DATA 621: BUSINESS ANALYTICS AND DATA MINING HOMEWORK#4: LOGISTIC REGRESSION

Group 2 - Gabriel Campos, Melissa Bowman, Alexander Khaykin, & Jennifer Abinette

Last edited December 03, 2023

Contents

Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

Deliverables

- 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.
- Assigned predictions (probabilities, classifications, cost) for the evaluation data set. Use 0.5 threshold.
- Include your R statistical programming code in an Appendix.

Write Up:

1. DATA EXPLORATION (25 Points)

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

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

2. DATA PREPARATION (25 Points)

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

a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS (25 Points)

Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS (25 Points)

Decide on the criteria for selecting the best multiple linear regression model and the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2 , (c) F-statistic, and (d) residual plots. For the binary logistic regression model, will you use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.

1 DATA EXPLORATION & PREPARATION

1.1 Import Data

1.1.1 Training Dataset

```
df_insur_train <-
  read.csv(paste0(url_git,"insurance_training_data.csv"))

head(df_insur_train)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14                No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M           PhD  Professional       14    Private  $14,230
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial  $14,940
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private   $4,010
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private  $15,440
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private  $18,000
```

```
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial  $17,430
##   TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11     Minivan     yes   $4,461        2      No       3      18
## 2   1     Minivan     yes       $0        0      No       0       1
## 3   4       z_SUV      no  $38,690        2      No       3      10
## 4   7     Minivan     yes       $0        0      No       0       6
## 5   1       z_SUV      no  $19,217        2     Yes       3      17
## 6   1 Sports Car      no       $0        0      No       0       7
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

```r
dim(df_insur_train)
```

```
## [1] 8161   26
```

In the training dataset, there are 8,161 rows and 26 columns. We will remove the INDEX column because it is a unique identifier and will not be used.The two outcome variables are:

- TARGET_FLAG - a 0/1 variable that indicates if a insurance client has been in a car accident
- TARGET_AMT - a numeric variable that of insurance claim payout per car accident

```r
df_insur_train <- df_insur_train %>%
  select(-INDEX)
```

1.1.2 Evaluation Dataset

```r
df_insur_eval <-
  read.csv(paste0(url_git,"insurance-evaluation-data.csv"))

head(df_insur_eval)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ  INCOME PARENT1
## 1     3          NA         NA        0  48        0  11 $52,881      No
## 2     9          NA         NA        1  40        1  11 $50,815     Yes
## 3    10          NA         NA        0  44        2  12 $43,486     Yes
## 4    18          NA         NA        0  35        2  NA $21,204     Yes
## 5    21          NA         NA        0  59        0  12 $87,460      No
## 6    30          NA         NA        0  46        0  14               No
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M     Bachelors       Manager       26    Private  $21,970
## 2       $0    z_No   M z_High School       Manager       21    Private  $18,930
## 3       $0    z_No z_F z_High School z_Blue Collar       30 Commercial   $5,900
## 4       $0    z_No   M z_High School      Clerical       74    Private   $9,230
## 5       $0    z_No   M z_High School       Manager       45    Private  $15,420
```

```
## 6 $207,519     Yes   M     Bachelors  Professional        7 Commercial  $25,660
##   TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1   1         Van     yes       $0        0      No       2      10
## 2   6     Minivan      no   $3,295        1      No       2       1
## 3  10       z_SUV      no       $0        0      No       0      10
## 4   6      Pickup      no       $0        0     Yes       0       4
## 5   1     Minivan     yes  $44,857        2      No       4       1
## 6   1 Panel Truck      no   $2,119        1      No       2      12
##              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
```

```
df_insur_eval <- df_insur_eval %>%
  select(-INDEX)
```

- There are 12 variables with discrete values and 13 variables with continuous values

1.2 Transformations

- We noticed that there are characters in several of the columns that need to be cleaned up before the analysis. These will be removed and if necessary the variable will be converted to the appropriate data type.

