# Car Crash Prediction

*Shyam BV*

*April 8, 2018*

## Contents

| VARIABLE NAME | DEFINITION | THEORETICAL EFFECT |
|---|---|---|
| INDEX | Identification Variable (do not use) | None |
| TARGET_FLAG | Was Car in a crash? 1=YES 0=NO | None |
| TARGET_AMT | If car was in a crash, what was the cost | None |
| AGE | Age of Driver | Very young people tend to be risky. Maybe very old people also. |
| BLUEBOOK | Value of Vehicle | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR_AGE | Vehicle Age | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR_TYPE | Type of Car | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR_USE | Vehicle Use | Commercial vehicles are driven more, so might increase probability of collision |
| CLM_FREQ | # Claims (Past 5 Years) | The more claims you filed in the past, the more you are likely to file in the future |
| EDUCATION | Max Education Level | Unknown effect, but in theory more educated people tend to drive more safely |
| HOMEKIDS | # Children at Home | Unknown effect |
| HOME_VAL | Home Value | In theory, home owners tend to drive more responsibly |
| INCOME | Income | In theory, rich people tend to get into fewer crashes |
| JOB | Job Category | In theory, white collar jobs tend to be safer |
| KIDSDRIV | # Driving Children | When teenagers drive your car, you are more likely to get into crashes |
| MSTATUS | Marital Status | In theory, married people drive more safely |
| MVR_PTS | Motor Vehicle Record Points | If you get lots of traffic tickets, you tend to get into more crashes |
| OLDCLAIM | Total Claims (Past 5 Years) | If your total payout over the past five years was high, this suggests future payouts will be high |
| PARENT1 | Single Parent | Unknown effect |
| RED_CAR | A Red Car | Urban legend says that red cars (especially red sports cars) are more risky. Is that true? |
| REVOKED | License Revoked (Past 7 Years) | If your license was revoked in the past 7 years, you probably are a more risky driver. |
| SEX | Gender | Urban legend says that women have less crashes then men. Is that true? |
| TIF | Time in Force | 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 | Unknown |
| YOJ | Years on Job | People who stay at a job for a long time are usually more safe |

Figure 1: Data Definition.

# 1 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.

---

Deliverables:

1. A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
2. Assigned predictions (probabilities, classifications, cost) for the evaluation data set. Use 0.5 threshold.
3. Include your R statistical programming code in an Appendix.

## 1.1 Data Exploration

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

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data and/or Histograms
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

Below is the summary of the dataset and a quick view of the dataset.

```
##      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   $0      : 615
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0           : 445
##  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
##  PARENT1        HOME_VAL    MSTATUS      SEX             EDUCATION
##  No :7084   $0      :2294   Yes :4894   M  :3786   <High School :1203
##  Yes:1077           : 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
##  Min.   : 1.000   Minivan    :2145   no :5783   $0     :5009
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :   4
##  Median : 4.000   Pickup     :1389              $1,391 :   4
##  Mean   : 5.351   Sports Car : 907              $4,263 :   4
##  3rd Qu.: 7.000   Van        : 750              $1,105 :   3
##  Max.   :25.000   z_SUV      :2294              $1,332 :   3
##                                                 (Other):3134
##     CLM_FREQ      REVOKED       MVR_PTS         CAR_AGE
##  Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              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
##  Highly Urban/ Urban  :6492
##  z_Highly Rural/ Rural:1669
##
##
##
##
##
```

```
##   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
## 1       $0    z_No   M           PhD  Professional       14    Private
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial
##   BLUEBOOK TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS
## 1  $14,230  11    Minivan     yes   $4,461        2      No       3
## 2  $14,940   1    Minivan     yes       $0        0      No       0
## 3   $4,010   4      z_SUV      no  $38,690        2      No       3
## 4  $15,440   7    Minivan     yes       $0        0      No       0
## 5  $18,000   1      z_SUV      no  $19,217        2     Yes       3
## 6  $17,430   1 Sports Car      no       $0        0      No       0
##   CAR_AGE          URBANICITY
## 1      18 Highly Urban/ Urban
## 2       1 Highly Urban/ Urban
## 3      10 Highly Urban/ Urban
## 4       6 Highly Urban/ Urban
## 5      17 Highly Urban/ Urban
## 6       7 Highly Urban/ Urban
```

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

```
## CAR_AGE          510
## URBANICITY         0
```

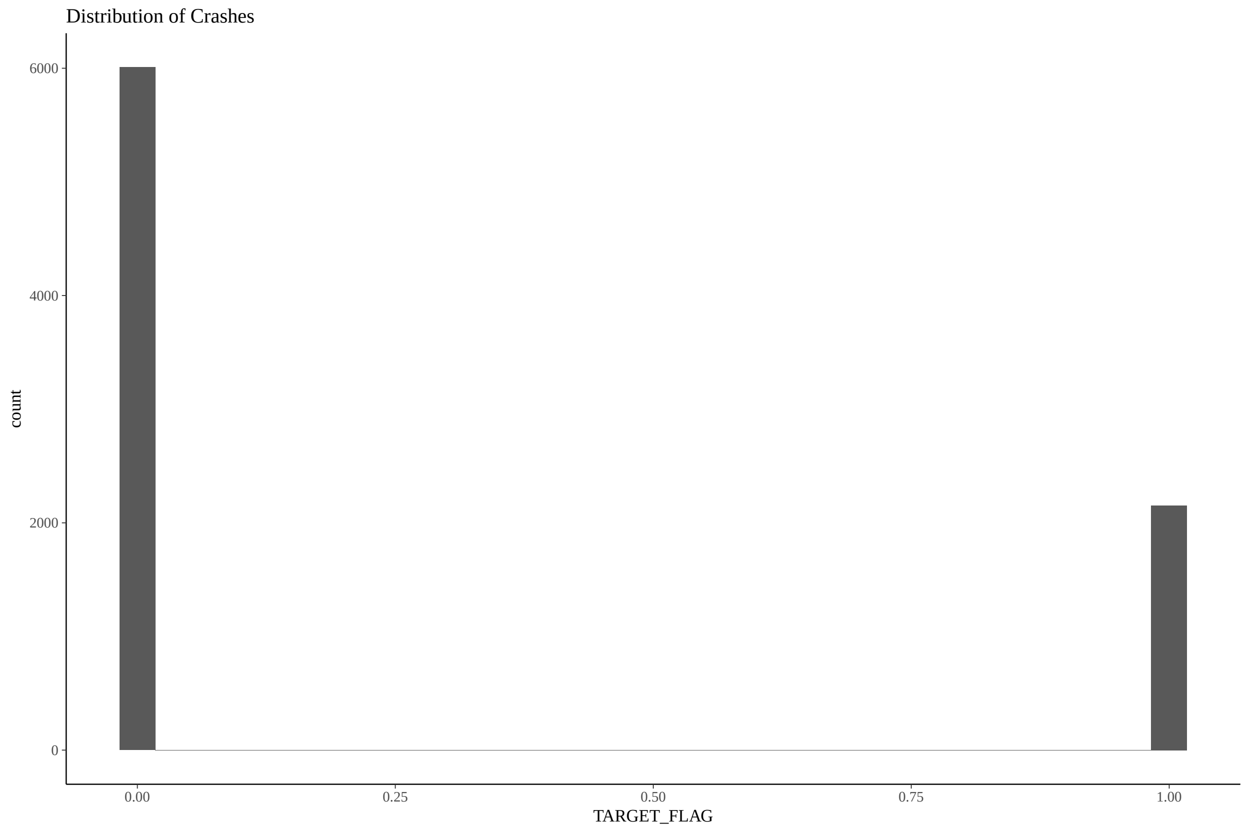
Below are the inference from the summary:

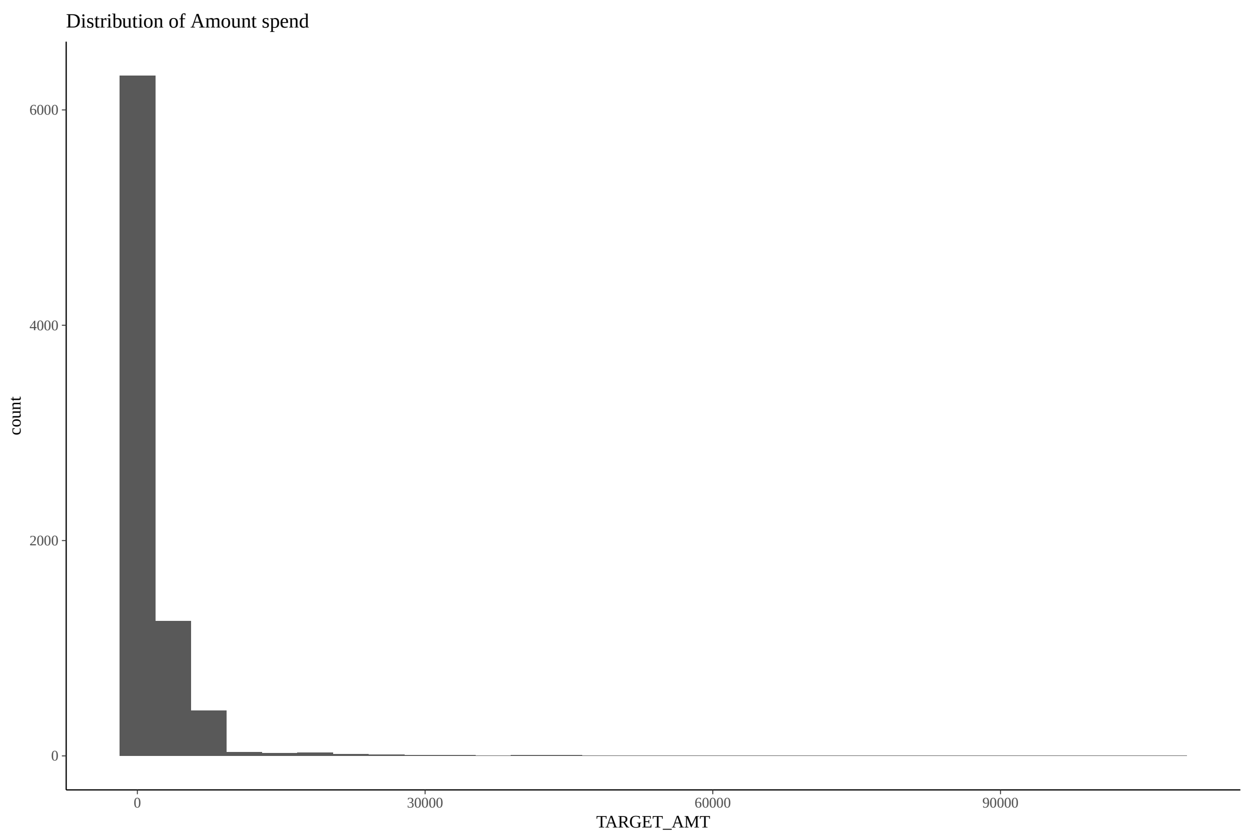
1. Index feature can be removed
2. `Age,YOJ, CAR_AGE` variable has `NA` data. It needs to be handelled appropriately.
3. `OLDCLAIM, BLUEBOOK,HOME_VAL, INCOME` has some blank data. And it has $ sign in it. So it is considered as factor. Need to clean the data.
4. `PARENT1, MSTATUS,SEX,EDUCATION,JOB,CAR_USE,CAR_TYPE,RED_CAR,REVOKED,URBANICITY` is coded as categorical variable. It needs to be changed as dummy variable in the model.
5. `CAR_AGE` has negative value. It needs to be corrected.

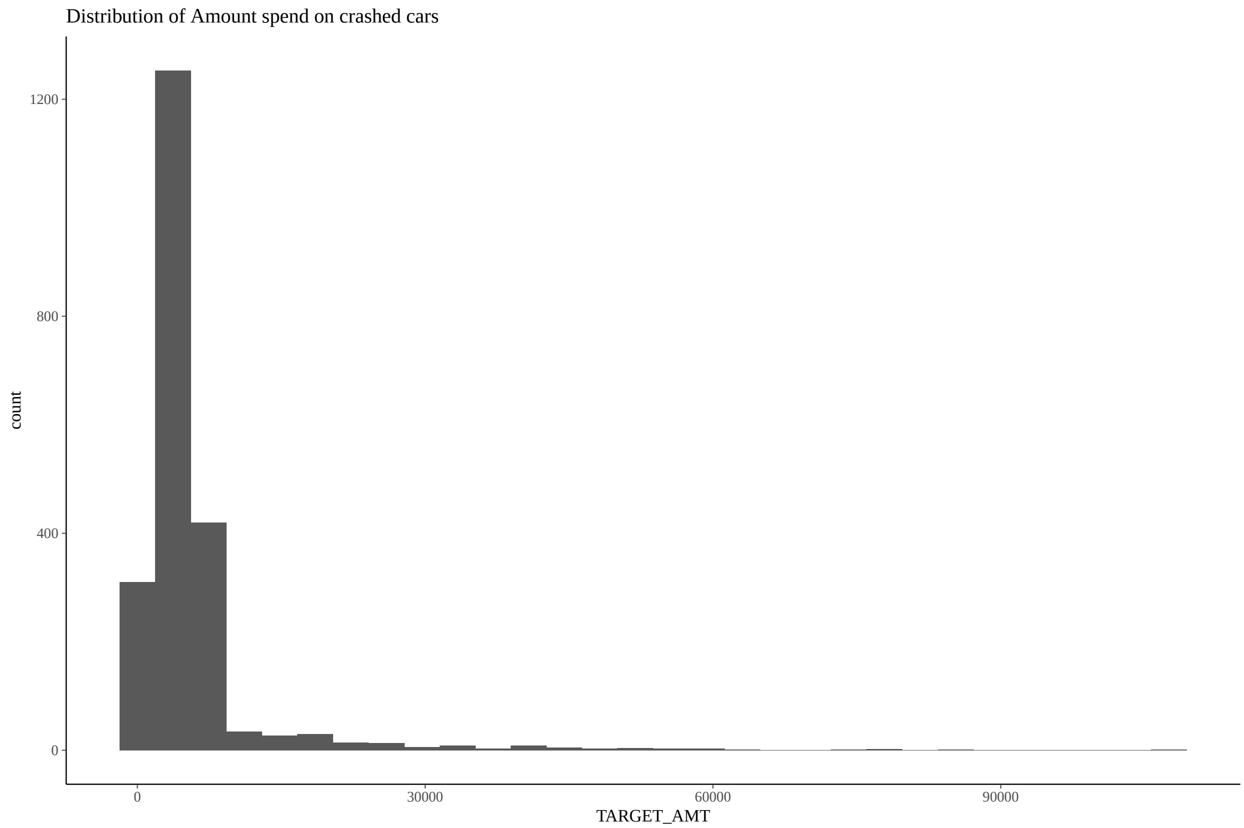
As lot of cleaning needs to be performed, we will draw necessary plots after data preparation.

### 1.1.1 Response variables

For this dataset, we have two response variables. They are `TARGET_FLAG` and `TARGET_AMT`. `TARGET_FLAG` mentions wheather the person will have a car crash or not.


