07 -0.16
AGE
-0.45 0.14 0.19 0.22 0.01 0.17 0 -0.03 -0.03 -0.08 0.18
KIDSDRIV
-0.07 0.46 0.05 -0.04 -0.01 0 -0.02 0 0.02 0.04 0.06 -0.05
TARGET_AMT
0.05 -0.05 0.06 -0.02 -0.06 -0.09 0.03 0 -0.04 0.08 0.12 0.14 -0.06
TARGET_FLAG
0.54 0.09 -0.11 0.11 -0.07 -0.14 -0.18 0.05 -0.11 -0.08 0.14 0.22 0.23 -0.11
TARGET_AMT
KIDSDRIV
AGE
HOMEKIDS
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
CLM_FREQ
MVR_PTS
CAR_AGE
Corr
1.0
0.5
0.0
-0.5
-1.0


REVOKEDYes
RED_CARyes
CAR_TYPEz_SUV
CAR_TYPEVan
CAR_TYPESports Car
CAR_TYPEPickup
CAR_TYPEPanel Truck
CAR_USEPrivate
JOBz_Blue Collar
JOBStudent
JOBProfessional
JOBManager
JOBLawyer
JOBHome Maker
JOBDoctor
JOBClerical
EDUCATIONz_High School
EDUCATIONPhD
EDUCATIONMasters
EDUCATIONBachelors
SEXz_F
MSTATUSz_No
PARENT1Yes
PARENT1No
PARENT1Yes
MSTATUSz_No
SEXz_F
EDUCATIONBachelors
EDUCATIONMasters
EDUCATIONPhD
EDUCATIONz_High School
JOBClerical
JOBDoctor
JOBHome Maker
JOBLawyer
JOBManager
JOBProfessional
JOBStudent
JOBz_Blue Collar
CAR_USEPrivate
CAR_TYPEPanel Truck
CAR_TYPEPickup
CAR_TYPESports Car
CAR_TYPEVan
CAR_TYPEz_SUV
RED_CARyes
REVOKEDYes
URBANICITYz_Highly Rural/ Rural
Corr
1.0
0.5
0.0
-0.5
-1.0

Some observations from the above charts:

- positive correlation:
  Income and HomeVal
  Income and BlueBook
  SexF and CarType SUV
  Phd Degree and Job as Doctor Master's Degree and Job as Lawyer Income and Education Income and Urbanicity Urban
- Negative correlation:
  Age and HomeKids
  HomeKids and CarAge
  Urbanicity Rural and Claim frequency Urbanicity Rural and BlueBook

# Evaluation dataset

Procedures described above were also applied to the evaluation set.

# Data Prep

Looking at the plots we see we have to make a few changes to some variables. We'll make HOMEKIDS boolean instead of a factor. For the rows where AGE and CAR_AGE are less than zero, we make them equal to 0. For blank JOBS we label those as "Unknown". Finally, change Education to 1 if PhD and Masters.

# Missing Data

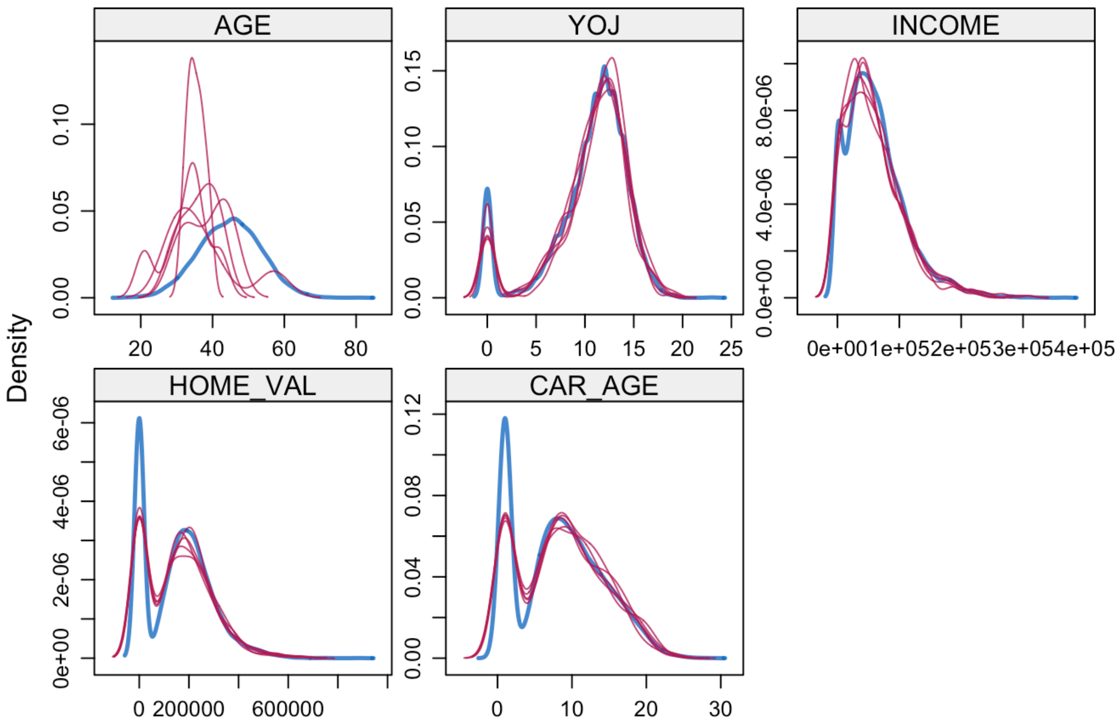
We have missing data for income, yoj, home_val, and car_age variables.

```
summary(Train_Data)
```