```
df_insur_train <- df_insur_train %>%
  mutate(INCOME = gsub("\\$", "", INCOME), HOME_VAL = gsub("\\$", "", HOME_VAL),
         BLUEBOOK = gsub("\\$", "", BLUEBOOK), OLDCLAIM = gsub("\\$", "",
                                                               OLDCLAIM)) %>%
  mutate(INCOME = gsub(",", "", INCOME), HOME_VAL = gsub(",", "", HOME_VAL),
         BLUEBOOK = gsub(",", "", BLUEBOOK), OLDCLAIM = gsub(",", "",
                                                             OLDCLAIM)) %>%
  mutate(INCOME = as.numeric(INCOME), HOME_VAL = as.numeric(HOME_VAL),
         BLUEBOOK = as.numeric(BLUEBOOK), OLDCLAIM = as.numeric(OLDCLAIM))
```

```
df_insur_train <- df_insur_train %>%
  mutate(MSTATUS = gsub("z_","", MSTATUS), SEX = gsub("z_","", SEX),
         EDUCATION = gsub("z_","", EDUCATION), JOB = gsub("z_","", JOB),
         CAR_TYPE = gsub("z_","", CAR_TYPE), URBANICITY = gsub("z_","",
                                                               URBANICITY))
```

- Applied same to evaluation data

```
df_insur_eval <- df_insur_eval %>%
  mutate(INCOME = gsub("\\$", "", INCOME), HOME_VAL = gsub("\\$", "", HOME_VAL),
         BLUEBOOK = gsub("\\$", "", BLUEBOOK), OLDCLAIM = gsub("\\$", "",
                                                               OLDCLAIM)) %>%
  mutate(INCOME = gsub(",", "", INCOME), HOME_VAL = gsub(",", "", HOME_VAL),
         BLUEBOOK = gsub(",", "", BLUEBOOK), OLDCLAIM = gsub(",", "",
```

```
                                                            OLDCLAIM)) %>%
  mutate(INCOME = as.numeric(INCOME), HOME_VAL = as.numeric(HOME_VAL),
         BLUEBOOK = as.numeric(BLUEBOOK), OLDCLAIM = as.numeric(OLDCLAIM))
```

```
df_insur_eval <- df_insur_eval %>%
  mutate(MSTATUS = gsub("z_","", MSTATUS), SEX = gsub("z_","", SEX),
         EDUCATION = gsub("z_","", EDUCATION), JOB = gsub("z_","", JOB),
         CAR_TYPE = gsub("z_","", CAR_TYPE), URBANICITY = gsub("z_","",
                                                              URBANICITY))
```

- We will recode JOB into White Collar(Clerical, Doctor, Lawyer, Manager, and Professional), Blue Collar, and None (Student, Homemaker)

```
df_insur_train <- df_insur_train %>%
  mutate(JOB = ifelse(JOB=="Blue Collar", "Blue Collar",
                      ifelse(JOB=="Student" | JOB=="Home Maker",
                             "None",
                             "White Collar")))
df_insur_eval <- df_insur_eval %>%
  mutate(JOB = ifelse(JOB=="Blue Collar", "Blue Collar",
                      ifelse(JOB=="Student" | JOB=="Home Maker",
                             "None", "White Collar")))
```

- We will also recode KIDSDRIV into a 0 or 1 (1+kids driving). Because there are a lot more insurance claims without kids dring than with kids driving.

```
df_insur_train <- df_insur_train %>%
  mutate(KIDSDRIV = ifelse(KIDSDRIV >= 1, 1, 0))

df_insur_eval <- df_insur_eval %>%
  mutate(KIDSDRIV = ifelse(KIDSDRIV >= 1, 1, 0))
```

- Also, recode the yes/mo labels for marital status, parent status, red car, and revoked license variables as 1/0.

```
df_insur_train <- df_insur_train %>%
  mutate(MSTATUS = ifelse(MSTATUS == "No", "0", "1"),
         PARENT1 = ifelse(PARENT1 == "No", "0", "1"),
         RED_CAR = ifelse(RED_CAR == "no", "0", "1"),
         REVOKED = ifelse(REVOKED == "No", "0", "1"))
df_insur_eval <- df_insur_eval %>%
  mutate(MSTATUS = ifelse(MSTATUS == "No", "0", "1"),
         PARENT1 = ifelse(PARENT1 == "No", "0", "1"),
         RED_CAR = ifelse(RED_CAR == "no", "0", "1"),
         REVOKED = ifelse(REVOKED == "No", "0", "1"))
```

- Lastly we will shorten the lables for Urbanicity and Turn Education into a factor with "< Highschool" as the reference variable.