Distribution of Crashes
count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
TARGET_FLAG


Distribution of Amount spend
count
6000
4000
2000
0
0
30000
60000
90000
TARGET_AMT




## 1.2 Data Preparation

Different data preparation needs to be performed. We will try to clean the data one by one.

### 1.2.1 Data Clearning

#### 1.2.1.1 Fixing $ value

As a first step, there are some columns which has $ symbol in the values. Lets fix it in the first step so we can have numeric values.

#### 1.2.1.2 Dropping Index column

As index column is not required, we will drop the index column.

```
##   TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1 HOME_VAL
## 1           0          0        0  60        0  11  67349      No        0
## 2           0          0        0  43        0  11  91449      No   257252
## 3           0          0        0  35        1  10  16039      No   124191
## 4           0          0        0  51        0  14     NA      No   306251
## 5           0          0        0  50        0  NA 114986      No   243925
## 6           1       2946        0  34        1  12 125301     Yes        0
##   MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK TIF
## 1    z_No   M           PhD  Professional       14    Private    14230  11
## 2    z_No   M z_High School z_Blue Collar       22 Commercial    14940   1
## 3     Yes z_F z_High School      Clerical        5    Private     4010   4
```

```
## 4     Yes   M  <High School z_Blue Collar       32    Private    15440   7
## 5     Yes z_F          PhD       Doctor       36    Private    18000   1
## 6    z_No z_F    Bachelors z_Blue Collar       46 Commercial    17430   1
##      CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1    Minivan     yes     4461        2      No       3      18
## 2    Minivan     yes        0        0      No       0       1
## 3      z_SUV      no    38690        2      No       3      10
## 4    Minivan     yes        0        0      No       0       6
## 5      z_SUV      no    19217        2     Yes       3      17
## 6 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
```

Summary of the dataset after performing cleaning the amount variables.

```
##   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   No :7084
##  1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   Yes:1077
##  Median :0.0000   Median :11.0   Median : 54028
##  Mean   :0.7212   Mean   :10.5   Mean   : 61898
##  3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
##  Max.   :5.0000   Max.   :23.0   Max.   :367030
##                   NA's   :454    NA's   :445
##     HOME_VAL      MSTATUS      SEX             EDUCATION
##  Min.   :     0   Yes :4894   M  :3786   <High School :1203
##  1st Qu.:     0   z_No:3267   z_F:4375   Bachelors    :2242
##  Median :161160                          Masters      :1658
##  Mean   :154867                          PhD          : 728
##  3rd Qu.:238724                          z_High School:2330
##  Max.   :885282
##  NA's   :464
##             JOB           TRAVTIME            CAR_USE         BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   Min.   : 1500
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   1st Qu.: 9280
##  Professional :1117   Median : 33.00                     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   Minivan    :2145   no :5783   Min.   :    0
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   1st Qu.:    0
```

```
##  Median : 4.000   Pickup     :1389             Median :    0
##  Mean   : 5.351   Sports Car : 907             Mean   : 4037
##  3rd Qu.: 7.000   Van        : 750             3rd Qu.: 4636
##  Max.   :25.000   z_SUV      :2294             Max.   :57037
##
##     CLM_FREQ      REVOKED       MVR_PTS         CAR_AGE
##  Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              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
##  Highly Urban/ Urban  :6492
##  z_Highly Rural/ Rural:1669
##
##
##
##
##
```

### 1.2.2 Fixing NA Values

In this dataset, there are missing values in `AGE,YOJ, CAR_AGE, INCOME, HOME_VAL` variables. Each needs to be imputed differently. Lets impute the values by each variable.

As a first step lets validate the records which are invalid or has `NA` on multiple columns.

1. We cannot have `CAR_AGE` as negative. So lets drop the observations.
2. If multiple variables like `HOMVE_VAL, INCOME, CAR_AGE, YOJ` are having NA we will drop those records.
3. Lets drop the observations which has `HOME_VAL` as NA. Because the median house value is more than the mean. If imputation is performed, then it might skew the variable. So we will drop NA records.

```
##   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   :0.2639   Mean   :  1497   Mean   :0.1726   Mean   :44.76
##  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   :4
##     HOMEKIDS          YOJ             INCOME        PARENT1
##  Min.   :0.0000   Min.   : 0.00   Min.   :     0   No :6672
##  1st Qu.:0.0000   1st Qu.: 9.00   1st Qu.: 28117   Yes:1024
##  Median :0.0000   Median :11.00   Median : 54124
##  Mean   :0.7265   Mean   :10.51   Mean   : 61896
##  3rd Qu.:1.0000   3rd Qu.:13.00   3rd Qu.: 86212
##  Max.   :5.0000   Max.   :23.00   Max.   :367030
##                   NA's   :427     NA's   :412
##     HOME_VAL      MSTATUS      SEX               EDUCATION
##  Min.   :     0   Yes :4610   M  :3569   <High School :1136
##  1st Qu.:     0   z_No:3086   z_F:4127   Bachelors    :2121
##  Median :161139                          Masters      :1552
##  Mean   :154860                          PhD          : 682
```

```
##  3rd Qu.:238724                        z_High School:2205
##  Max.   :885282
##
##            JOB         TRAVTIME            CAR_USE       BLUEBOOK
##  z_Blue Collar:1723   Min.   :  5.00   Commercial:2844   Min.   : 1500
##  Clerical     :1204   1st Qu.: 22.00   Private   :4852   1st Qu.: 9358
##  Professional :1052   Median : 33.00                     Median :14450
##  Manager      : 934   Mean   : 33.52                     Mean   :15721
##  Lawyer       : 795   3rd Qu.: 44.00                     3rd Qu.:20823
##  Student      : 667   Max.   :142.00                     Max.   :69740
##  (Other)      :1321
##       TIF               CAR_TYPE    RED_CAR      OLDCLAIM
##  Min.   : 1.000   Minivan    :2039   no :5452   Min.   :    0
##  1st Qu.: 1.000   Panel Truck: 632   yes:2244   1st Qu.:    0
##  Median : 4.000   Pickup     :1304              Median :    0
##  Mean   : 5.358   Sports Car : 855              Mean   : 4027
##  3rd Qu.: 7.000   Van        : 701              3rd Qu.: 4603
##  Max.   :25.000   z_SUV      :2165              Max.   :57037
##
##     CLM_FREQ      REVOKED       MVR_PTS         CAR_AGE
##  Min.   :0.0000   No :6753   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:0.0000   Yes: 943   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              Median : 1.000   Median : 8.000
##  Mean   :0.7947              Mean   : 1.685   Mean   : 8.321
##  3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000              Max.   :13.000   Max.   :28.000
##                                               NA's   :474
##                  URBANICITY
##  Highly Urban/ Urban  :6118
##  z_Highly Rural/ Rural:1578
##
##
##
##
##
```

#### 1.2.3 Imputation

As different columns `AGE,YOJ, CAR_AGE, INCOME, HOME_VAL` have NA variables, we need to fill those values with some sort of imputation. We will try different types of imputation.

##### 1.2.3.1 KNN Imputation

Everyone driving should have a minimum age of 18. And the observations which has `NA` seems to kids. So their age should be more than 21+. `KNN` imputation will search for similar records and use the value for missing records.

##### 1.2.3.2 Median Imputation

Another option to perform imputation is using median. We will fill all the missing values as median value of that column.

#### 1.2.3.3 Mice Imputation

mice short for Multivariate Imputation by Chained Equations is an R package that provides advanced features for missing value treatment. It uses a slightly uncommon way of implementing the imputation in 2-steps, using mice() to build the model and complete() to generate the completed data. The mice(df) function produces multiple complete copies of df, each with different imputations of the missing data.

### 1.2.4 Imputation of Categorical Variable

`JOB` variable has some blank values. As it is a text column, we cannot use previous methods. We will just create a new job category as `Other`.

```
##     KIDSDRIV           AGE          HOMEKIDS           YOJ
##  Min.   :0.0000   Min.   :16.00   Min.   :0.0000   Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.00
##  Median :0.0000   Median :45.00   Median :0.0000   Median :11.00
##  Mean   :0.1726   Mean   :44.76   Mean   :0.7265   Mean   :10.53
##  3rd Qu.:0.0000   3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.00
##  Max.   :4.0000   Max.   :81.00   Max.   :5.0000   Max.   :23.00
##
##      INCOME       PARENT1       HOME_VAL      MSTATUS      SEX
##  Min.   :     0   No :6672   Min.   :     0   Yes :4610   M  :3569
##  1st Qu.: 29696   Yes:1024   1st Qu.:     0   z_No:3086   z_F:4127
##  Median : 54124              Median :161139
##  Mean   : 61480              Mean   :154860
##  3rd Qu.: 83429              3rd Qu.:238724
##  Max.   :367030              Max.   :885282
##
##           EDUCATION               JOB         TRAVTIME
##  <High School :1136   z_Blue Collar:1723   Min.   :  5.00
##  Bachelors    :2121   Clerical     :1204   1st Qu.: 22.00
##  Masters      :1552   Professional :1052   Median : 33.00
##  PhD          : 682   Manager      : 934   Mean   : 33.52
##  z_High School:2205   Lawyer       : 795   3rd Qu.: 44.00
##                       Student      : 667   Max.   :142.00
##                       (Other)      :1321
##        CAR_USE        BLUEBOOK          TIF                CAR_TYPE
##  Commercial:2844   Min.   : 1500   Min.   : 1.000   Minivan    :2039
##  Private   :4852   1st Qu.: 9358   1st Qu.: 1.000   Panel Truck: 632
##                    Median :14450   Median : 4.000   Pickup     :1304
##                    Mean   :15721   Mean   : 5.358   Sports Car : 855
##                    3rd Qu.:20823   3rd Qu.: 7.000   Van        : 701
##                    Max.   :69740   Max.   :25.000   z_SUV      :2165
##
##  RED_CAR       OLDCLAIM        CLM_FREQ      REVOKED       MVR_PTS
##  no :5452   Min.   :    0   Min.   :0.0000   No :6753   Min.   : 0.000
##  yes:2244   1st Qu.:    0   1st Qu.:0.0000   Yes: 943   1st Qu.: 0.000
##             Median :    0   Median :0.0000              Median : 1.000
##             Mean   : 4027   Mean   :0.7947              Mean   : 1.685
##             3rd Qu.: 4603   3rd Qu.:2.0000              3rd Qu.: 3.000
##             Max.   :57037   Max.   :5.0000              Max.   :13.000
##
##     CAR_AGE                       URBANICITY    TARGET_FLAG
##  Min.   : 0.000   Highly Urban/ Urban  :6118   Min.   :0.0000
```

```
##  1st Qu.: 4.000   z_Highly Rural/ Rural:1578   1st Qu.:0.0000
##  Median : 8.000                                Median :0.0000
##  Mean   : 8.301                                Mean   :0.2639
##  3rd Qu.:12.000                                3rd Qu.:1.0000
##  Max.   :28.000                                Max.   :1.0000
##
##    TARGET_AMT
##  Min.   :     0
##  1st Qu.:     0
##  Median :     0
##  Mean   :  1497
##  3rd Qu.:  1036
##  Max.   :107586
##
```

#### 1.2.5 Feature Engineering and Transformation

We need to perform some transformations and add new features on the input dataset. This will provide more information to the model.

##### 1.2.5.1 Binary Variables Creation

We will convert add some binary variables. This information has been provided in the question. Below variables will be added to the dataset.

1. New variable can have kids or No kids.
2. Education less than High school and greater than high school, so creating a binary variable.
3. In theory, home owners tend to drive more responsibly - So creating a binary variable.
4. If Old claims are performed, then he has higher chances of crash - creating a binary variable.
5. If CLM_FREQ is hig, then there are higher chaces of crash.
6. If a home ownership is there, then less chances of crash.

```
##     KIDSDRIV          AGE          HOMEKIDS           YOJ
##  Min.   :0.0000   Min.   :16.00   Min.   :0.0000   Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.00
##  Median :0.0000   Median :45.00   Median :0.0000   Median :11.00
##  Mean   :0.1726   Mean   :44.76   Mean   :0.7265   Mean   :10.53
##  3rd Qu.:0.0000   3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.00
##  Max.   :4.0000   Max.   :81.00   Max.   :5.0000   Max.   :23.00
##
##      INCOME       PARENT1       HOME_VAL      MSTATUS      SEX
##  Min.   :     0   No :6672   Min.   :     0   Yes :4610   M  :3569
##  1st Qu.: 29696   Yes:1024   1st Qu.:     0   z_No:3086   z_F:4127
##  Median : 54124              Median :161139
##  Mean   : 61480              Mean   :154860
##  3rd Qu.: 83429              3rd Qu.:238724
##  Max.   :367030              Max.   :885282
##
##    EDUCATION                  JOB         TRAVTIME           CAR_USE
##  Min.   :0.0000   z_Blue Collar:1723   Min.   :  5.00   Commercial:2844
##  1st Qu.:1.0000   Clerical     :1204   1st Qu.: 22.00   Private   :4852
##  Median :1.0000   Professional :1052   Median : 33.00
##  Mean   :0.8524   Manager      : 934   Mean   : 33.52
##  3rd Qu.:1.0000   Lawyer       : 795   3rd Qu.: 44.00
```

```
##  Max.   :1.0000   Student      : 667   Max.   :142.00
##                   (Other)      :1321
##     BLUEBOOK          TIF               CAR_TYPE    RED_CAR
##  Min.   : 1500   Min.   : 1.000   Minivan    :2039   no :5452
##  1st Qu.: 9358   1st Qu.: 1.000   Panel Truck: 632   yes:2244
##  Median :14450   Median : 4.000   Pickup     :1304
##  Mean   :15721   Mean   : 5.358   Sports Car : 855
##  3rd Qu.:20823   3rd Qu.: 7.000   Van        : 701
##  Max.   :69740   Max.   :25.000   z_SUV      :2165
##
##     OLDCLAIM        CLM_FREQ      REVOKED       MVR_PTS
##  Min.   :    0   Min.   :0.0000   No :6753   Min.   : 0.000
##  1st Qu.:    0   1st Qu.:0.0000   Yes: 943   1st Qu.: 0.000
##  Median :    0   Median :0.0000              Median : 1.000
##  Mean   : 4027   Mean   :0.7947              Mean   : 1.685
##  3rd Qu.: 4603   3rd Qu.:2.0000              3rd Qu.: 3.000
##  Max.   :57037   Max.   :5.0000              Max.   :13.000
##
##     CAR_AGE                            URBANICITY    TARGET_FLAG
##  Min.   : 0.000   Highly Urban/ Urban  :6118   Min.   :0.0000
##  1st Qu.: 4.000   z_Highly Rural/ Rural:1578   1st Qu.:0.0000
##  Median : 8.000                                Median :0.0000
##  Mean   : 8.301                                Mean   :0.2639
##  3rd Qu.:12.000                                3rd Qu.:1.0000
##  Max.   :28.000                                Max.   :1.0000
##
##    TARGET_AMT      KIDSDRIV_BIN     HOMEKIDS_BIN     OLDCLAIM_BIN
##  Min.   :      0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:      0   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :      0   Median :1.0000   Median :1.0000   Median :0.0000
##  Mean   :   1497   Mean   :0.8794   Mean   :0.6463   Mean   :0.3833
##  3rd Qu.:   1036   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :107586   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##     HOME_OWN
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2981
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

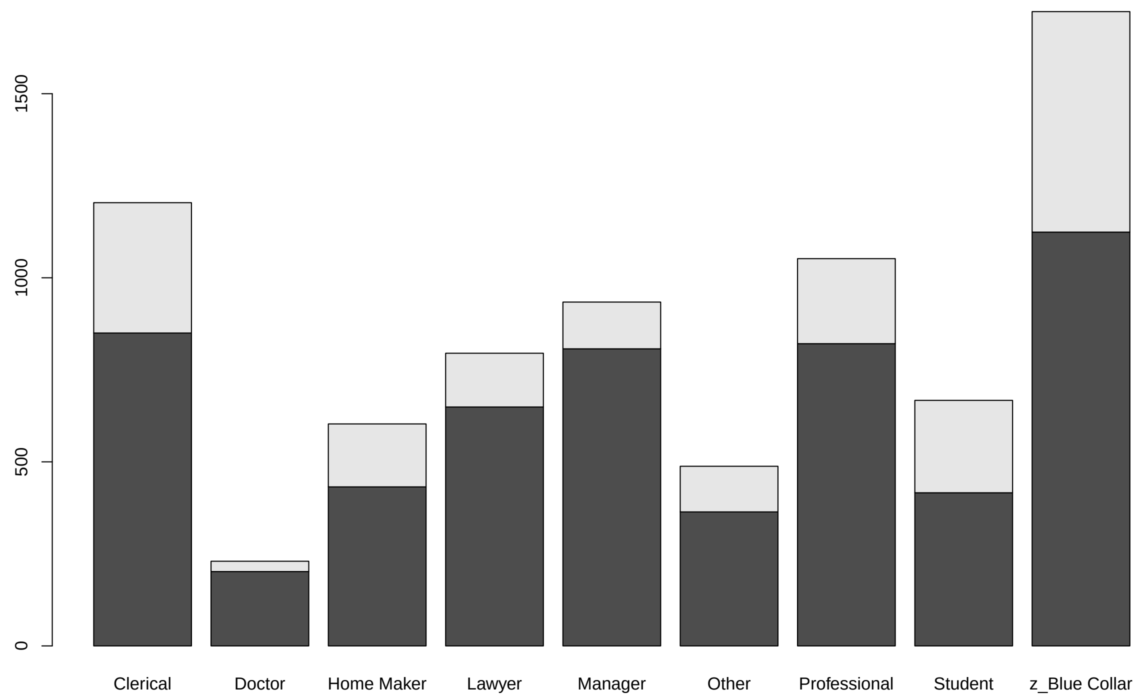
As a next step, we will also transform INCOME varaiable to different bins. We will split into three parts, low income class, middle class and high income.