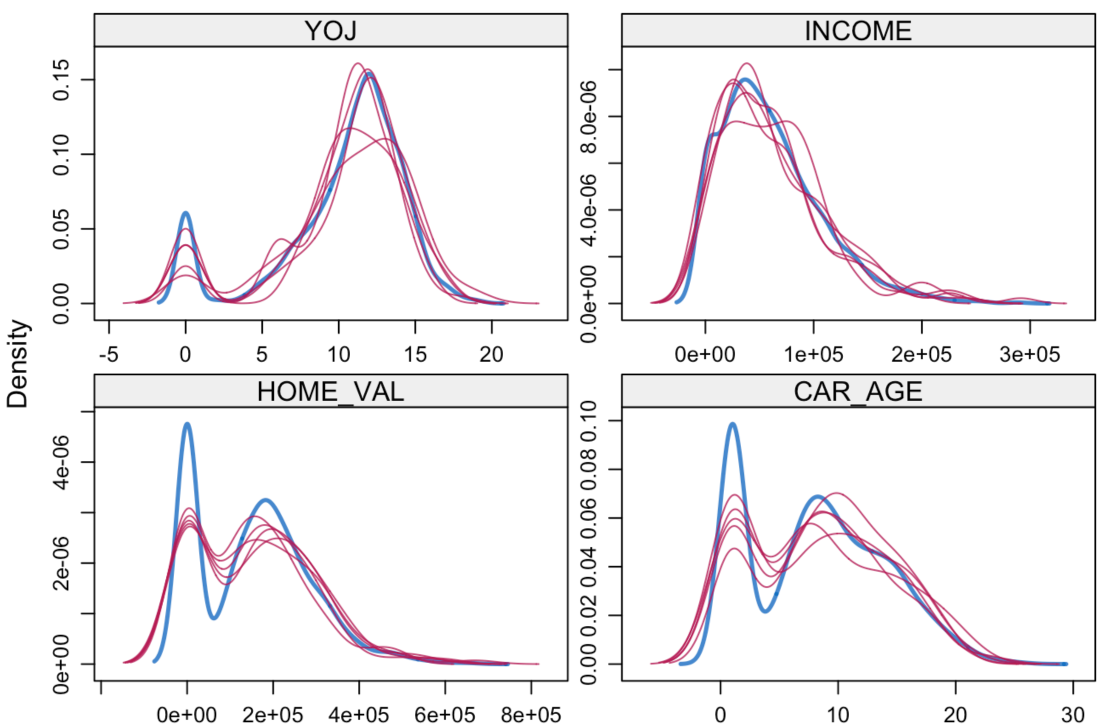
```
##   TARGET_FLAG       TARGET_AMT        KIDSDRIV          AGE       
##  Min.   :0.0000   Min.   :     0   Min.   :0.0000   Min.   :16.00  
##  1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00  
##  Median :0.0000   Median :     0   Median :0.0000   Median :45.00  
##  Mean   :0.2638   Mean   :  1504   Mean   :0.1711   Mean   :44.79  
##  3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00  
##  Max.   :1.0000   Max.   :107586   Max.   :4.0000   Max.   :81.00  
##                                                     NA's   :6      
##     HOMEKIDS           YOJ           INCOME         PARENT1         
##  Min.   :0.0000   Min.   : 0.0   Min.   :     0   Length:8161       
##  1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   Class :character  
##  Median :0.0000   Median :11.0   Median : 54028   Mode  :character  
##  Mean   :0.3519   Mean   :10.5   Mean   : 61898                     
##  3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986                     
##  Max.   :1.0000   Max.   :23.0   Max.   :367030                     
##                   NA's   :454    NA's   :445                        
##     HOME_VAL        MSTATUS              SEX              EDUCATION     
##  Min.   :     0   Length:8161        Length:8161        Min.   :0.0000  
##  1st Qu.:     0   Class :character   Class :character   1st Qu.:0.0000  
##  Median :161160   Mode  :character   Mode  :character   Median :1.0000  
##  Mean   :154867                                         Mean   :0.7076  
##  3rd Qu.:238724                                         3rd Qu.:1.0000  
##  Max.   :885282                                         Max.   :1.0000  
##  NA's   :464                                                            
##             JOB          TRAVTIME        CAR_USE             BLUEBOOK    
##  z_Blue Collar:1825   Min.   :  5.00   Length:8161        Min.   : 1500  
##  Clerical     :1271   1st Qu.: 22.00   Class :character   1st Qu.: 9280  
##  Professional :1117   Median : 33.00   Mode  :character   Median :14440  
##  Manager      : 988   Mean   : 33.49                      Mean   :15710  
##  Lawyer       : 835   3rd Qu.: 44.00                      3rd Qu.:20850  
##  Student      : 712   Max.   :142.00                      Max.   :69740  
##  (Other)      :1413                                                      
##       TIF            CAR_TYPE           RED_CAR            OLDCLAIM    
##  Min.   : 1.000   Length:8161        Length:8161        Min.   :    0  
##  1st Qu.: 1.000   Class :character   Class :character   1st Qu.:    0  
##  Median : 4.000   Mode  :character   Mode  :character   Median :    0  
##  Mean   : 5.351                                         Mean   : 4037  
##  3rd Qu.: 7.000                                         3rd Qu.: 4636  
##  Max.   :25.000                                         Max.   :57037  
##                                                                        
##     CLM_FREQ        REVOKED             MVR_PTS         CAR_AGE      
##  Min.   :0.0000   Length:8161        Min.   : 0.000   Min.   : 0.000  
##  1st Qu.:0.0000   Class :character   1st Qu.: 0.000   1st Qu.: 1.000  
##  Median :0.0000   Mode  :character   Median : 1.000   Median : 8.000  
##  Mean   :0.7986                      Mean   : 1.696   Mean   : 8.329  
##  3rd Qu.:2.0000                      3rd Qu.: 3.000   3rd Qu.:12.000  
##  Max.   :5.0000                      Max.   :13.000   Max.   :28.000  
##                                                       NA's   :510     
##   URBANICITY       
##  Length:8161       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
## 
```

We assume the missing data are Missing at Random and choose to impute. The reason we want to impute the missing data rather than replacing with mean or median because of large number of missing values. If we're replacing with mean or median on the large number of missing values, can result in loss of variation in data. We're imputing the missing data using the MICE package. The method of predictive mean matching (PMM) is selected for continuous variables.

```
## Warning: Number of logged events: 8
```




```
## Warning: Number of logged events: 10
```




# Building Models

# Binary Model 1 - All Variables

Our first model will seek to create a baseline using binary response variable, using a logistic regression model that contains all of our features.

```
##
## Call:
## glm(formula = factor(TARGET_FLAG) ~ ., family = binomial, data = subset(complete_train_data,
##     select = -c(TARGET_AMT)))
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.357e-01  3.070e-01  -1.745 0.081007 .
## KIDSDRIV                        3.495e-01  5.980e-02   5.846 5.05e-09 ***
## AGE                             2.251e-03  4.131e-03   0.545 0.585868
## HOMEKIDS                        2.958e-01  9.717e-02   3.044 0.002335 **
## YOJ                            -1.427e-02  8.349e-03  -1.709 0.087488 .
## INCOME                         -3.902e-06  1.085e-06  -3.597 0.000322 ***
## PARENT1Yes                      2.302e-01  1.207e-01   1.908 0.056390 .
## HOME_VAL                       -1.284e-06  3.402e-07  -3.775 0.000160 ***
## MSTATUSz_No                     5.458e-01  8.654e-02   6.306 2.85e-10 ***
## SEXz_F                         -9.138e-02  1.120e-01  -0.816 0.414588
## EDUCATION                      -1.261e-01  1.349e-01  -0.935 0.349596
## JOBDoctor                      -8.496e-01  2.661e-01  -3.193 0.001408 **
## JOBHome Maker                  -2.713e-01  1.422e-01  -1.908 0.056391 .
## JOBLawyer                      -4.434e-01  1.799e-01  -2.465 0.013697 *
## JOBManager                     -1.078e+00  1.399e-01  -7.709 1.27e-14 ***
## JOBProfessional                -3.875e-01  1.182e-01  -3.279 0.001041 **
## JOBStudent                     -2.387e-01  1.315e-01  -1.816 0.069435 .
## JOBUnknown                     -4.847e-01  1.962e-01  -2.471 0.013477 *
## JOBz_Blue Collar               -1.027e-01  1.066e-01  -0.963 0.335505
## TRAVTIME                        1.442e-02  1.879e-03   7.675 1.65e-14 ***
## CAR_USEPrivate                 -7.352e-01  8.697e-02  -8.454  < 2e-16 ***
## BLUEBOOK                       -2.090e-05  5.272e-06  -3.964 7.37e-05 ***
## TIF                            -5.536e-02  7.338e-03  -7.545 4.54e-14 ***
## CAR_TYPEPanel Truck             5.821e-01  1.602e-01   3.634 0.000279 ***
## CAR_TYPEPickup                  5.696e-01  9.982e-02   5.706 1.16e-08 ***
## CAR_TYPESports Car              1.014e+00  1.297e-01   7.817 5.42e-15 ***
## CAR_TYPEVan                     6.238e-01  1.258e-01   4.957 7.14e-07 ***
## CAR_TYPEz_SUV                   7.597e-01  1.111e-01   6.840 7.90e-12 ***
## RED_CARyes                     -1.721e-02  8.631e-02  -0.199 0.841988
## OLDCLAIM                       -1.411e-05  3.907e-06  -3.611 0.000305 ***
## CLM_FREQ                        1.955e-01  2.851e-02   6.858 6.97e-12 ***
## REVOKEDYes                      8.932e-01  9.125e-02   9.788  < 2e-16 ***
## MVR_PTS                         1.133e-01  1.361e-02   8.324  < 2e-16 ***
## CAR_AGE                        -1.985e-02  6.885e-03  -2.882 0.003947 **
## URBANICITYz_Highly Rural/ Rural -2.375e+00  1.124e-01 -21.125  < 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: 7306.3  on 8126  degrees of freedom
## AIC: 7376.3
##
## Number of Fisher Scoring iterations: 5
```

The AIC result from the binomial model can be derived using the logit link function.