```
##      [0, 38666]  [38668, 71121] [71141, 367030]
##            2566            2565            2565
```

#### 1.2.5.2 JOB analysis

Job plays a major role in accidents. Genearlly a person in white-collar is less likely to have an accident compared to blue-collar or a car driver. Because white-collar person works in a secured office and may not travel much.

Below is the distribution of the accidents. Doctors are very less likely to cause an accident.




We can group all the white-collar and blue-collar jobs. Here 'Clerical','Doctor','Lawyer','Manager','Professional','Other' are considered as white-collar job. We will convert all the values as white collar and leave out Home_maker and students.

#### 1.2.6 Correlation Charts

##### 1.2.6.1 TARGET_FLAG Plots

As a next step we will draw some correlation matrix and analyze individual charts. As the dataset has many variables, we will spilt it into different plots.

Above plots suggests that there are some room for improvemnt by performing binning.

#### 1.2.7 Numerical variables transformation

Some of the other predictor variables are not correctly distributed. So we might need to perform transformations to correct the variables.

#### 1.2.8 Adding Dummy Variables

As a next step, there are different factor variables with text. Those need to be converted to dummy variables. This is an important step in preparing the dataset.

Finally we have created dummy variables for all the factor predictor variables. We have also performed the drop-off step. This dummy variables inclusion has increased the variable count.

#### 1.2.9 Correlation matrix

Below is the correlation matrix of the dataset.

```
##           Var1         Var2      Freq
## 1 OLDCLAIM_BIN     CLM_FREQ 0.8693796
## 2     CLM_FREQ OLDCLAIM_BIN 0.8693796
## 3   RED_CAR_no      SEX_z_F 0.6675273
## 4      SEX_z_F   RED_CAR_no 0.6675273
## 5 OLDCLAIM_BIN     OLDCLAIM 0.5813259
## 6     OLDCLAIM OLDCLAIM_BIN 0.5813259
```

#### 1.2.10 TRAN TEST Split

As a final step before we build our models, we need to validate the models which we will build. However, there is no test dataset. We will split the dataset into two parts and use the test dataset to validate our model.

### 1.3 Build Models and evaluation

After performing all the data cleaning, transformations and feature engineering, we will build different models on car crash classification and cost of an accident(regression).

#### 1.3.1 TARGET_FLAG - Crash prediction

Car crash is an a binary response variable. Whether the crash happened or not. Our Models has to predict the binary variable. So these type of models will be a classification problem.

##### 1.3.1.1 Model 1 - GLM Stepwise selection

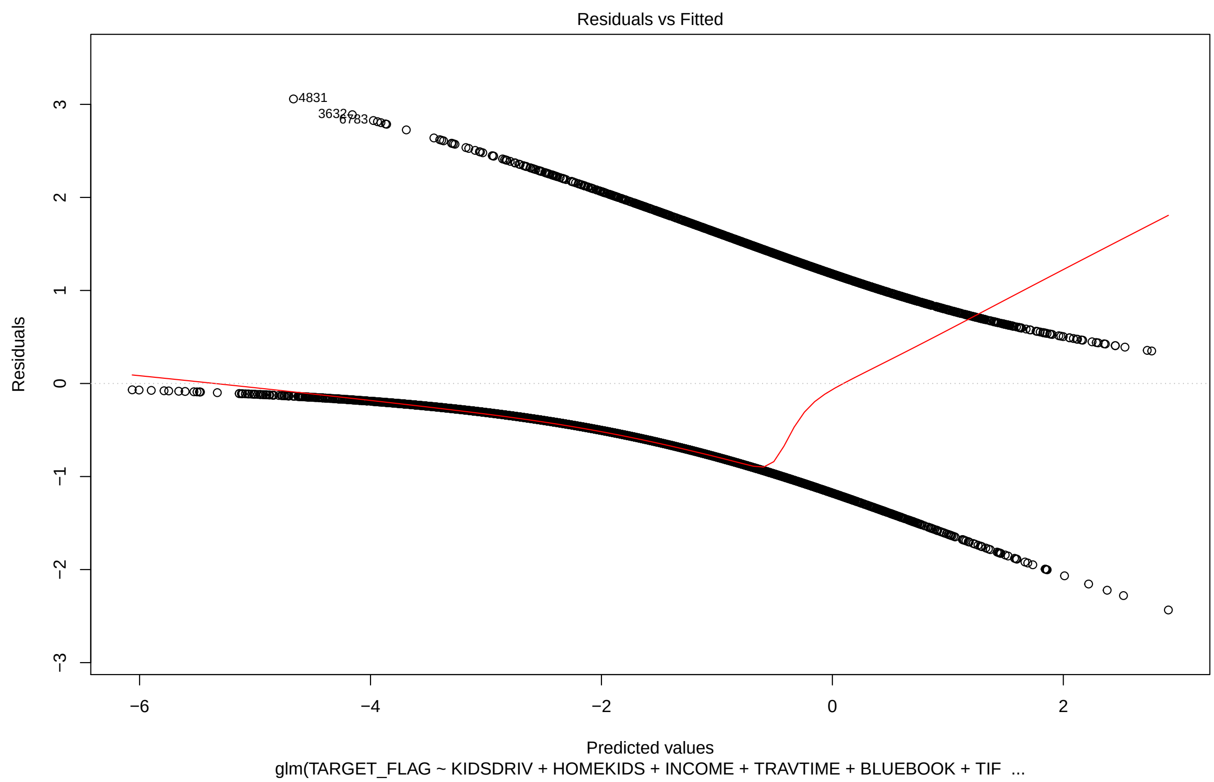
We will create a GLM binomial model with `logit` link function. As there are different variables which not statistcially significant, we will perform backward stepwise variable reduction.

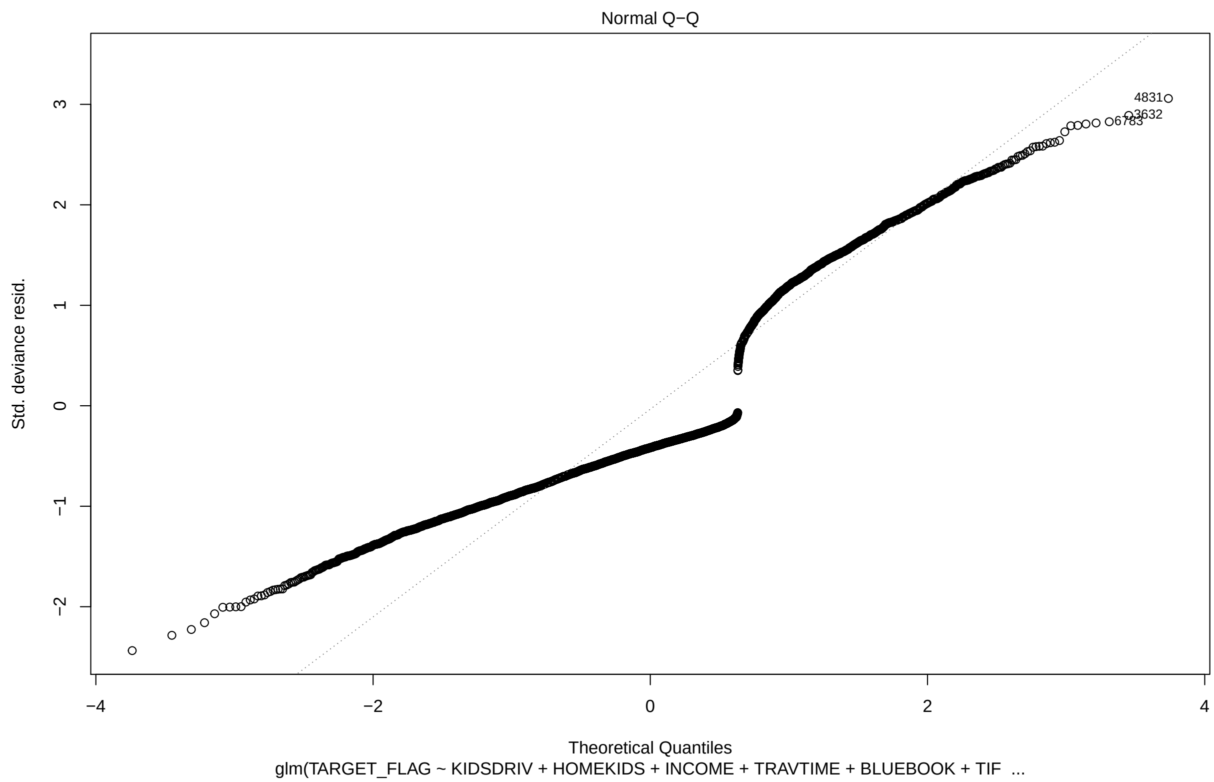
Below are the different evaluation metrics we will perform to validate the model.

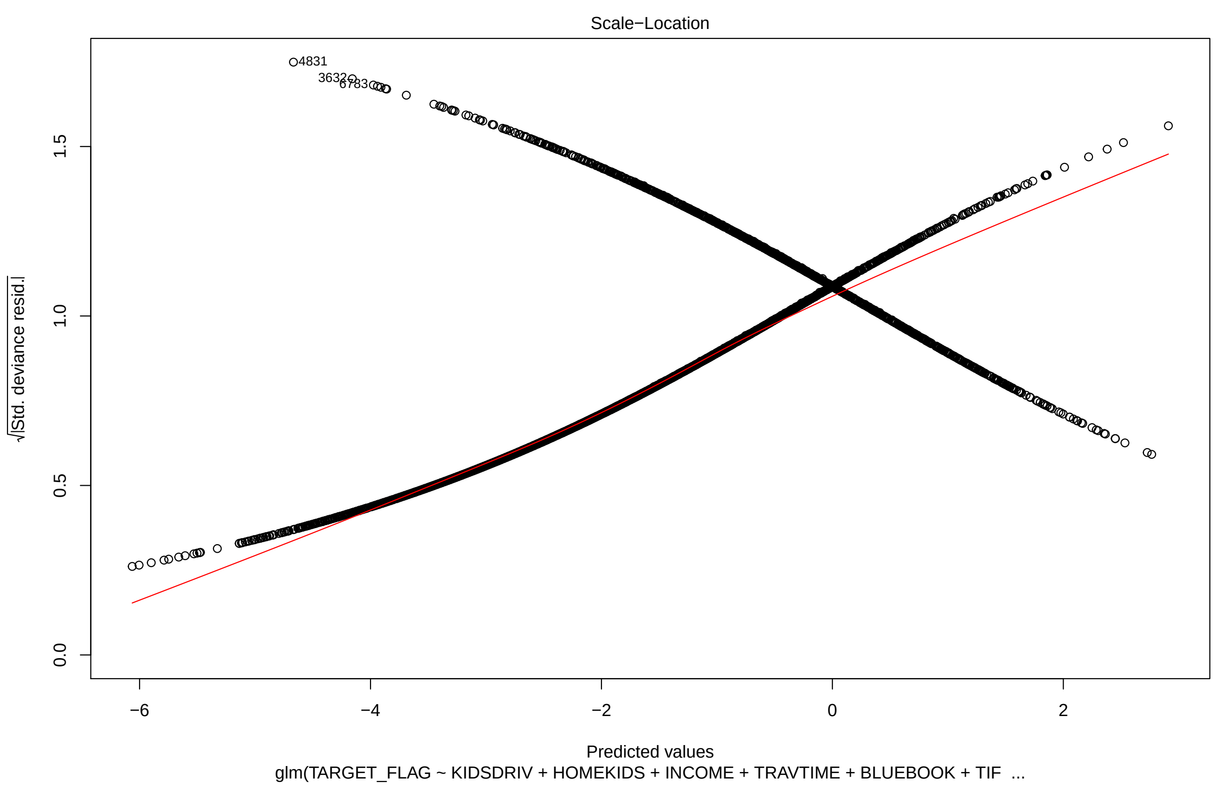
```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + TRAVTIME +
##     BLUEBOOK + TIF + OLDCLAIM + MVR_PTS + CAR_AGE + HOMEKIDS_BIN +
##     OLDCLAIM_BIN + HOME_OWN + PARENT1_Yes + EDUCATION_0 + MSTATUS_Yes +
##     INCOME_BIN_0 + CAR_USE_Commercial + CAR_TYPE_z_SUV + `CAR_TYPE_Sports Car` +
##     CAR_TYPE_Van + `CAR_TYPE_Panel Truck` + CAR_TYPE_Pickup +
##     REVOKED_Yes + `URBANICITY_z_Highly Rural/ Rural`, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4343  -0.7295  -0.4164   0.6618   3.0583
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -8.189e-01  2.524e-01  -3.244 0.001179
## KIDSDRIV                           2.839e-01  7.216e-02   3.934 8.36e-05
```

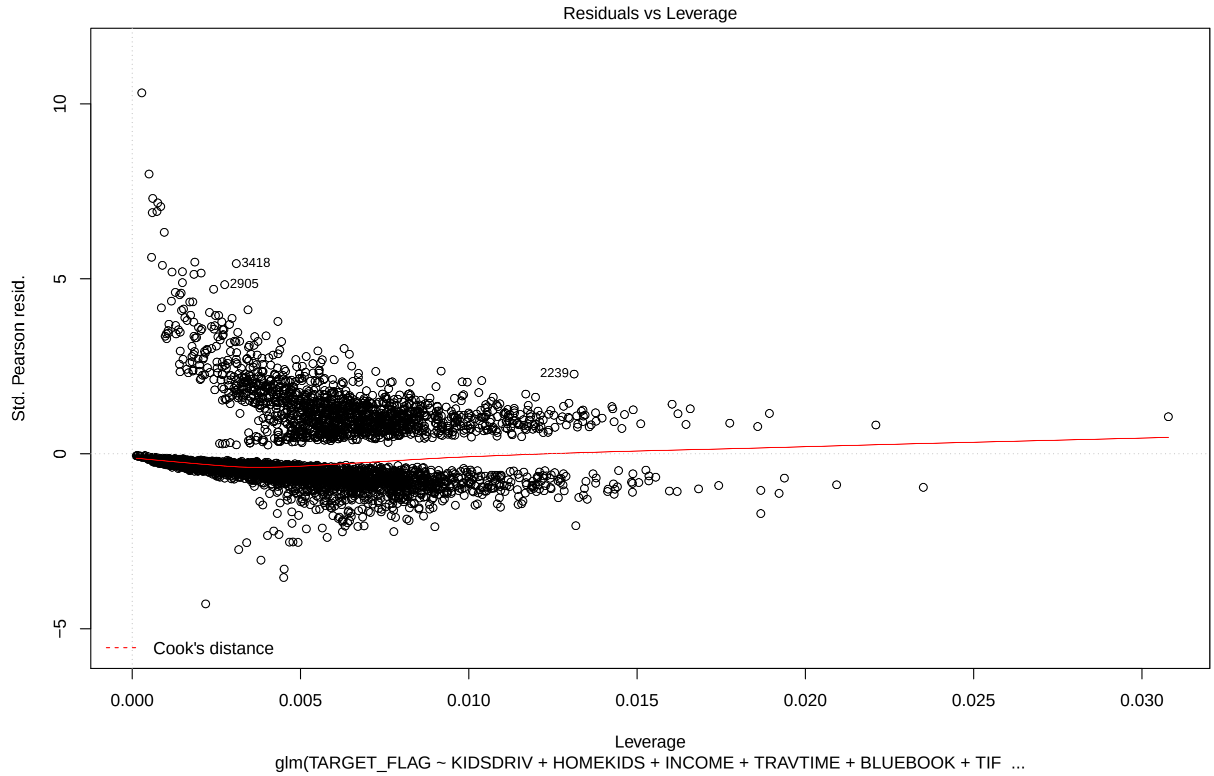
```
## HOMEKIDS                             -1.151e-01  6.405e-02  -1.797 0.072324
## INCOME                               -7.705e-06  1.114e-06  -6.919 4.56e-12
## TRAVTIME                              1.546e-02  2.284e-03   6.770 1.29e-11
## BLUEBOOK                             -2.483e-05  5.813e-06  -4.272 1.94e-05
## TIF                                  -5.009e-02  9.014e-03  -5.557 2.74e-08
## OLDCLAIM                             -1.688e-05  5.125e-06  -3.295 0.000986
## MVR_PTS                               9.799e-02  1.737e-02   5.640 1.70e-08
## CAR_AGE                              -2.254e-02  7.431e-03  -3.033 0.002418
## HOMEKIDS_BIN                         -4.775e-01  1.634e-01  -2.923 0.003469
## OLDCLAIM_BIN                          5.125e-01  9.612e-02   5.332 9.72e-08
## HOME_OWN                              2.948e-01  9.380e-02   3.143 0.001675
## PARENT1_Yes                           2.877e-01  1.471e-01   1.957 0.050396
## EDUCATION_0                           3.491e-01  1.101e-01   3.170 0.001525
## MSTATUS_Yes                          -4.965e-01  1.073e-01  -4.626 3.73e-06
## INCOME_BIN_0                         -1.427e-01  9.601e-02  -1.487 0.137125
## CAR_USE_Commercial                    9.426e-01  8.878e-02  10.618  < 2e-16
## CAR_TYPE_z_SUV                        6.484e-01  1.022e-01   6.347 2.20e-10
## `CAR_TYPE_Sports Car`                 8.281e-01  1.322e-01   6.265 3.72e-10
## CAR_TYPE_Van                          5.110e-01  1.457e-01   3.508 0.000452
## `CAR_TYPE_Panel Truck`                5.050e-01  1.722e-01   2.932 0.003366
## CAR_TYPE_Pickup                       3.687e-01  1.196e-01   3.083 0.002052
## REVOKED_Yes                           9.009e-01  1.127e-01   7.997 1.28e-15
## `URBANICITY_z_Highly Rural/ Rural` -2.213e+00  1.381e-01 -16.028  < 2e-16
##
## (Intercept)                        **
## KIDSDRIV                           ***
## HOMEKIDS                           .
## INCOME                             ***
## TRAVTIME                           ***
## BLUEBOOK                           ***
## TIF                                ***
## OLDCLAIM                           ***
## MVR_PTS                            ***
## CAR_AGE                            **
## HOMEKIDS_BIN                       **
## OLDCLAIM_BIN                       ***
## HOME_OWN                           **
## PARENT1_Yes                        .
## EDUCATION_0                        **
## MSTATUS_Yes                        ***
## INCOME_BIN_0
## CAR_USE_Commercial                 ***
## CAR_TYPE_z_SUV                     ***
## `CAR_TYPE_Sports Car`              ***
## CAR_TYPE_Van                       ***
## `CAR_TYPE_Panel Truck`             **
## CAR_TYPE_Pickup                    **
## REVOKED_Yes                        ***
## `URBANICITY_z_Highly Rural/ Rural` ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