# Binary Model 2 - Hand Pick Model

We can see that from saturated model above that variables *AGE*, *YOJ*, *PARENT1*, *SEX*, *EDUCATION*, *JOB*, and *RED_CAR* have p values greater than 0.05. These variables will be dropped to build the next model.

Also, we see some predictors are skewed and so we take log of them to build model 2.

```
##
## Call:
## glm(formula = factor(TARGET_FLAG) ~ KIDSDRIV + HOMEKIDS + log(INCOME +
##     1) + log(HOME_VAL + 1) + MSTATUS + log(TRAVTIME) + CAR_USE +
##     log(BLUEBOOK) + TIF + CAR_TYPE + log(OLDCLAIM + 1) + CLM_FREQ +
##     REVOKED + MVR_PTS + log(CAR_AGE + 1) + URBANICITY, family = binomial,
##     data = subset(complete_train_data, select = -c(TARGET_AMT)))
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     2.046758   0.535653   3.821 0.000133 ***
## KIDSDRIV                        0.335379   0.057712   5.811 6.20e-09 ***
## HOMEKIDS                        0.415878   0.066314   6.271 3.58e-10 ***
## log(INCOME + 1)                -0.065998   0.009875  -6.683 2.34e-11 ***
## log(HOME_VAL + 1)              -0.021055   0.006203  -3.394 0.000688 ***
## MSTATUSz_No                     0.622958   0.071276   8.740  < 2e-16 ***
## log(TRAVTIME)                   0.414613   0.050846   8.154 3.51e-16 ***
## CAR_USEPrivate                 -0.868973   0.069507 -12.502  < 2e-16 ***
## log(BLUEBOOK)                  -0.370916   0.053708  -6.906 4.98e-12 ***
## TIF                            -0.051603   0.007254  -7.114 1.13e-12 ***
## CAR_TYPEPanel Truck             0.321091   0.131765   2.437 0.014816 *
## CAR_TYPEPickup                  0.508296   0.096798   5.251 1.51e-07 ***
## CAR_TYPESports Car              0.888776   0.105753   8.404  < 2e-16 ***
## CAR_TYPEVan                     0.529241   0.117897   4.489 7.16e-06 ***
## CAR_TYPEz_SUV                   0.707195   0.084156   8.403  < 2e-16 ***
## log(OLDCLAIM + 1)               0.024291   0.012290   1.976 0.048099 *
## CLM_FREQ                        0.083958   0.042795   1.962 0.049775 *
## REVOKEDYes                      0.723364   0.080234   9.016  < 2e-16 ***
## MVR_PTS                         0.109194   0.013872   7.871 3.51e-15 ***
## log(CAR_AGE + 1)               -0.274182   0.035892  -7.639 2.19e-14 ***
## URBANICITYz_Highly Rural/ Rural -2.234275   0.112526 -19.856  < 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: 7435.2  on 8140  degrees of freedom
## AIC: 7477.2
##
## Number of Fisher Scoring iterations: 5
```

# Binary Model 3 - Forward Step Model

We will now build a model using forward selection in order to compare if using forward selection is better than hand picking values to create a model.

We can use the same stepAIC function to build the third model. The forward selection approach starts from the null model and adds a variable that improves the model the most, one at a time, until the stopping criterion is met. We can see the result is different compared to the backward selection approach. The AIC is a little higher.

```
##
## Call:
## glm(formula = factor(TARGET_FLAG) ~ 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,
##     family = binomial, data = subset(complete_train_data, select = -c(TARGET_AMT)))
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.357e-01  3.070e-01  -1.745 0.081007 .
## KIDSDRIV                        3.495e-01  5.980e-02   5.846 5.05e-09 ***
## AGE                             2.251e-03  4.131e-03   0.545 0.585868
## HOMEKIDS                        2.958e-01  9.717e-02   3.044 0.002335 **
## YOJ                            -1.427e-02  8.349e-03  -1.709 0.087488 .
## INCOME                         -3.902e-06  1.085e-06  -3.597 0.000322 ***
## PARENT1Yes                      2.302e-01  1.207e-01   1.908 0.056390 .
## HOME_VAL                       -1.284e-06  3.402e-07  -3.775 0.000160 ***
## MSTATUSz_No                     5.458e-01  8.654e-02   6.306 2.85e-10 ***
## SEXz_F                         -9.138e-02  1.120e-01  -0.816 0.414588
## EDUCATION                      -1.261e-01  1.349e-01  -0.935 0.349596
## JOBDoctor                      -8.496e-01  2.661e-01  -3.193 0.001408 **
## JOBHome Maker                  -2.713e-01  1.422e-01  -1.908 0.056391 .
## JOBLawyer                      -4.434e-01  1.799e-01  -2.465 0.013697 *
## JOBManager                     -1.078e+00  1.399e-01  -7.709 1.27e-14 ***
## JOBProfessional                -3.875e-01  1.182e-01  -3.279 0.001041 **
## JOBStudent                     -2.387e-01  1.315e-01  -1.816 0.069435 .
## JOBUnknown                     -4.847e-01  1.962e-01  -2.471 0.013477 *
## JOBz_Blue Collar               -1.027e-01  1.066e-01  -0.963 0.335505
## TRAVTIME                        1.442e-02  1.879e-03   7.675 1.65e-14 ***
## CAR_USEPrivate                 -7.352e-01  8.697e-02  -8.454  < 2e-16 ***
## BLUEBOOK                       -2.090e-05  5.272e-06  -3.964 7.37e-05 ***
## TIF                            -5.536e-02  7.338e-03  -7.545 4.54e-14 ***
## CAR_TYPEPanel Truck             5.821e-01  1.602e-01   3.634 0.000279 ***
## CAR_TYPEPickup                  5.696e-01  9.982e-02   5.706 1.16e-08 ***
## CAR_TYPESports Car              1.014e+00  1.297e-01   7.817 5.42e-15 ***
## CAR_TYPEVan                     6.238e-01  1.258e-01   4.957 7.14e-07 ***
## CAR_TYPEz_SUV                   7.597e-01  1.111e-01   6.840 7.90e-12 ***
## RED_CARyes                     -1.721e-02  8.631e-02  -0.199 0.841988
## OLDCLAIM                       -1.411e-05  3.907e-06  -3.611 0.000305 ***
## CLM_FREQ                        1.955e-01  2.851e-02   6.858 6.97e-12 ***
## REVOKEDYes                      8.932e-01  9.125e-02   9.788  < 2e-16 ***
## MVR_PTS                         1.133e-01  1.361e-02   8.324  < 2e-16 ***
## CAR_AGE                        -1.985e-02  6.885e-03  -2.882 0.003947 **
## URBANICITYz_Highly Rural/ Rural -2.375e+00  1.124e-01 -21.125  < 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: 7306.3  on 8126  degrees of freedom
## AIC: 7376.3
##
## Number of Fisher Scoring iterations: 5
```