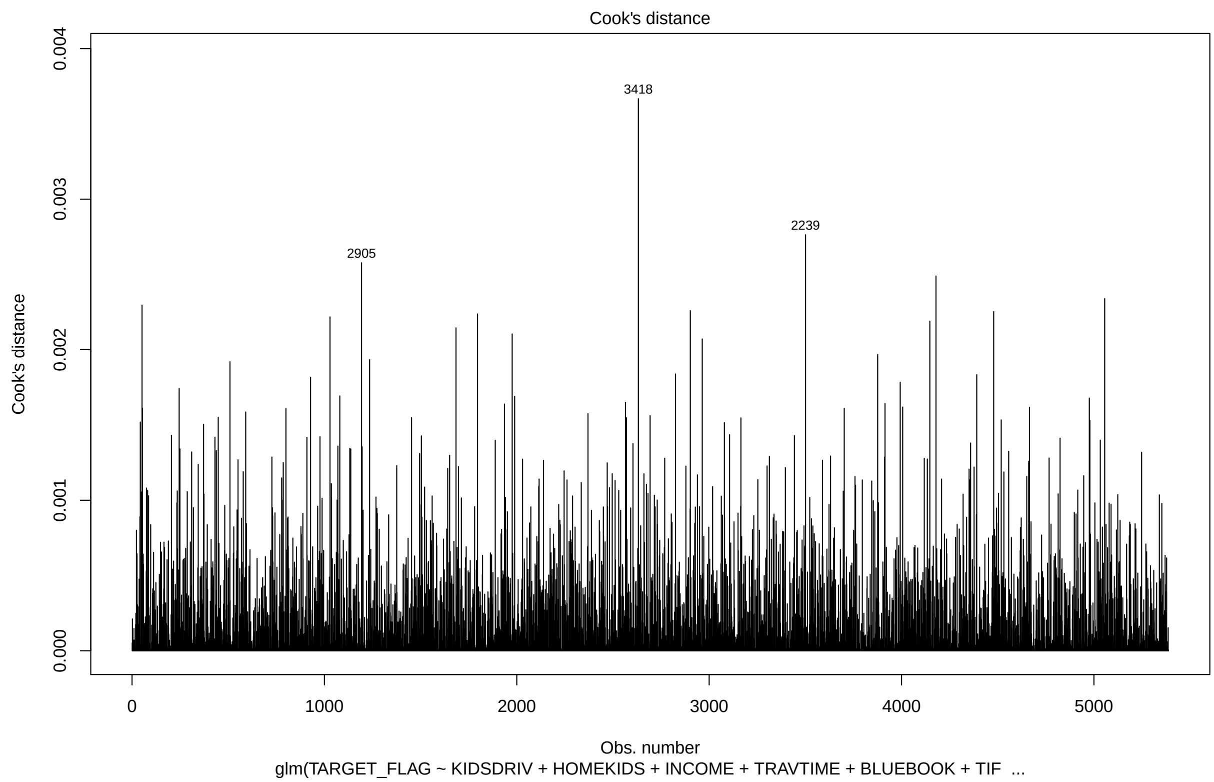
```
##     Null deviance: 6218.4  on 5386  degrees of freedom
## Residual deviance: 4903.3  on 5362  degrees of freedom
## AIC: 4953.3
##
## Number of Fisher Scoring iterations: 5
```




Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + TRAVTIME + BLUEBOOK + TIF ...
4831
3632
6783
−4
−2
0
2
4
−2
−1
0
1
2
3

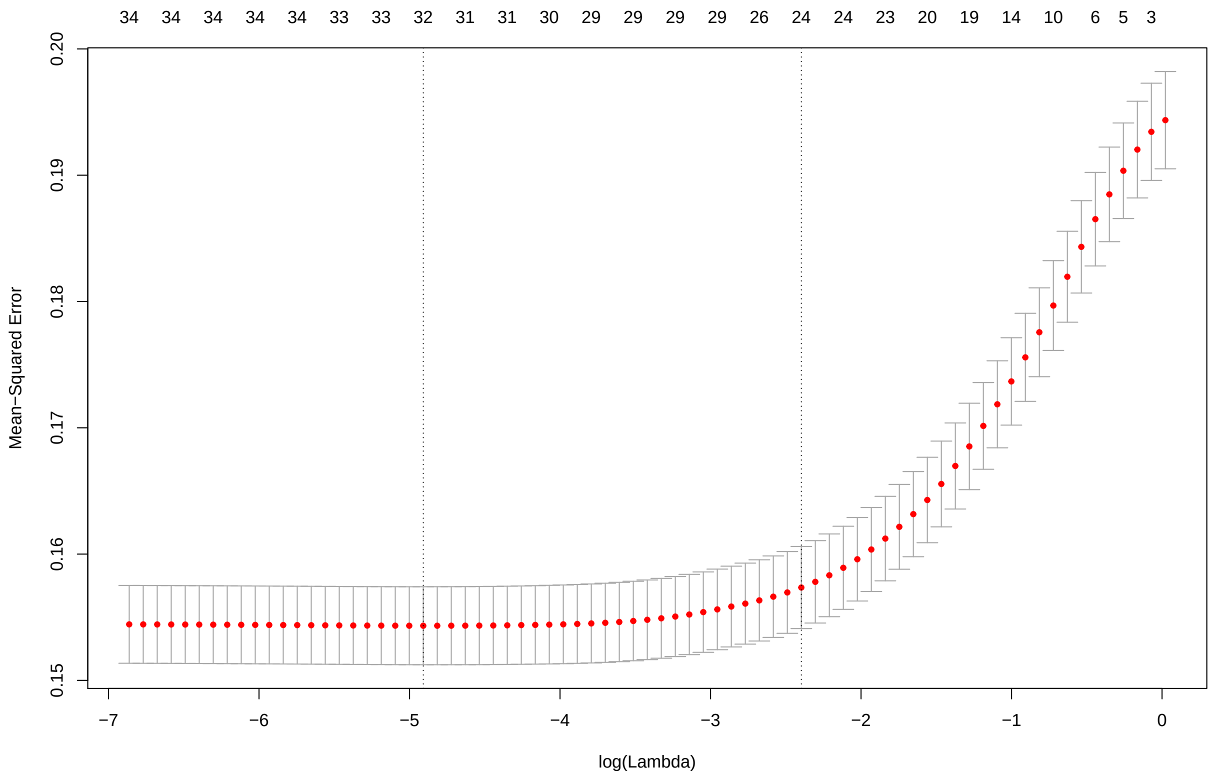

Scale−Location
√|Std. deviance resid.|
Predicted values
glm(TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + TRAVTIME + BLUEBOOK + TIF ...
4831
3632
6783
−6
−4
−2
0
2
0.0
0.5
1.0
1.5


Residuals vs Leverage
Std. Pearson resid.
Leverage
glm(TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + TRAVTIME + BLUEBOOK + TIF ...
Cook's distance
3418
2905
2239


Cook's distance
Cook's distance
Obs. number
glm(TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + TRAVTIME + BLUEBOOK + TIF ...
3418
2905
2239

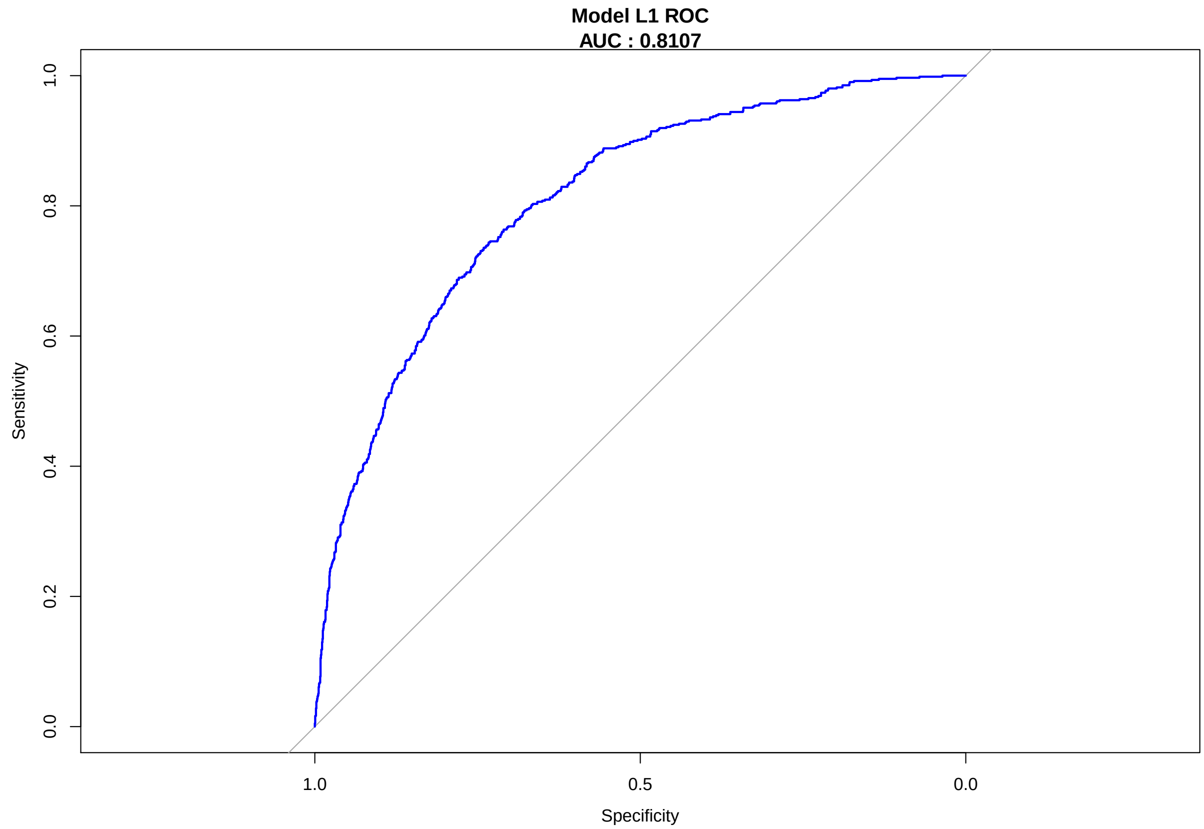
#### 1.3.1.2 Model 2- Lasso Binary regression using GLMNET

In this type of model, we will create a LASSO binary regression using GLMNET package. In this approach, we will shrink the variable coefficents to 0 by selecting the appropirate lambda value.




```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1616  403
##          1   84  206
##
##                Accuracy : 0.7891
##                  95% CI : (0.7719, 0.8056)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : 2.162e-09
##
##                   Kappa : 0.3472
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.33826
##             Specificity : 0.95059
##          Pos Pred Value : 0.71034
##          Neg Pred Value : 0.80040
##              Prevalence : 0.26375
##          Detection Rate : 0.08922
##    Detection Prevalence : 0.12560
##       Balanced Accuracy : 0.64442
```

```
##
##        'Positive' Class : 1
##
```




#### 1.3.1.3 Model 3 - Bayesian Logistic Regression

In this model, we will run Bayesian type logistic regression. Bayesian model calculates the prior and posterior probability using Markov Chain Monte Carlo(MCMC) method.