# Binary Model 4 - Stepwise Step Model

We also can use the same stepAIC function to build the fourth model using stepwise regression. The stepwise regression method involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration. This is exactly same result as the backward step model.

```
##
## Call:
## glm(formula = factor(TARGET_FLAG) ~ KIDSDRIV + HOMEKIDS + YOJ +
##     INCOME + PARENT1 + HOME_VAL + MSTATUS + JOB + TRAVTIME +
##     CAR_USE + BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY, family = binomial,
##     data = subset(complete_train_data, select = -c(TARGET_AMT)))
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -5.931e-01  1.990e-01  -2.981 0.002877 **
## KIDSDRIV                       3.577e-01  5.865e-02   6.098 1.07e-09 ***
## HOMEKIDS                       2.636e-01  8.737e-02   3.016 0.002558 **
## YOJ                           -1.343e-02  8.216e-03  -1.635 0.102114
## INCOME                        -3.850e-06  1.080e-06  -3.563 0.000366 ***
## PARENT1Yes                     2.324e-01  1.206e-01   1.928 0.053898 .
## HOME_VAL                      -1.270e-06  3.395e-07  -3.740 0.000184 ***
## MSTATUSz_No                    5.408e-01  8.633e-02   6.264 3.76e-10 ***
## JOBDoctor                     -7.308e-01  2.438e-01  -2.997 0.002727 **
## JOBHome Maker                 -2.540e-01  1.384e-01  -1.836 0.066395 .
## JOBLawyer                     -3.288e-01  1.431e-01  -2.297 0.021605 *
## JOBManager                    -1.035e+00  1.342e-01  -7.714 1.22e-14 ***
## JOBProfessional               -3.734e-01  1.175e-01  -3.178 0.001481 **
## JOBStudent                    -2.339e-01  1.312e-01  -1.782 0.074677 .
## JOBUnknown                    -3.789e-01  1.652e-01  -2.293 0.021837 *
## JOBz_Blue Collar              -1.021e-01  1.065e-01  -0.959 0.337387
## TRAVTIME                       1.447e-02  1.878e-03   7.701 1.35e-14 ***
## CAR_USEPrivate                -7.338e-01  8.688e-02  -8.446  < 2e-16 ***
## BLUEBOOK                      -2.245e-05  4.735e-06  -4.741 2.13e-06 ***
## TIF                           -5.525e-02  7.336e-03  -7.532 5.01e-14 ***
## CAR_TYPEPanel Truck            6.291e-01  1.493e-01   4.214 2.51e-05 ***
## CAR_TYPEPickup                 5.693e-01  9.972e-02   5.709 1.13e-08 ***
## CAR_TYPESports Car             9.615e-01  1.074e-01   8.953  < 2e-16 ***
## CAR_TYPEVan                    6.490e-01  1.215e-01   5.342 9.21e-08 ***
## CAR_TYPEz_SUV                  7.050e-01  8.587e-02   8.210  < 2e-16 ***
## OLDCLAIM                      -1.408e-05  3.908e-06  -3.603 0.000315 ***
## CLM_FREQ                       1.959e-01  2.849e-02   6.874 6.23e-12 ***
## REVOKEDYes                     8.932e-01  9.120e-02   9.794  < 2e-16 ***
## MVR_PTS                        1.128e-01  1.360e-02   8.298  < 2e-16 ***
## CAR_AGE                       -1.770e-02  6.497e-03  -2.724 0.006443 **
## URBANICITYz_Highly Rural/ Rural -2.375e+00  1.124e-01 -21.127  < 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: 7308.3  on 8130  degrees of freedom
## AIC: 7370.3
##
## Number of Fisher Scoring iterations: 5
```

Judging by AIC, the stepwise approach reduces the dimensionality and improves fit, given its lower estimated prediction error. This suggests that, in addition to being a simple model, the stepwise method works better to create an overall better fit to the data.

The analysis of deviance table shows further confirms that dropping these statistical insignificant variables {*AGE*, *SEX*, *EDUCATION*, *RED_CAR*} in model 4.

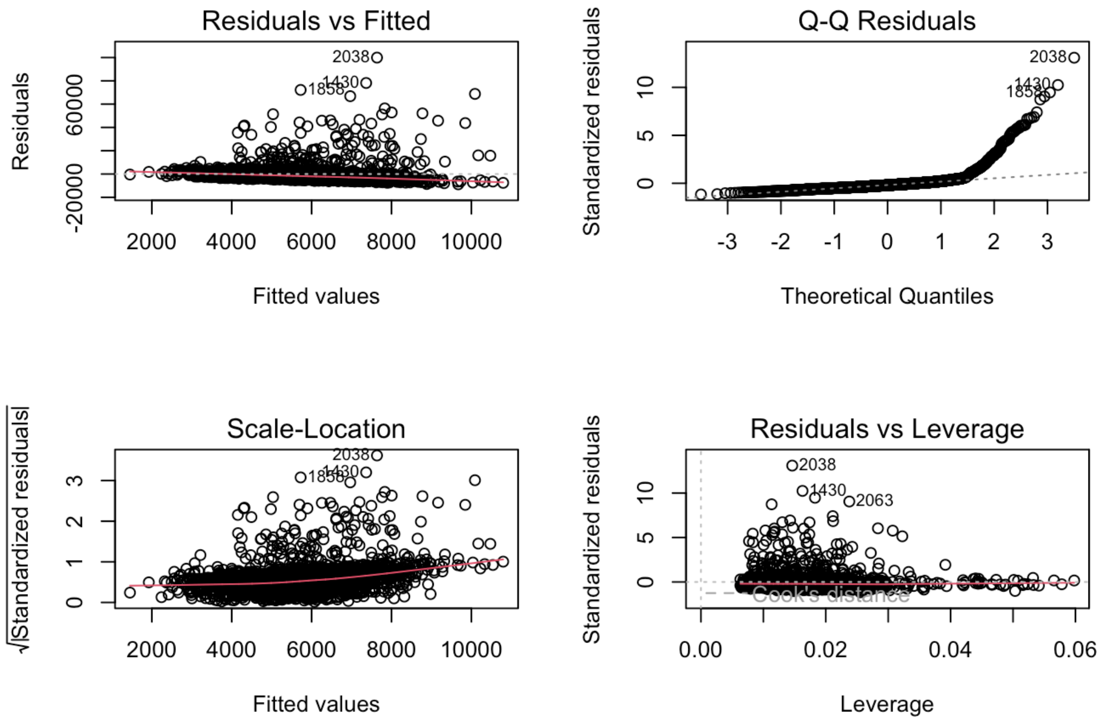
```
## Analysis of Deviance Table
##
## Model 1: factor(TARGET_FLAG) ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
## Model 2: factor(TARGET_FLAG) ~ 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
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      8130     7308.3
## 2      8126     7306.3  4   2.0169   0.7327
```

We next ran some multivariate regression models, using the features to predict the numeric response variable, TARGET_AMT, which gives the costs associated with a car's accident, creating a multiple linear regression model to predict the response variable.

# Multi Linear Regression Model 1 - All Variables

For the multi linear regression we want to know what is going to be the insurance cost if a person has crashed their car. We are going to build a multi linear regression model which includes all the data, from there we will keep the variables that have significance and use that to build subsequence models. We will first need to create a dataset specifically for a multi linear regression as we only care about if a customer has crashed their car.

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = mlr_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8832  -3165  -1507    441  99949
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.537e+03  1.761e+03   2.576   0.0101 *
## KIDSDRIV                       -1.488e+02  3.081e+02  -0.483   0.6292
## AGE                             2.060e+01  2.200e+01   0.936   0.3493
## HOMEKIDS                        5.980e+02  5.729e+02   1.044   0.2967
## YOJ                             1.844e+01  4.821e+01   0.383   0.7021
## INCOME                         -8.726e-03  6.632e-03  -1.316   0.1884
## PARENT1Yes                      1.043e+02  6.713e+02   0.155   0.8765
## HOME_VAL                        1.856e-03  2.030e-03   0.914   0.3608
## MSTATUSz_No                     8.865e+02  5.187e+02   1.709   0.0876 .
## SEXz_F                         -1.431e+03  6.569e+02  -2.178   0.0295 *
## EDUCATION                      -1.357e+03  8.794e+02  -1.543   0.1229
## JOBDoctor                      -1.540e+03  1.772e+03  -0.869   0.3851
## JOBHome Maker                  -2.595e+02  8.274e+02  -0.314   0.7538
## JOBLawyer                      -1.506e+02  1.130e+03  -0.133   0.8939
## JOBManager                     -1.015e+03  9.129e+02  -1.112   0.2664
## JOBProfessional                 8.676e+02  6.846e+02   1.267   0.2052
## JOBStudent                     -2.117e+02  7.340e+02  -0.288   0.7731
## JOBUnknown                     -1.382e+02  1.204e+03  -0.115   0.9087
## JOBz_Blue Collar                2.287e+02  5.875e+02   0.389   0.6971
## TRAVTIME                        1.357e+00  1.109e+01   0.122   0.9026
## CAR_USEPrivate                 -3.710e+02  4.978e+02  -0.745   0.4562
## BLUEBOOK                        1.285e-01  3.057e-02   4.205 2.72e-05 ***
## TIF                            -1.750e+01  4.257e+01  -0.411   0.6811
## CAR_TYPEPanel Truck            -6.553e+02  9.549e+02  -0.686   0.4926
## CAR_TYPEPickup                 -6.263e+01  5.929e+02  -0.106   0.9159
## CAR_TYPESports Car              1.051e+03  7.493e+02   1.403   0.1608
## CAR_TYPEVan                     6.086e+01  7.681e+02   0.079   0.9369
## CAR_TYPEz_SUV                   8.737e+02  6.663e+02   1.311   0.1899
## RED_CARyes                     -1.886e+02  4.964e+02  -0.380   0.7040
## OLDCLAIM                        2.403e-02  2.261e-02   1.063   0.2880
## CLM_FREQ                       -1.179e+02  1.580e+02  -0.746   0.4558
## REVOKEDYes                     -1.123e+03  5.162e+02  -2.176   0.0296 *
## MVR_PTS                         1.152e+02  6.841e+01   1.684   0.0923 .
## CAR_AGE                        -7.005e+01  4.021e+01  -1.742   0.0817 .
## URBANICITYz_Highly Rural/ Rural -1.026e+02  7.566e+02  -0.136   0.8922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7692 on 2118 degrees of freedom
## Multiple R-squared:  0.02875,    Adjusted R-squared:  0.01316
## F-statistic: 1.844 on 34 and 2118 DF,  p-value: 0.002191
```

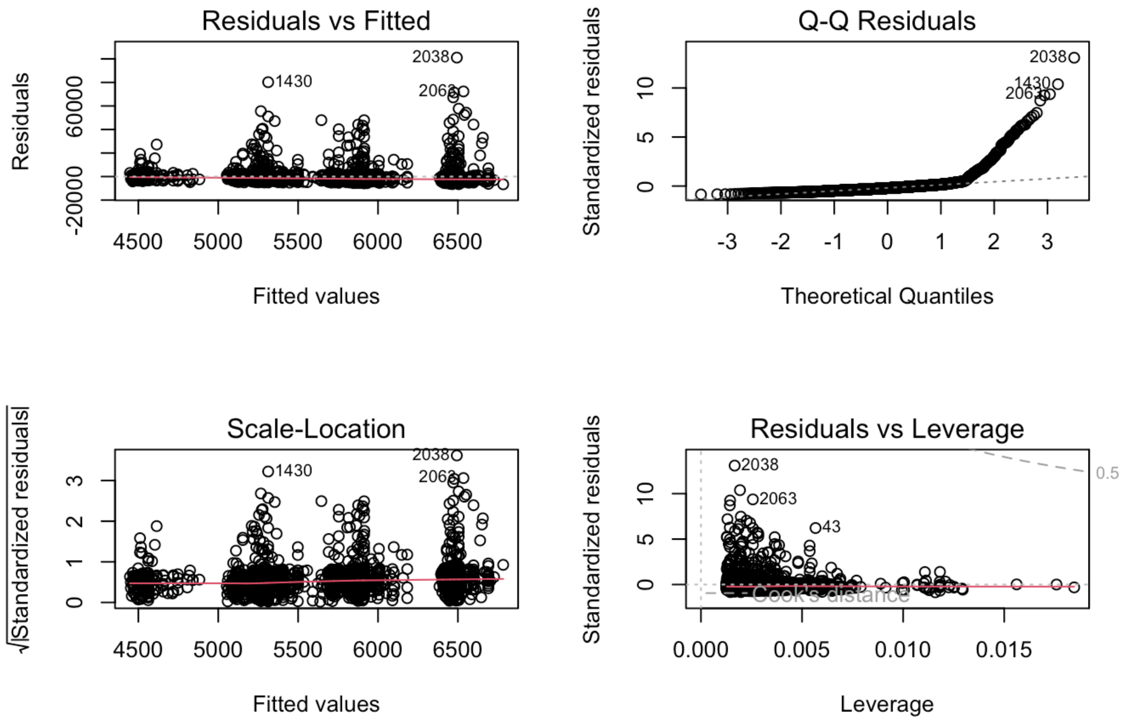



Looking at the plot above we can see that the Residual vs Fitted graph has a large variance for a couple of outliers while a majority of the points have very low residuals. Also looking at the Normal Q-Q graph we can see that it is not a normal distribution. This is quite good for a model which utilizes all of the variables and we would like to see if we can improve the model by selecting variables that are significant.

Some variables that we would like to use for the next model are **KIDSDRIV**, **SEX**, **CAR_USE**, **REVOKED**, and **CAR_AGE**. These variables makes a lot of are usually thought of as the variables which can increase the cost of insurance

# Multi Linear Regression Model 2 - Hand Picking Variables

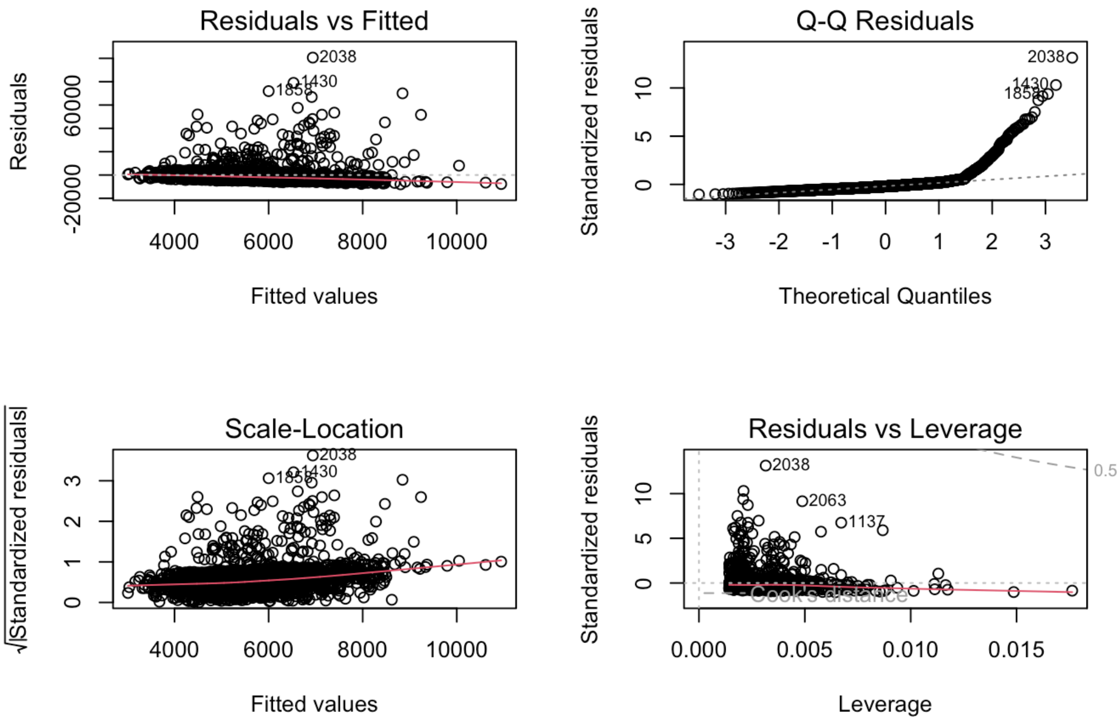
```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + SEX + CAR_USE + REVOKED +
##     CAR_AGE, data = mlr_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6587  -3072  -1614    171 101093
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6510.524    360.192  18.075   <2e-16 ***
## KIDSDRIV         92.865    267.381   0.347   0.7284
## SEXz_F         -591.894    358.393  -1.652   0.0988 .
## CAR_USEPrivate -600.417    357.914  -1.678   0.0936 .
## REVOKEDYes     -750.528    413.747  -1.814   0.0698 .
## CAR_AGE          -5.725     30.312  -0.189   0.8502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7731 on 2147 degrees of freedom
## Multiple R-squared:  0.005369,   Adjusted R-squared:  0.003053
## F-statistic: 2.318 on 5 and 2147 DF,  p-value: 0.04124
```




With the second model we can see that we have a lower R Squared value in our new model compared to the first model which included all the variables. We can also see that there is still a large variance with the Residual vs Fitted plot. We will next try to use a stepwise function to find the best model from all the variables.

# Multi Linear Regression Model 3 - Stepwise Function