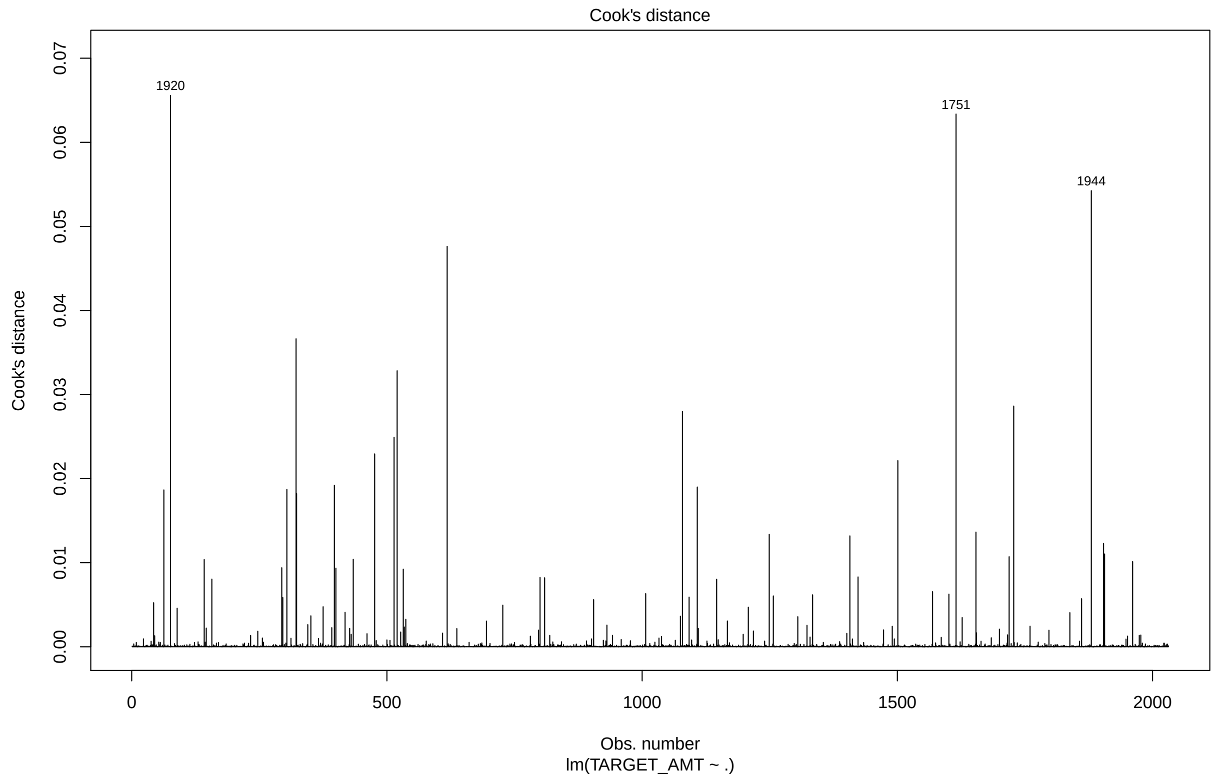
`rstanarm` package provides functions to run Bayesian type models.

### 1.3.2 TARGET_AMT - COST prediction

Previously we have predicted the car crash using the available variables. As a next step, if the accident happens, we will build models to predict the cost of car to pay for that accident.

#### 1.3.2.1 Model 1 - Stepwise selection

As a inital step we will build a simple stepwise model as a base. This will have all the variables and automatic stepwise selection process.



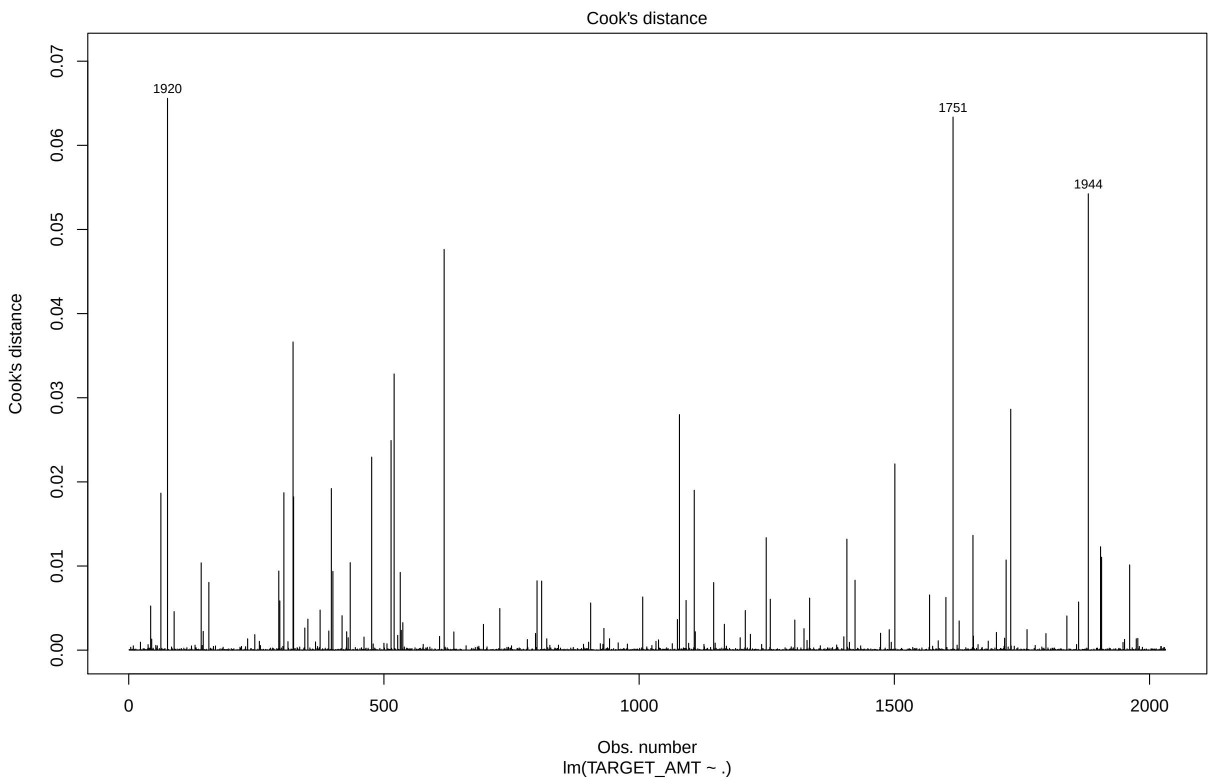

```
## [1] 1944 1751 1920
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = outliers_remove(train_target_amt,
##     model_11_amt_step))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9619  -3181  -1511    499 100148
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          5.432e+03  1.852e+03   2.934  0.00339
## KIDSDRIV                            -8.752e+02  6.608e+02  -1.324  0.18553
## AGE                                  1.754e+01  2.285e+01   0.767  0.44289
## HOMEKIDS                             6.690e+01  3.046e+02   0.220  0.82616
## YOJ                                  3.770e+01  6.003e+01   0.628  0.53006
## INCOME                              -4.202e-03  8.616e-03  -0.488  0.62579
## HOME_VAL                            -7.677e-04  3.761e-03  -0.204  0.83829
## TRAVTIME                             4.878e+00  1.149e+01   0.425  0.67123
## BLUEBOOK                             1.350e-01  3.200e-02   4.217 2.59e-05
## TIF                                 -8.000e+00  4.416e+01  -0.181  0.85624
## OLDCLAIM                             2.446e-02  2.554e-02   0.958  0.33825
## CLM_FREQ                            -1.156e+02  2.469e+02  -0.468  0.63966
## MVR_PTS                              1.312e+02  7.355e+01   1.784  0.07465
## CAR_AGE                             -4.845e+01  3.797e+01  -1.276  0.20208
## KIDSDRIV_BIN                        -1.373e+03  1.110e+03  -1.238  0.21603
```

```
## HOMEKIDS_BIN                         -1.723e+02  8.569e+02  -0.201  0.84062
## OLDCLAIM_BIN                          6.395e+01  7.001e+02   0.091  0.92723
## HOME_OWN                             -8.011e+02  8.815e+02  -0.909  0.36358
## SEX_z_F                              -1.562e+03  6.807e+02  -2.294  0.02189
## PARENT1_Yes                           3.839e+01  7.006e+02   0.055  0.95630
## EDUCATION_0                           4.185e+02  5.247e+02   0.798  0.42521
## MSTATUS_Yes                          -1.015e+03  5.682e+02  -1.786  0.07419
## INCOME_BIN_2                          8.296e+01  5.419e+02   0.153  0.87835
## INCOME_BIN_0                         -4.899e+02  5.078e+02  -0.965  0.33476
## `JOB_z_Blue Collar`                  -2.164e+02  5.136e+02  -0.421  0.67356
## `JOB_Home Maker`                     -1.280e+02  8.207e+02  -0.156  0.87609
## JOB_Student                          -8.274e+01  7.981e+02  -0.104  0.91745
## CAR_USE_Commercial                    4.917e+02  5.283e+02   0.931  0.35213
## CAR_TYPE_z_SUV                        1.055e+03  6.945e+02   1.519  0.12891
## `CAR_TYPE_Sports Car`                 1.323e+03  7.787e+02   1.700  0.08935
## CAR_TYPE_Van                          4.169e+01  8.003e+02   0.052  0.95846
## `CAR_TYPE_Panel Truck`               -4.533e+02  9.888e+02  -0.458  0.64669
## CAR_TYPE_Pickup                      -5.787e+01  6.229e+02  -0.093  0.92598
## RED_CAR_no                            3.383e+01  5.178e+02   0.065  0.94792
## REVOKED_Yes                          -9.183e+02  5.490e+02  -1.672  0.09459
## `URBANICITY_z_Highly Rural/ Rural` -6.785e+01  7.829e+02  -0.087  0.93095
##
## (Intercept)                          **
## KIDSDRIV
## AGE
## HOMEKIDS
## YOJ
## INCOME
## HOME_VAL
## TRAVTIME
## BLUEBOOK                             ***
## TIF
## OLDCLAIM
## CLM_FREQ
## MVR_PTS                              .
## CAR_AGE
## KIDSDRIV_BIN
## HOMEKIDS_BIN
## OLDCLAIM_BIN
## HOME_OWN
## SEX_z_F                              *
## PARENT1_Yes
## EDUCATION_0
## MSTATUS_Yes                          .
## INCOME_BIN_2
## INCOME_BIN_0
## `JOB_z_Blue Collar`
## `JOB_Home Maker`
## JOB_Student
## CAR_USE_Commercial
## CAR_TYPE_z_SUV
## `CAR_TYPE_Sports Car`                .
## CAR_TYPE_Van
## `CAR_TYPE_Panel Truck`
```

```
## CAR_TYPE_Pickup
## RED_CAR_no
## REVOKED_Yes                     .
## `URBANICITY_z_Highly Rural/ Rural`
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7773 on 1992 degrees of freedom
## Multiple R-squared:  0.0284, Adjusted R-squared:  0.01133
## F-statistic: 1.664 on 35 and 1992 DF,  p-value: 0.008865
```

Created model is not very good for this particular dataset. As the response variable is skewed, we will transform the response variable and perform then create a model.




```
## [1] 1944 1751 1920
```

```
##
## Call:
## lm(formula = log(TARGET_AMT) ~ BLUEBOOK + OLDCLAIM + CLM_FREQ +
##     MVR_PTS + SEX_z_F + MSTATUS_Yes + REVOKED_Yes, data = train_target_amt_nooutlier)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6947 -0.3971  0.0304  0.4082  3.2350
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.785e+00  2.601e-01  26.089  < 2e-16 ***
## BLUEBOOK     1.641e-01  2.712e-02   6.050 1.73e-09 ***
```

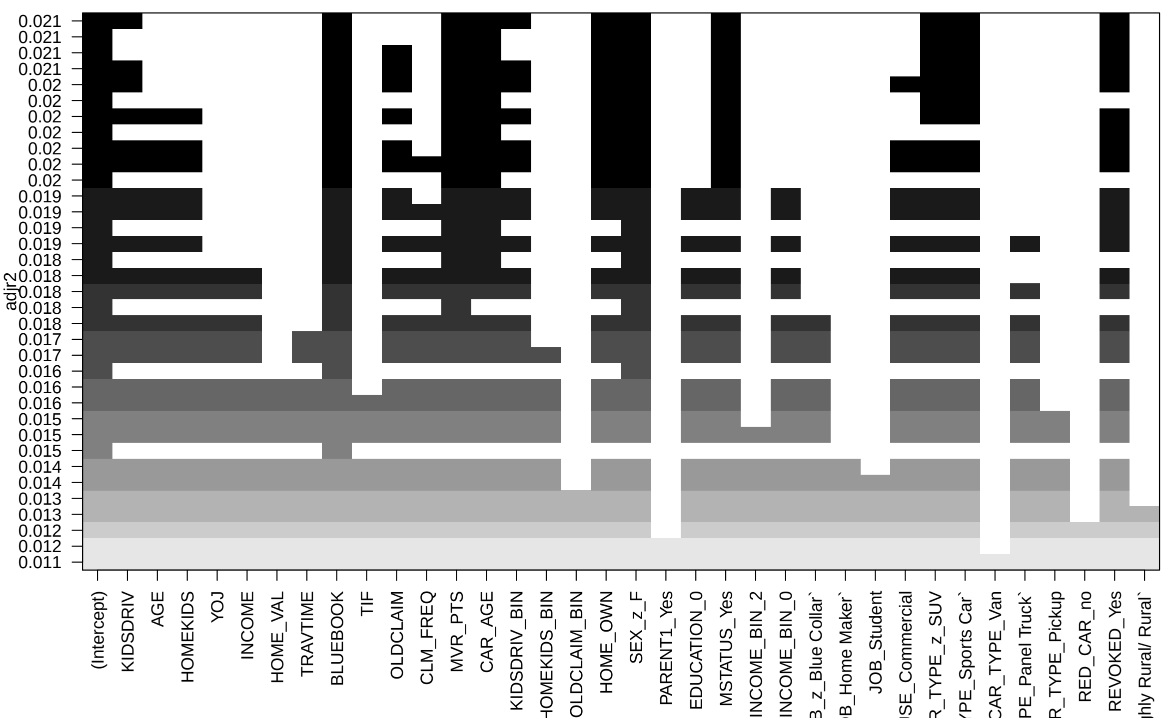
```
## OLDCLAIM     4.954e-06  2.404e-06   2.060   0.0395 *
## CLM_FREQ    -3.262e-02  1.672e-02  -1.951   0.0512 .
## MVR_PTS      1.571e-02  7.280e-03   2.158   0.0310 *
## SEX_z_F     -5.725e-02  3.600e-02  -1.591   0.1119
## MSTATUS_Yes -6.449e-02  3.567e-02  -1.808   0.0707 .
## REVOKED_Yes -8.923e-02  5.467e-02  -1.632   0.1028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.802 on 2020 degrees of freedom
## Multiple R-squared:  0.02546,    Adjusted R-squared:  0.02208
## F-statistic: 7.538 on 7 and 2020 DF,  p-value: 5.443e-09
```

Above model is better than the previous model. However, it has only less variables and the adjusted R2 is not very high. We will try other models and see.

#### 1.3.2.2 Model 2 - Regsubsets

In this model, we will perform automatic selection of the variables using `regsubsets`.

```
## [1] "Adjusted R2:0.0207120371661899"
```




Automatic selection of variables did not improve much on the adj-R2. We will try other different models.

#### 1.3.2.3 Model 3 - Ridge Regression

In this attempt, we will perform Ridge regression. Ridge regression uses L2 regulaization and reduces the coeffecients.

```
## Ridge Regression
##
## 2031 samples
##   35 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 2031, 2031, 2031, 2031, 2031, 2031, ...
## Resampling results across tuning parameters:
##
##   lambda  RMSE      Rsquared     MAE
##   0e+00   8061.362  0.004439284  3876.282
##   1e-04   8061.312  0.004440042  3876.194
##   1e-01   8041.648  0.004880798  3841.835
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was lambda = 0.1.
```

It seems the results are not signficant. Rsquared has not immproved much. So this is also not the best model.

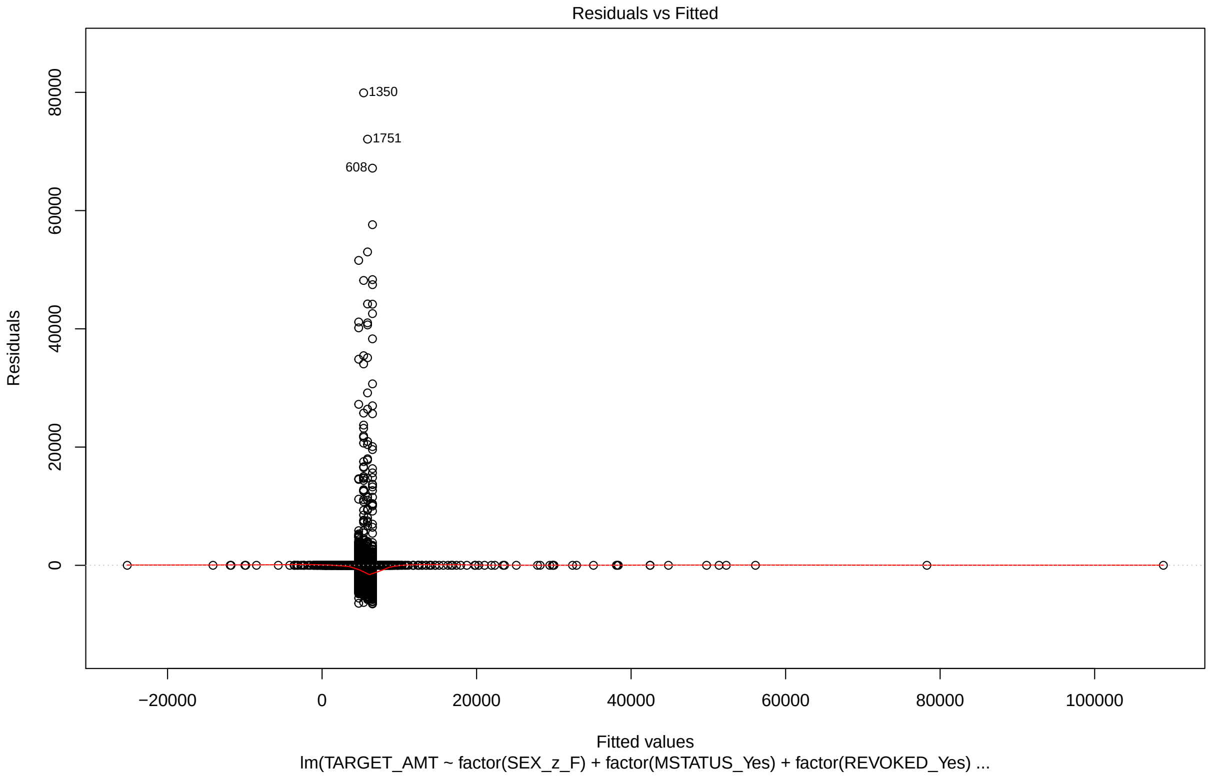
#### 1.3.2.4 Model 4 - Regression Splines

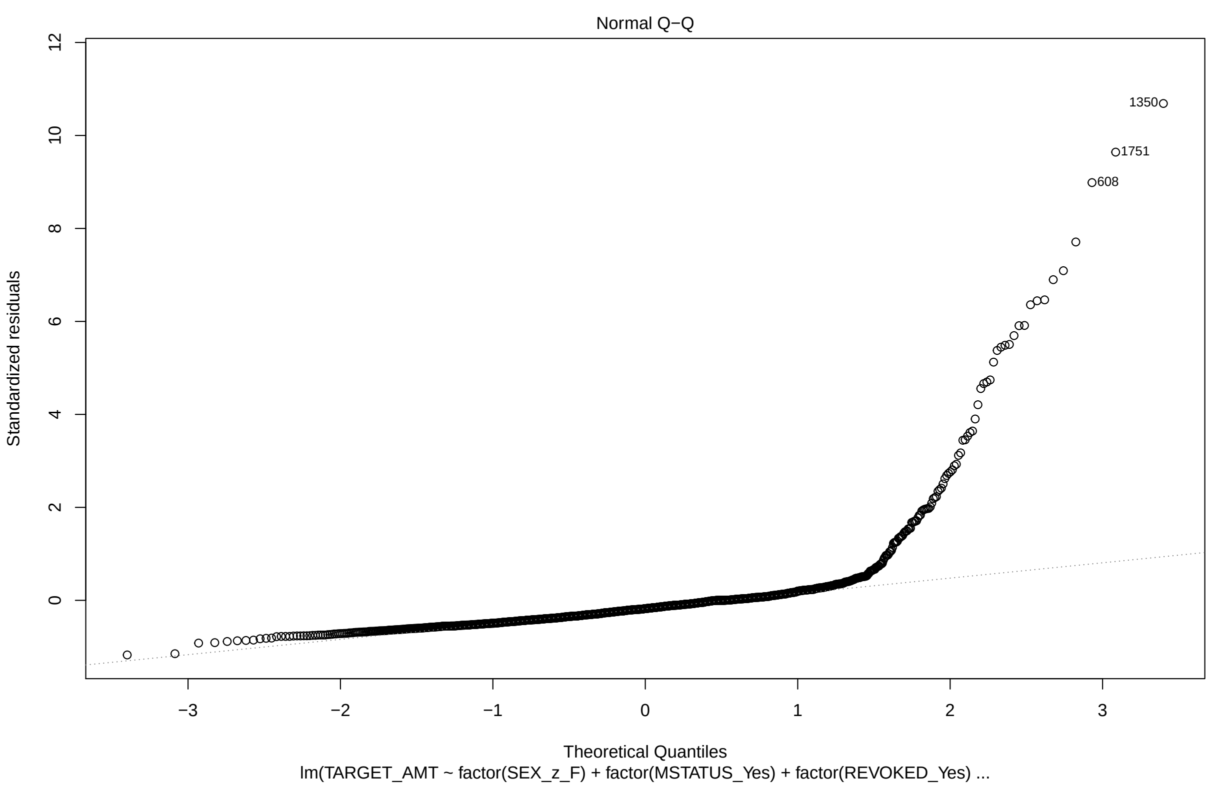
This time we will try a nonlinear model with regression splines. Splines provide a way to smoothly interpolate between fixed points called knots.

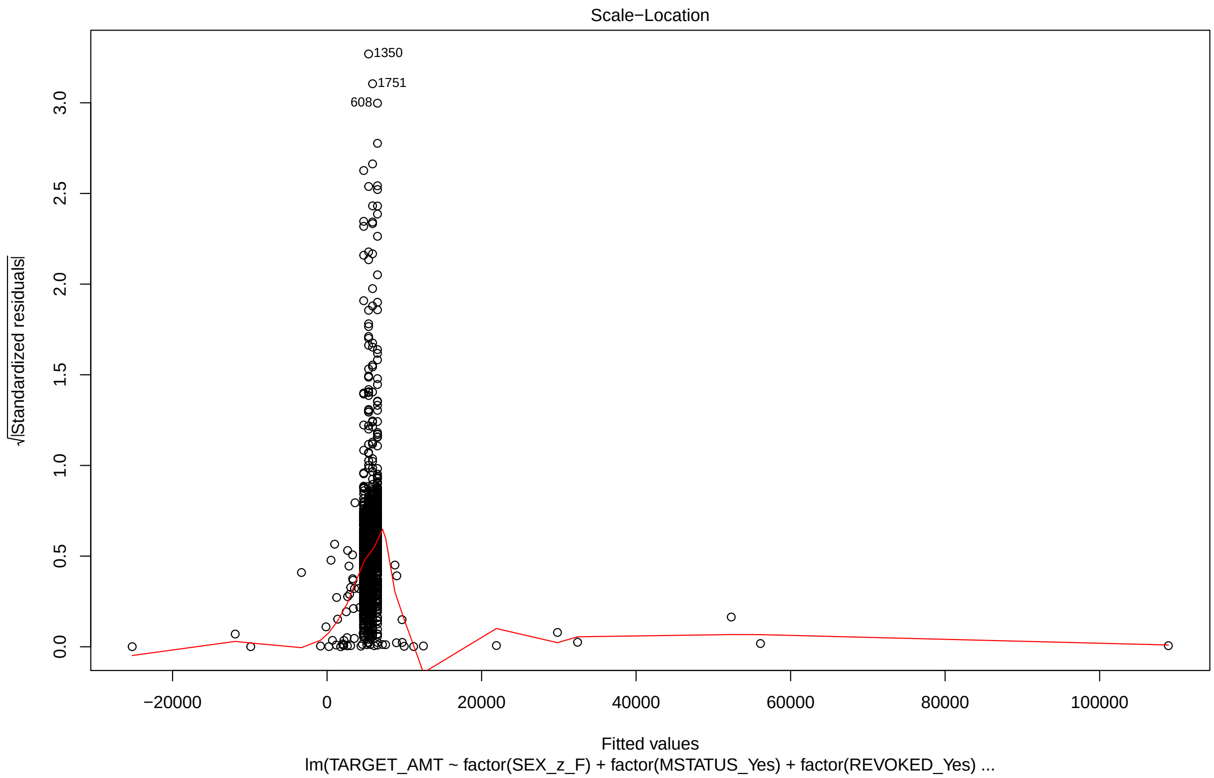
```
## [1] "Adjusted R2: 0.081990025388321"
```

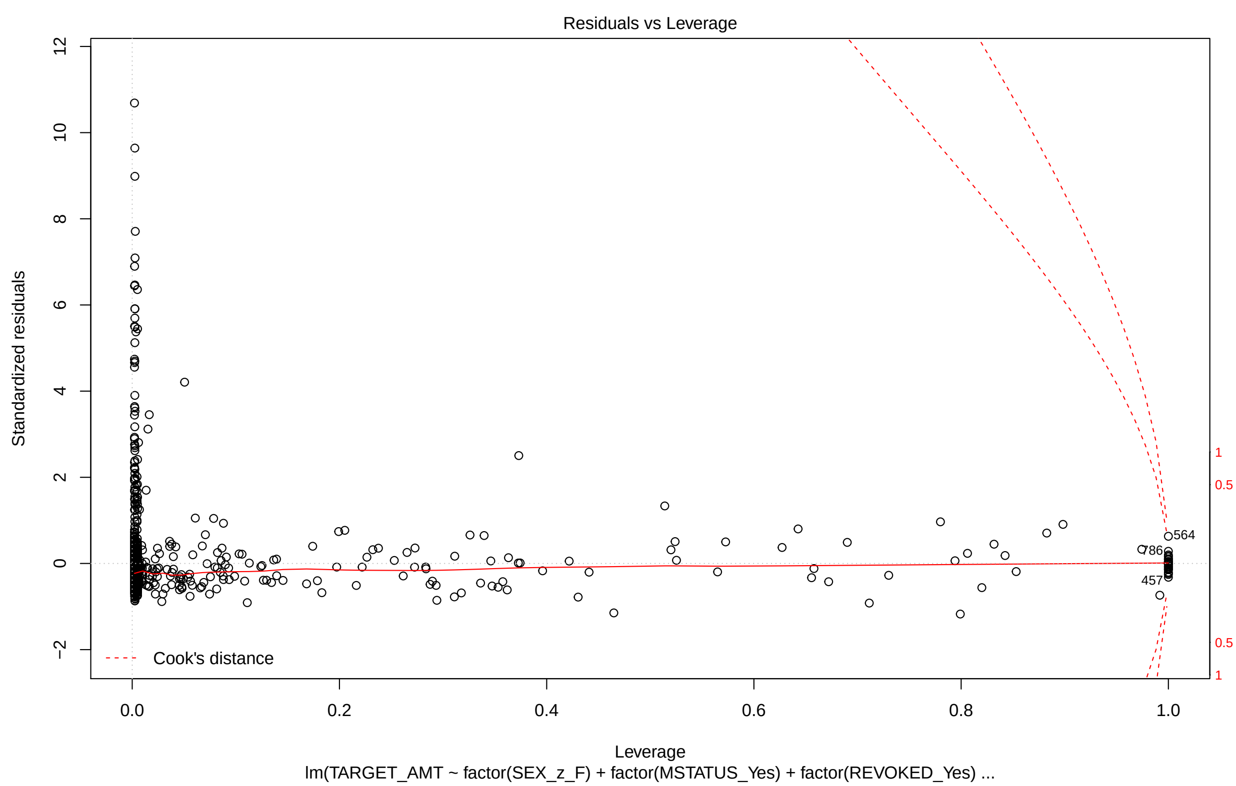
```
## [1] "F-statistic: 1.27387454516885" "F-statistic: 662"
## [3] "F-statistic: 1368"
```