```
## 
## Call:
## lm(formula = TARGET_AMT ~ MSTATUS + SEX + BLUEBOOK + REVOKED + 
##     MVR_PTS, data = mlr_data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -7964  -3154  -1542    359 100647 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4108.50301  459.26911   8.946  < 2e-16 ***
## MSTATUSz_No  513.91766  331.20096   1.552   0.1209    
## SEXz_F      -661.58714  333.93883  -1.981   0.0477 *  
## BLUEBOOK       0.10689    0.02002   5.339 1.03e-07 ***
## REVOKEDYes  -697.99672  409.40606  -1.705   0.0884 .  
## MVR_PTS      127.80337   64.17872   1.991   0.0466 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7675 on 2147 degrees of freedom
## Multiple R-squared:  0.01992,    Adjusted R-squared:  0.01764 
## F-statistic: 8.728 on 5 and 2147 DF,  p-value: 3.314e-08
```



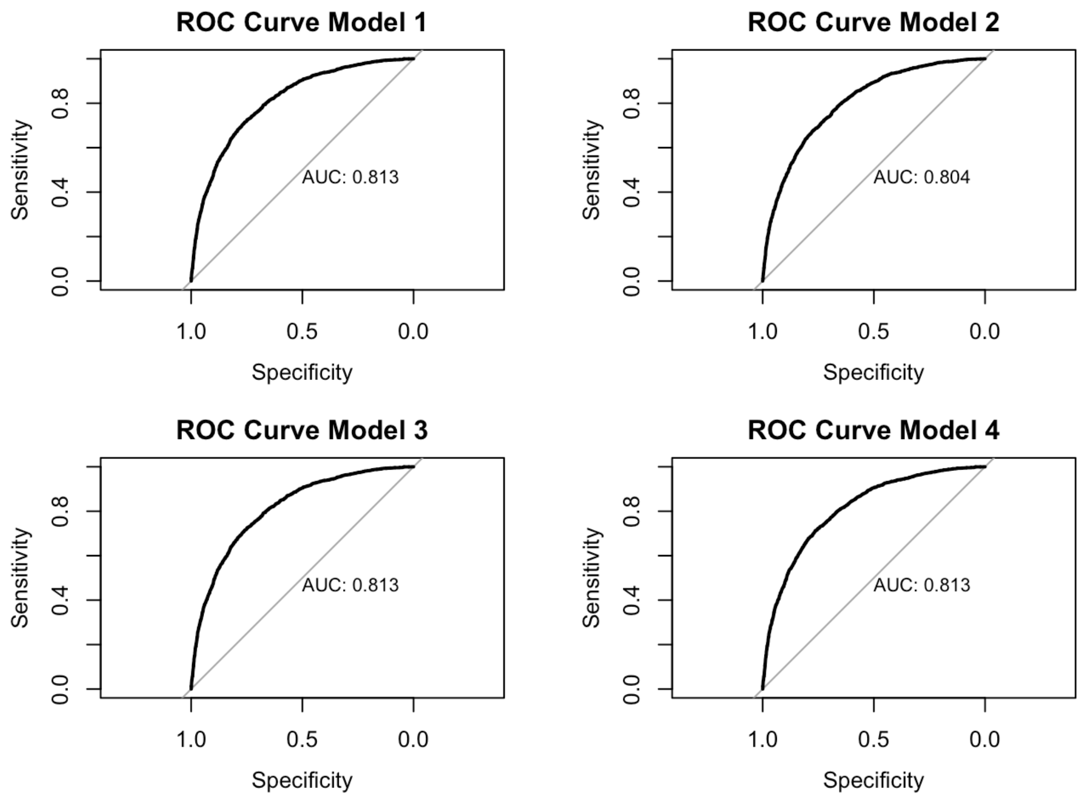

Using the backward step function we can see that the model choose **MSTATUS**, **SEX**, **BLUEBOOK**, **REVOKED**, and **MVR_PTS**. Looking at the residual vs fitted plot we still see a variance caused by outliers.

# Model Selection

## Logistic Regression

we will compare various metrics for all 4 models. We check models' confusion matrix, accuracy, classification error rate, precision, sensitivity, specificity, F1 score, AUC, and ROC curves.

First, let's plot the ROC curves for all 4 models and then calculate the various metrics.




| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Accuracy | 0.7905894 | 0.7856880 | 0.7905894 | 0.7908345 |
| Class. Error Rate | 0.2094106 | 0.2143120 | 0.2094106 | 0.2091655 |
| Sensitivity | 0.4217371 | 0.3901533 | 0.4217371 | 0.4217371 |
| Specificity | 0.9227696 | 0.9274301 | 0.9227696 | 0.9231025 |
| Precision | 0.6618076 | 0.6583072 | 0.6618076 | 0.6627737 |
| F1 | 0.5151773 | 0.4899388 | 0.5151773 | 0.5154698 |
| AUC | 0.8129532 | 0.8040969 | 0.8129532 | 0.8128175 |

By looking at the ROC curves, model 1, 3, and 4 are showing the same area under curve value. So, it’s hard to justify which model is the best. Fortunately, we have the various calculated metrics to provide us more details which model is the best. Based on that, we can say that the model 4 performs the highest in all metrics except Class. Error Rate.

# Multi Linear Regression

We will be looking at all the models created and looking at the metrics like R Squared Value, RMSE, F-Statistics, and Residual Plots in order to determine which is the best model which represents our data. We will then compare the best model selected against the evaluation data set in order to see if the model truly represents the dataset

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| R Square | 0.0287517 | 0.0053690 | 0.0199215 |
| F Stat | 1.8440845 | 2.3178745 | 8.7281624 |
| R Adj Square | 0.0131604 | 0.0030526 | 0.0176390 |

Based on the table above the the best model would be model 3 based on the summary and the residual vs fitted plot

# Predictions

# Predictions

Logistics Model:

```
prediction_binary = predict(m4b, complete_eval_data, type="response")
complete_eval_data$TARGET_FLAG = prediction_binary
complete_eval_data$TARGET_FLAG <- ifelse(complete_eval_data$TARGET_FLAG > 0.5, 1, 0)
print(head(complete_eval_data,10))
```

```
##    TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1 HOME_VAL
## 1            0         NA        0  48        0  11  52881      No        0
## 2            0         NA        1  40        1  11  50815     Yes        0
## 3            0         NA        0  44        1  12  43486     Yes        0
## 4            0         NA        0  35        1  10  21204     Yes        0
## 5            0         NA        0  59        0  12  87460      No        0
## 6            0         NA        0  46        0  14  40764      No   207519
## 7            0         NA        0  60        0  12  37940      No   182739
## 8            0         NA        0  54        0  12  33212      No   158432
## 9            0         NA        2  36        1  12 130540     Yes   344195
## 10           0         NA        0  50        0   8 167469      No        0
##    MSTATUS SEX EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK TIF
## 1     z_No   M         1       Manager       26    Private    21970   1
## 2     z_No   M         1       Manager       21    Private    18930   6
## 3     z_No z_F         1 z_Blue Collar       30 Commercial     5900  10
## 4     z_No   M         1      Clerical       74    Private     9230   6
## 5     z_No   M         1       Manager       45    Private    15420   1
## 6      Yes   M         1  Professional        7 Commercial    25660   1
## 7      Yes z_F         1 z_Blue Collar       16 Commercial    11290   1
## 8      Yes   M         1 z_Blue Collar       27 Commercial    24000   4
## 9     z_No z_F         1 z_Blue Collar        5 Commercial    27200   4
## 10    z_No z_F         0        Doctor       22    Private    34150   4
##       CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1          Van     yes        0        0      No       2      10
## 2      Minivan      no     3295        1      No       2       1
## 3        z_SUV      no        0        0      No       0      10
## 4       Pickup      no        0        0     Yes       0       4
## 5      Minivan     yes    44857        2      No       4       1
## 6  Panel Truck      no     2119        1      No       2      12
## 7   Sports Car      no        0        0      No       0       1
## 8  Panel Truck      no        0        0      No       5      12
## 9      Minivan      no        0        0      No       0       9
## 10  Sports Car      no        0        0      No       3       1
##             URBANICITY
## 1    Highly Urban/ Urban
## 2    Highly Urban/ Urban
## 3  z_Highly Rural/ Rural
## 4  z_Highly Rural/ Rural
## 5    Highly Urban/ Urban
## 6    Highly Urban/ Urban
## 7    Highly Urban/ Urban
## 8    Highly Urban/ Urban
## 9  z_Highly Rural/ Rural
## 10   Highly Urban/ Urban
```

Multi Linear Model:

```
prediction_linear = predict(mlr3, complete_eval_data)
complete_eval_data$TARGET_AMT = ifelse(complete_eval_data$TARGET_FLAG ==1, prediction_linear, 0)
print(head(complete_eval_data,10))
```

```
##    TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1 HOME_VAL
## 1            0          0        0  48        0  11  52881      No        0
## 2            0          0        1  40        1  11  50815     Yes        0
## 3            0          0        0  44        1  12  43486     Yes        0
## 4            0          0        0  35        1  10  21204     Yes        0
## 5            0          0        0  59        0  12  87460      No        0
## 6            0          0        0  46        0  14  40764      No   207519
## 7            0          0        0  60        0  12  37940      No   182739
## 8            0          0        0  54        0  12  33212      No   158432
## 9            0          0        2  36        1  12 130540     Yes   344195
## 10           0          0        0  50        0   8 167469      No        0
##    MSTATUS SEX EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK TIF
## 1     z_No   M         1       Manager       26    Private    21970   1
## 2     z_No   M         1       Manager       21    Private    18930   6
## 3     z_No z_F         1 z_Blue Collar       30 Commercial     5900  10
## 4     z_No   M         1      Clerical       74    Private     9230   6
## 5     z_No   M         1       Manager       45    Private    15420   1
## 6      Yes   M         1  Professional        7 Commercial    25660   1
## 7      Yes z_F         1 z_Blue Collar       16 Commercial    11290   1
## 8      Yes   M         1 z_Blue Collar       27 Commercial    24000   4
## 9     z_No z_F         1 z_Blue Collar        5 Commercial    27200   4
## 10    z_No z_F         0        Doctor       22    Private    34150   4
##       CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1          Van     yes        0        0      No       2      10
## 2      Minivan      no     3295        1      No       2       1
## 3        z_SUV      no        0        0      No       0      10
## 4       Pickup      no        0        0     Yes       0       4
## 5      Minivan     yes    44857        2      No       4       1
## 6  Panel Truck      no     2119        1      No       2      12
## 7   Sports Car      no        0        0      No       0       1
## 8  Panel Truck      no        0        0      No       5      12
## 9      Minivan      no        0        0      No       0       9
## 10  Sports Car      no        0        0      No       3       1
##              URBANICITY
## 1    Highly Urban/ Urban
## 2    Highly Urban/ Urban
## 3  z_Highly Rural/ Rural
## 4  z_Highly Rural/ Rural
## 5    Highly Urban/ Urban
## 6    Highly Urban/ Urban
## 7    Highly Urban/ Urban
## 8    Highly Urban/ Urban
## 9  z_Highly Rural/ Rural
## 10   Highly Urban/ Urban
```

```
write.csv(complete_eval_data, 'predictions.csv')
```

# Appendix

```
Train_Data <- read.csv("https://raw.githubusercontent.com/ahussan/DATA_621_Group1/main/HW4/insurance_trai
ning_data.csv")
Eval_Data <- read.csv("https://raw.githubusercontent.com/ahussan/DATA_621_Group1/main/HW4/insurance-evalu
ation-data.csv")
Train_Data <- Train_Data[,-1]
str(Train_Data)
summary(Train_Data)
Train_Data$INCOME <- gsub(",","",(Train_Data$INCOME))
Train_Data$INCOME <- sub('.', '', Train_Data$INCOME)
Train_Data$INCOME <-trimws(Train_Data$INCOME, which = c("both"), whitespace = "[ \t\r\n]")
Train_Data$INCOME <- as.numeric(Train_Data$INCOME)
Train_Data$HOME_VAL <- gsub(",","",(Train_Data$HOME_VAL))
Train_Data$HOME_VAL <- sub('.', '', Train_Data$HOME_VAL)
Train_Data$HOME_VAL <-trimws(Train_Data$HOME_VAL, which = c("both"), whitespace = "[ \t\r\n]")
Train_Data$HOME_VAL <- as.numeric(Train_Data$HOME_VAL)
#
Train_Data$BLUEBOOK <- gsub(",","",(Train_Data$BLUEBOOK))
Train_Data$BLUEBOOK <- sub('.', '', Train_Data$BLUEBOOK)
Train_Data$BLUEBOOK <-trimws(Train_Data$BLUEBOOK, which = c("both"), whitespace = "[ \t\r\n]")
Train_Data$BLUEBOOK <- as.numeric(Train_Data$BLUEBOOK)
#
Train_Data$OLDCLAIM <- gsub(",","",(Train_Data$OLDCLAIM))
Train_Data$OLDCLAIM <- sub('.', '', Train_Data$OLDCLAIM)
Train_Data$OLDCLAIM <-trimws(Train_Data$OLDCLAIM, which = c("both"), whitespace = "[ \t\r\n]")
Train_Data$OLDCLAIM <- as.numeric(Train_Data$OLDCLAIM)
#
summary(Train_Data[,c(7,9,16,20)])
colSums(is.na(Train_Data))
Train_Data1<-dplyr::select_if(Train_Data, is.numeric)
Train_Data1 %>%
  keep(is.numeric) %>%
  tidyr::gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density()
Train_Data$HOMEKIDS[Train_Data$HOMEKIDS != 0 ] <- 1
Eval_Data$HOMEKIDS[Eval_Data$HOMEKIDS != 0 ] <- 1
Train_Data$CAR_AGE[Train_Data$AGE < 0 ] <- 0
Eval_Data$CAR_AGE[Eval_Data$AGE < 0 ] <- 0
Train_Data$CAR_AGE[Train_Data$CAR_AGE < 0 ] <- 0
Eval_Data$CAR_AGE[Eval_Data$CAR_AGE < 0 ] <- 0
Train_Data$JOB <- as.character(Train_Data$JOB)
Train_Data$JOB[Train_Data$JOB == ""] <- "Unknown"
Train_Data$JOB <- as.factor(Train_Data$JOB)
Eval_Data$JOB <- as.character(Eval_Data$JOB)
Eval_Data$JOB[Eval_Data$JOB == ""] <- "Unknown"
Eval_Data$JOB <- as.factor(Eval_Data$JOB)
summary(Train_Data)
impute_train_data <- mice(Train_Data, m=5, maxit=20, method='pmm', seed=321, print = FALSE)
densityplot(impute_train_data)
complete_train_data <- complete(impute_train_data,2)
impute_eval_data <- mice(Eval_Data, m=5, maxit=20, method='pmm', seed=321, print = FALSE)
densityplot(impute_eval_data)
complete_eval_data <- complete(impute_eval_data,2)
m1b <- glm(factor(TARGET_FLAG) ~ ., family=binomial, data=subset(complete_train_data, select=-c(TARGET_AM
T)))
summary(m1b)
m2b <- glm(factor(TARGET_FLAG) ~ KIDSDRIV+HOMEKIDS+log(INCOME+1)+log(HOME_VAL+1)+MSTATUS+log(TRAVTIME)+CA
R_USE+log(BLUEBOOK)+TIF    +CAR_TYPE+log(OLDCLAIM+1)+CLM_FREQ+REVOKED+MVR_PTS +log(CAR_AGE+1)+URBANICITY,
family=binomial, data=subset(complete_train_data, select=-c(TARGET_AMT)))
summary(m2b)
```