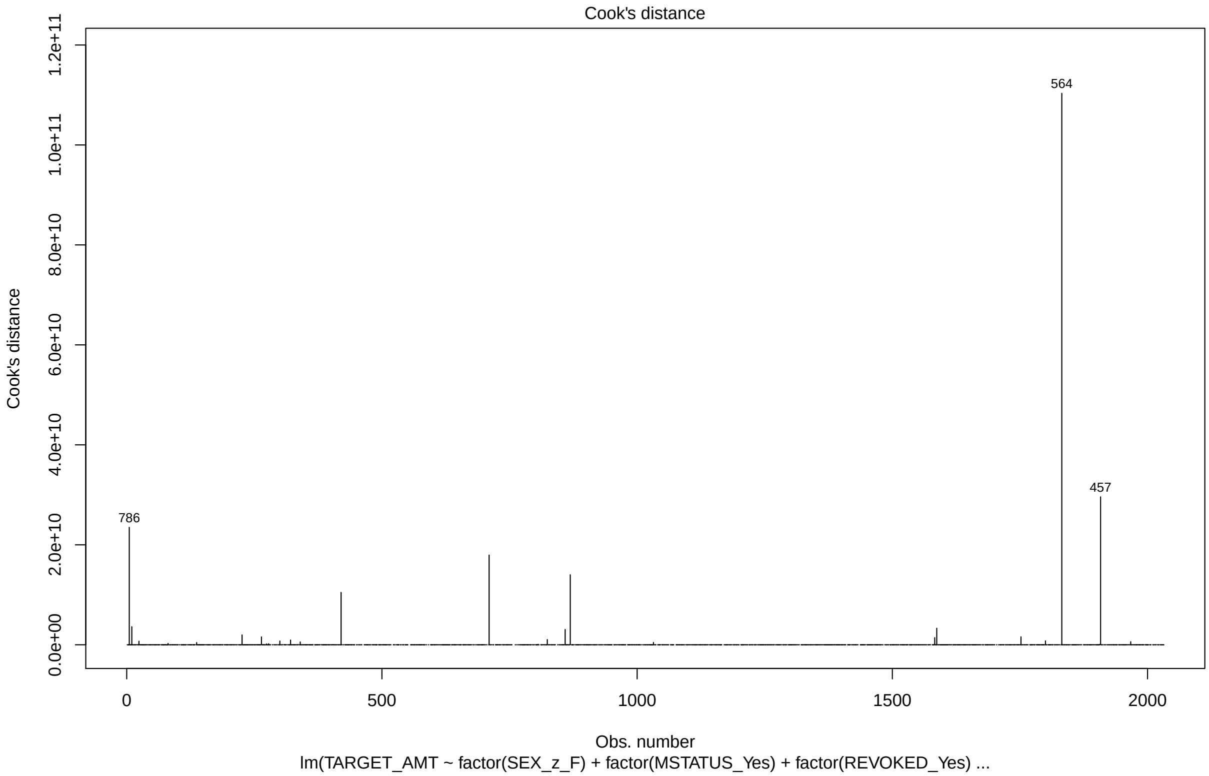
```
## [1] "RMSE: 6487.29566076278"
```


Residuals vs Fitted
Residuals
80000
60000
40000
20000
0
1350
1751
608
−20000
0
20000
40000
60000
80000
100000
Fitted values
lm(TARGET_AMT ~ factor(SEX_z_F) + factor(MSTATUS_Yes) + factor(REVOKED_Yes) ...


Normal Q−Q
Standardized residuals
12
10
8
6
4
2
0
1350
1751
608
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(TARGET_AMT ~ factor(SEX_z_F) + factor(MSTATUS_Yes) + factor(REVOKED_Yes) ...


Scale−Location
√|Standardized residuals|
1350
1751
608
Fitted values
lm(TARGET_AMT ~ factor(SEX_z_F) + factor(MSTATUS_Yes) + factor(REVOKED_Yes) ...


Residuals vs Leverage
Standardized residuals
564
786
457
Cook's distance
Leverage
lm(TARGET_AMT ~ factor(SEX_z_F) + factor(MSTATUS_Yes) + factor(REVOKED_Yes) ...




Above model is build from the base model from stepwise selection. When we add splines, then we get better adjusted R2 compared to other models. However, the residual plots show that there is some autocorrelation. So we will reject this model.

## 1.4 Model Selection

We have build different models and evaluated them. In this section, we will select the final model and add other metrics to it.

### 1.4.1 TARGET_FLAG Model

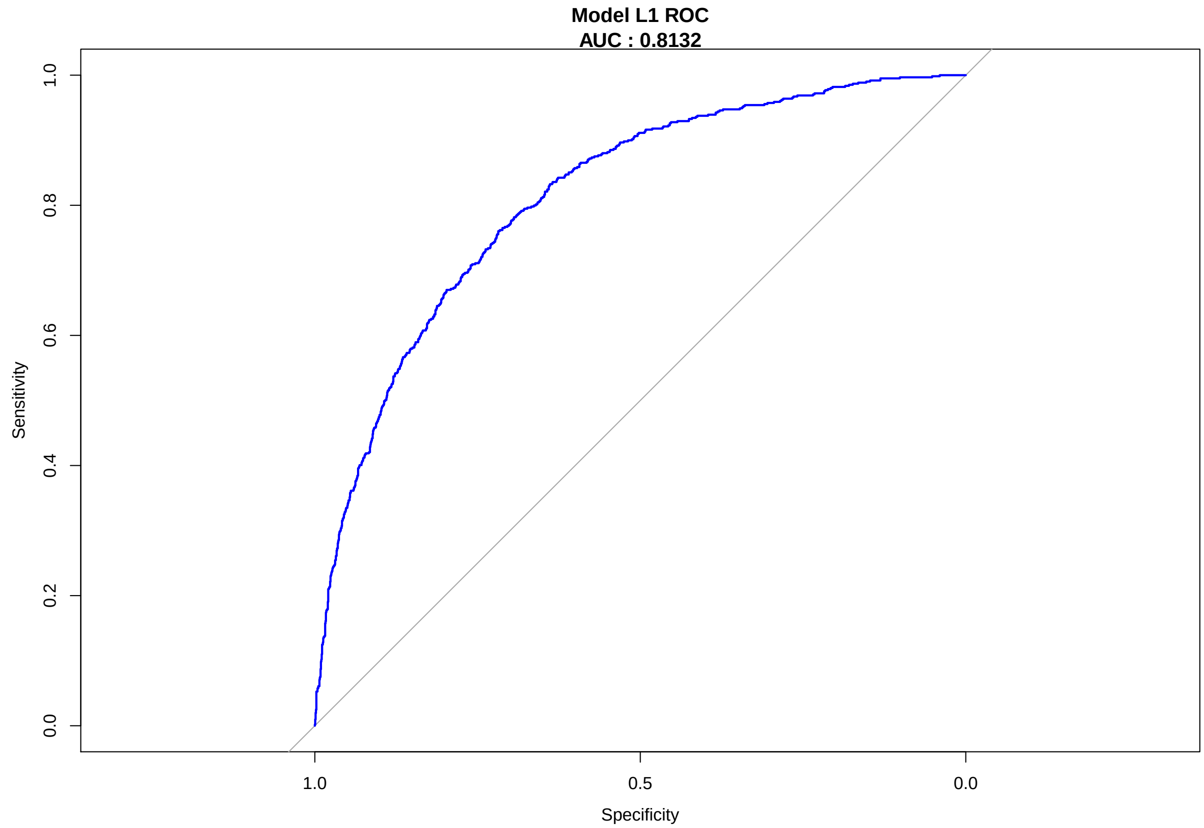
We have build basic model, stepwise model, Lasso logistic regression and regsubsets model. It seems stepwise model is performing good and more interpretable. Lets analyze the model further.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + TRAVTIME +
##     BLUEBOOK + TIF + OLDCLAIM + MVR_PTS + CAR_AGE + HOMEKIDS_BIN +
##     OLDCLAIM_BIN + HOME_OWN + PARENT1_Yes + EDUCATION_0 + MSTATUS_Yes +
##     INCOME_BIN_0 + CAR_USE_Commercial + CAR_TYPE_z_SUV + `CAR_TYPE_Sports Car` +
##     CAR_TYPE_Van + `CAR_TYPE_Panel Truck` + CAR_TYPE_Pickup +
##     REVOKED_Yes + `URBANICITY_z_Highly Rural/ Rural`, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4343  -0.7295  -0.4164   0.6618   3.0583
```