```
m3b <- stepAIC(m1b, direction='forward', trace=FALSE)
summary(m3b)
m4b <- stepAIC(m1b, direction='both', trace=FALSE)
summary(m4b)
anova(m4b,m1b, test="Chi")
mlr_data = complete_train_data %>%
  filter(TARGET_FLAG == 1) %>%
  dplyr::select(-1)
mlr1 = lm(TARGET_AMT ~ ., data = mlr_data)
summary(mlr1)
par(mfrow=c(2,2))
plot(mlr1)
mlr2 = lm(TARGET_AMT ~ KIDSDRIV + SEX + CAR_USE + REVOKED + CAR_AGE, data=mlr_data)
summary(mlr2)
par(mfrow=c(2,2))
plot(mlr2)
mlr3 = stepAIC(mlr1, direction='backward', trace=FALSE)
summary(mlr3)
par(mfrow=c(2,2))
plot(mlr3)
par(mfrow=c(2,2))
plot(roc(complete_train_data$TARGET_FLAG,  predict(m1b, complete_train_data, interval = "prediction")), p
rint.auc = TRUE, main='ROC Curve Model 1')
plot(roc(complete_train_data$TARGET_FLAG,  predict(m2b, complete_train_data, interval = "prediction")), p
rint.auc = TRUE, main='ROC Curve Model 2')
plot(roc(complete_train_data$TARGET_FLAG,  predict(m3b, complete_train_data, interval = "prediction")), p
rint.auc = TRUE, main='ROC Curve Model 3')
plot(roc(complete_train_data$TARGET_FLAG,  predict(m4b, complete_train_data, interval = "prediction")), p
rint.auc = TRUE, main='ROC Curve Model 4')
CM1 <- confusionMatrix(as.factor(as.integer(fitted(m1b) > .5)), as.factor(m1b$y), positive = "1")
CM2 <- confusionMatrix(as.factor(as.integer(fitted(m2b) > .5)), as.factor(m2b$y), positive = "1")
CM3 <- confusionMatrix(as.factor(as.integer(fitted(m3b) > .5)), as.factor(m3b$y), positive = "1")
CM4 <- confusionMatrix(as.factor(as.integer(fitted(m4b) > .5)), as.factor(m4b$y), positive = "1")
Roc1 <- roc(complete_train_data$TARGET_FLAG,  predict(m1b, complete_train_data, interval = "prediction"))
Roc2 <- roc(complete_train_data$TARGET_FLAG,  predict(m2b, complete_train_data, interval = "prediction"))
Roc3 <- roc(complete_train_data$TARGET_FLAG,  predict(m3b, complete_train_data, interval = "prediction"))
Roc4 <- roc(complete_train_data$TARGET_FLAG,  predict(m4b, complete_train_data, interval = "prediction"))
metrics1 <- c(CM1$overall[1], "Class. Error Rate" = 1 - as.numeric(CM1$overall[1]), CM1$byClass[c(1, 2,
5, 7)], AUC = Roc1$auc)
metrics2 <- c(CM2$overall[1], "Class. Error Rate" = 1 - as.numeric(CM2$overall[1]), CM2$byClass[c(1, 2,
5, 7)], AUC = Roc2$auc)
metrics3 <- c(CM3$overall[1], "Class. Error Rate" = 1 - as.numeric(CM3$overall[1]), CM3$byClass[c(1, 2,
5, 7)], AUC = Roc3$auc)
metrics4 <- c(CM4$overall[1], "Class. Error Rate" = 1 - as.numeric(CM4$overall[1]), CM4$byClass[c(1, 2,
5, 7)], AUC = Roc4$auc)
kable(cbind(metrics1, metrics2, metrics3, metrics4), col.names = c("Model 1", "Model 2", "Model 3", "Mode
l 4"))  %>%
  kable_styling(full_width = T)
m1.summary = summary(mlr1)
m2.summary = summary(mlr2)
m3.summary = summary(mlr3)
m1.square = m1.summary$r.squared
m2.square = m2.summary$r.squared
m3.square = m3.summary$r.squared
m1.fstat = as.numeric(m1.summary$fstatistic[1])
m2.fstat = as.numeric(m2.summary$fstatistic[1])
m3.fstat = as.numeric(m3.summary$fstatistic[1])
m1.r = m1.summary$adj.r.squared
m2.r = m2.summary$adj.r.squared
m3.r = m3.summary$adj.r.squared
metrics1 = c('R Square'=m1.square, 'F Stat'=m1.fstat, 'R Adj Square'=m1.r)
metrics2 = c('R Square'=m2.square, 'F Stat'=m2.fstat, 'R Square'=m2.r)
```

```
metrics3 = c('R Square'=m3.square, 'F Stat'=m3.fstat, 'R Square'=m3.r)
kable(cbind(metrics1, metrics2, metrics3), col.names = c("Model 1", "Model 2", "Model 3"))  %>%
  kable_styling(full_width = T)
prediction_binary = predict(m4b, complete_eval_data, type="response")
complete_eval_data$TARGET_FLAG = prediction_binary
complete_eval_data$TARGET_FLAG <- ifelse(complete_eval_data$TARGET_FLAG > 0.5, 1, 0)
print(head(complete_eval_data,10))
prediction_linear = predict(mlr3, complete_eval_data)
complete_eval_data$TARGET_AMT = ifelse(complete_eval_data$TARGET_FLAG ==1, prediction_linear, 0)
print(head(complete_eval_data,10))
write.csv(complete_eval_data, 'predictions.csv')
```