```
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -8.189e-01  2.524e-01  -3.244 0.001179
## KIDSDRIV                             2.839e-01  7.216e-02   3.934 8.36e-05
## HOMEKIDS                            -1.151e-01  6.405e-02  -1.797 0.072324
## INCOME                              -7.705e-06  1.114e-06  -6.919 4.56e-12
## TRAVTIME                             1.546e-02  2.284e-03   6.770 1.29e-11
## BLUEBOOK                            -2.483e-05  5.813e-06  -4.272 1.94e-05
## TIF                                 -5.009e-02  9.014e-03  -5.557 2.74e-08
## OLDCLAIM                            -1.688e-05  5.125e-06  -3.295 0.000986
## MVR_PTS                              9.799e-02  1.737e-02   5.640 1.70e-08
## CAR_AGE                             -2.254e-02  7.431e-03  -3.033 0.002418
## HOMEKIDS_BIN                        -4.775e-01  1.634e-01  -2.923 0.003469
## OLDCLAIM_BIN                         5.125e-01  9.612e-02   5.332 9.72e-08
## HOME_OWN                             2.948e-01  9.380e-02   3.143 0.001675
## PARENT1_Yes                          2.877e-01  1.471e-01   1.957 0.050396
## EDUCATION_0                          3.491e-01  1.101e-01   3.170 0.001525
## MSTATUS_Yes                         -4.965e-01  1.073e-01  -4.626 3.73e-06
## INCOME_BIN_0                        -1.427e-01  9.601e-02  -1.487 0.137125
## CAR_USE_Commercial                   9.426e-01  8.878e-02  10.618  < 2e-16
## CAR_TYPE_z_SUV                       6.484e-01  1.022e-01   6.347 2.20e-10
## `CAR_TYPE_Sports Car`                8.281e-01  1.322e-01   6.265 3.72e-10
## CAR_TYPE_Van                         5.110e-01  1.457e-01   3.508 0.000452
## `CAR_TYPE_Panel Truck`               5.050e-01  1.722e-01   2.932 0.003366
## CAR_TYPE_Pickup                      3.687e-01  1.196e-01   3.083 0.002052
## REVOKED_Yes                          9.009e-01  1.127e-01   7.997 1.28e-15
## `URBANICITY_z_Highly Rural/ Rural` -2.213e+00  1.381e-01 -16.028  < 2e-16
##
## (Intercept)                         **
## KIDSDRIV                            ***
## HOMEKIDS                            .
## INCOME                              ***
## TRAVTIME                            ***
## BLUEBOOK                            ***
## TIF                                 ***
## OLDCLAIM                            ***
## MVR_PTS                             ***
## CAR_AGE                             **
## HOMEKIDS_BIN                        **
## OLDCLAIM_BIN                        ***
## HOME_OWN                            **
## PARENT1_Yes                         .
## EDUCATION_0                         **
## MSTATUS_Yes                         ***
## INCOME_BIN_0
## CAR_USE_Commercial                  ***
## CAR_TYPE_z_SUV                      ***
## `CAR_TYPE_Sports Car`               ***
## CAR_TYPE_Van                        ***
## `CAR_TYPE_Panel Truck`              **
## CAR_TYPE_Pickup                     **
## REVOKED_Yes                         ***
## `URBANICITY_z_Highly Rural/ Rural` ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6218.4  on 5386  degrees of freedom
## Residual deviance: 4903.3  on 5362  degrees of freedom
## AIC: 4953.3
##
## Number of Fisher Scoring iterations: 5
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1570  358
##          1  130  251
##
##                Accuracy : 0.7887
##                  95% CI : (0.7714, 0.8051)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : 2.913e-09
##
##                   Kappa : 0.3815
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4122
##             Specificity : 0.9235
##          Pos Pred Value : 0.6588
##          Neg Pred Value : 0.8143
##              Prevalence : 0.2638
##          Detection Rate : 0.1087
##    Detection Prevalence : 0.1650
##       Balanced Accuracy : 0.6678
##
##        'Positive' Class : 1
##
```




So the model perfoms well on the test dataset. Similar transformations needs to be performed on new dataset and predict the car crash.

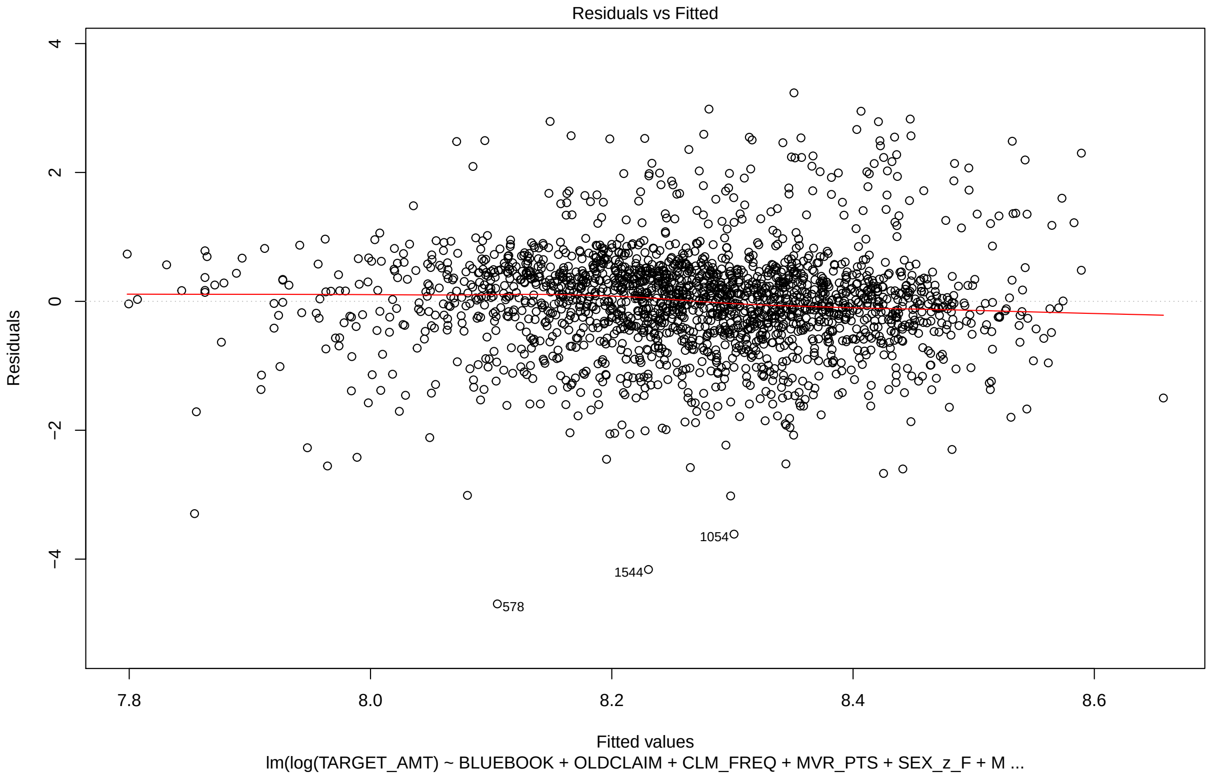
### 1.4.2 TARGET_AMT Model

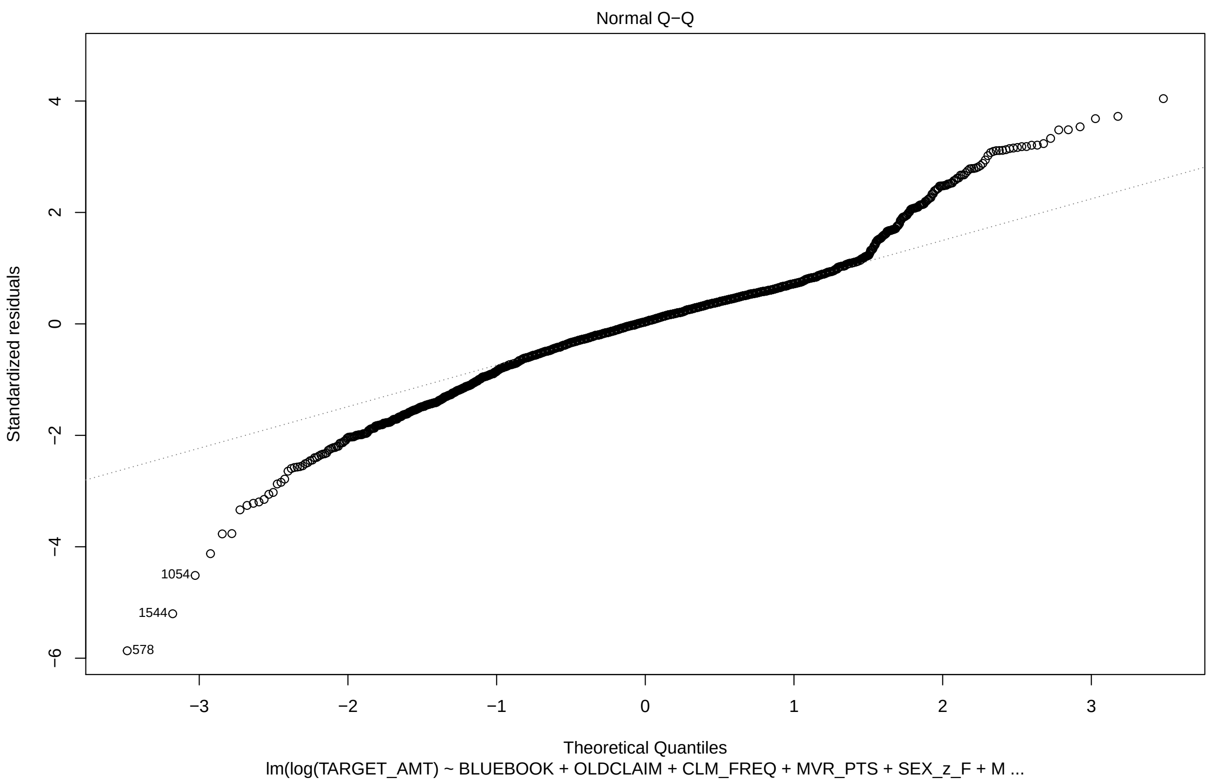
We have build basic model, stepwise model, regsubsets, ridge regression and regression splines model. By comparing ll the models, we can see stepwise model and regression splines model are performing better. However, all the models seems to do fairly bad. As the TARGET_AMT is fairly complex, I'll select the general linear regression with reduced variables.

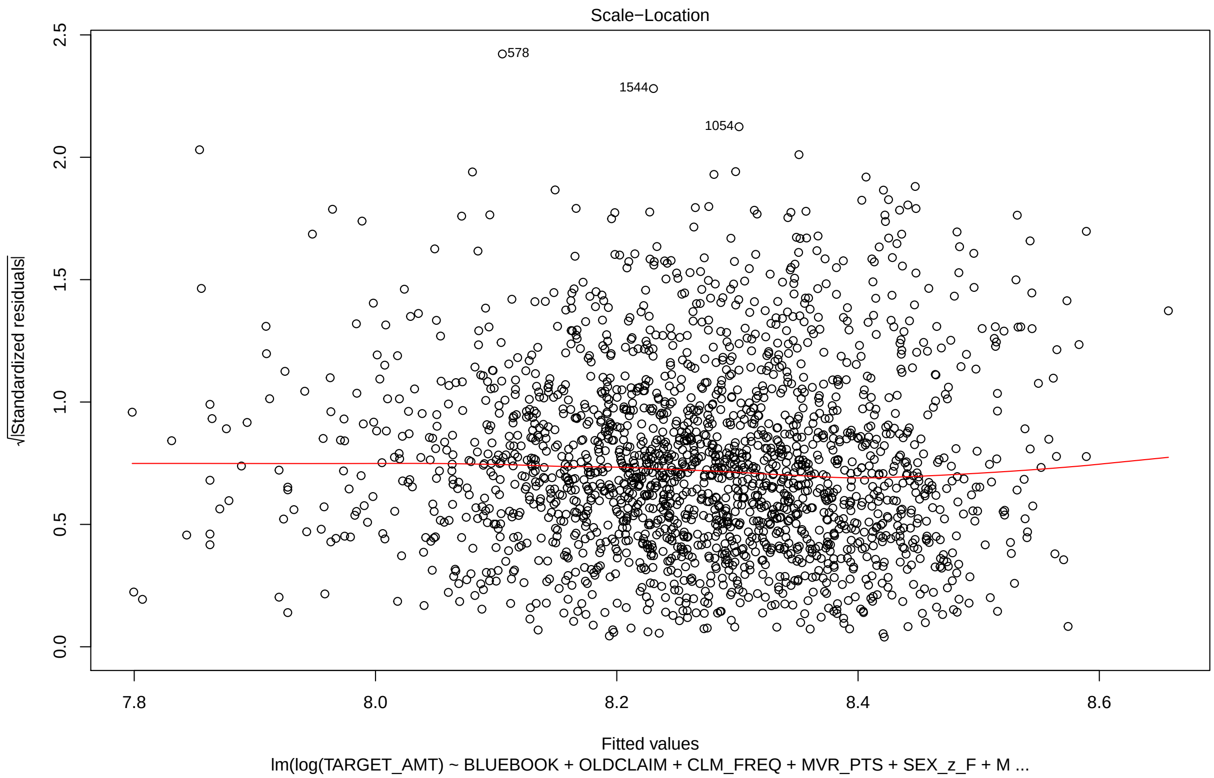
```
## [1] "Adjusted R2: 0.0220804007438259"
```

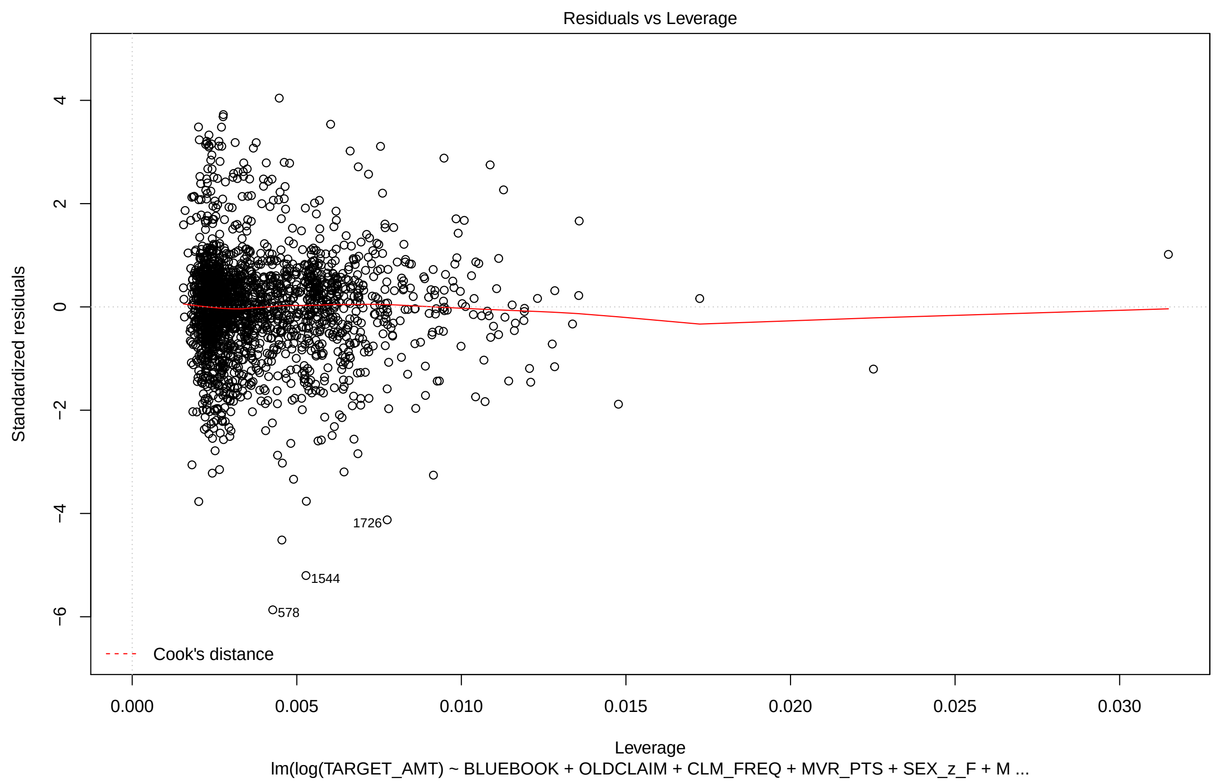
```
## [1] "F-statistic: 7.53821969790011"
```

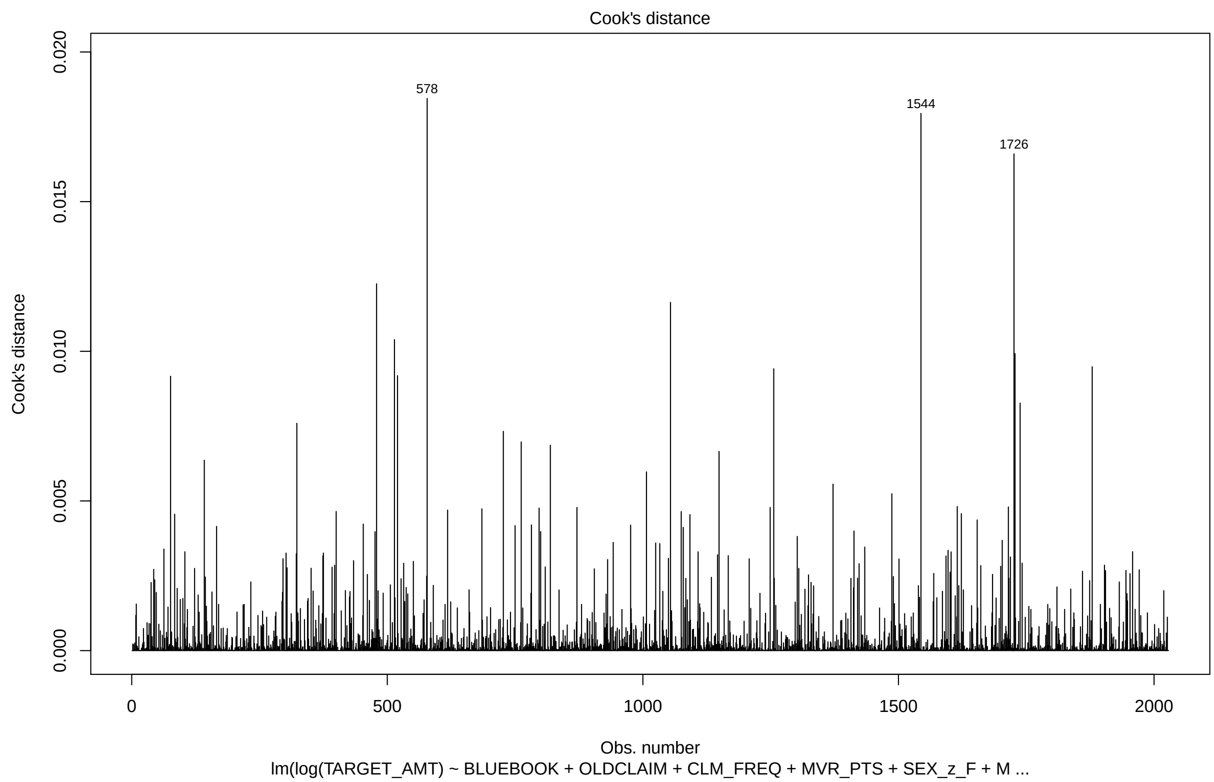
```
## [1] "RMSE: 8447.76961249729"
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(TARGET_AMT) ~ BLUEBOOK + OLDCLAIM + CLM_FREQ + MVR_PTS + SEX_z_F + M ...
1054
1544
578


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(TARGET_AMT) ~ BLUEBOOK + OLDCLAIM + CLM_FREQ + MVR_PTS + SEX_z_F + M ...
1054
1544
578


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
7.8
8.0
8.2
8.4
8.6
578
1544
1054
Fitted values
lm(log(TARGET_AMT) ~ BLUEBOOK + OLDCLAIM + CLM_FREQ + MVR_PTS + SEX_z_F + M ...


Residuals vs Leverage
Standardized residuals
−6
−4
−2
0
2
4
0.000
0.005
0.010
0.015
0.020
0.025
0.030
1726
1544
578
Cook's distance
Leverage
lm(log(TARGET_AMT) ~ BLUEBOOK + OLDCLAIM + CLM_FREQ + MVR_PTS + SEX_z_F + M ...




## 1.5 Prediction of evaluation dataset

Finally we will predict the values of evaluation dataset using the models which we freezed.

### 1.5.1 Target Flag

```
## predicted_model11
##    0    1
## 1928  381
```

We are predicting there will be around 381 crashes.

### 1.5.2 Target Amt

```
## [1] 2544.552
```

## 1.6 Summary

1. We have performed data cleaning on the necessary columsn.
2. Performed a detailed exploratory data analysis.
3. Transformed the variables and added additional features.
4. Build various models for predicting TARGET_FLAG and TARGET_AMT.
5. Evaluated various metrics on the dataset and predicted the evaluation datasets.