```r
df_insur_train <- df_insur_train %>%
  mutate(URBANICITY = ifelse(URBANICITY == "Highly Urban/ Urban",
                             "Urban", "Rural")) %>%
  mutate(EDUCATION = factor(EDUCATION,levels = c("<High School",
                                                  "High School",
                                                  "Bachelors",
                                                  "Masters",
                                                  "PhD")))

df_insur_eval <- df_insur_eval %>%
  mutate(URBANICITY = ifelse(URBANICITY == "Highly Urban/ Urban",
                             "Urban", "Rural")) %>%
  mutate(EDUCATION = factor(EDUCATION,levels = c("<High School",
                                                  "High School",
                                                  "Bachelors",
                                                  "Masters",
                                                  "PhD")))
```

1.3 Missing Data Imputation

1.3.1 Training Dataset

```r
#loop to count the NAs for each column
for (i in colnames(df_insur_train)){
  print(paste(i,"  ", sum(is.na(df_insur_train[,i])),sep = ""))
}
```

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

- There are NAs in three variable columns, 6 in AGE, 454 in YOJ (Years on the job) , and 510 in CAR_AGE. For these variable we will impute the median so as not to create an over fitting problem. Also, there was an irrational value of negative 3 for CAR_AGE, we replaced it with zero.

```
df_insur_train <- df_insur_train %>%
  mutate(AGE = ifelse(is.na(AGE),
                      median(AGE, na.rm = TRUE),
                      AGE), YOJ = ifelse(is.na(YOJ),
                                         median(YOJ, na.rm = TRUE), YOJ),
         CAR_AGE = ifelse(is.na(CAR_AGE),
                          median(CAR_AGE, na.rm = TRUE), CAR_AGE),
         HOME_VAL = ifelse(is.na(HOME_VAL),
                           median(HOME_VAL,
                                  na.rm = TRUE), HOME_VAL),
         INCOME = ifelse(is.na(INCOME),
                         median(INCOME, na.rm = TRUE),
                         INCOME)) %>%
  mutate(CAR_AGE = ifelse(CAR_AGE < 0, 0, CAR_AGE))

summary(df_insur_train$CAR_AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   8.000   8.308  12.000  28.000
```

1.3.2 Evaluation Dataset

```
#loop to count the NAs for each column
for (i in colnames(df_insur_eval)){
  print(paste(i,"  ", sum(is.na(df_insur_eval[,i])),sep = ""))
}
```

```
## [1] "TARGET_FLAG  2141"
## [1] "TARGET_AMT  2141"
## [1] "KIDSDRIV  0"
## [1] "AGE  1"
## [1] "HOMEKIDS  0"
## [1] "YOJ  94"
## [1] "INCOME  125"
## [1] "PARENT1  0"
## [1] "HOME_VAL  111"
## [1] "MSTATUS  0"
## [1] "SEX  0"
## [1] "EDUCATION  0"
## [1] "JOB  0"
## [1] "TRAVTIME  0"
## [1] "CAR_USE  0"
## [1] "BLUEBOOK  0"
## [1] "TIF  0"
## [1] "CAR_TYPE  0"
## [1] "RED_CAR  0"
## [1] "OLDCLAIM  0"
```

```
## [1] "CLM_FREQ  0"
## [1] "REVOKED  0"
## [1] "MVR_PTS  0"
## [1] "CAR_AGE  129"
## [1] "URBANICITY  0"
```

- There are NAs in five variable columns, 1 in AGE, 94 in YOJ (Years on the job) , 125 in INCOME, 111 HOME_VAL, and 129 in CAR_AGE. For these variable we will impute the median so as not to create an over fitting problem.

```r
df_insur_eval <- df_insur_eval %>%
  mutate(AGE = ifelse(is.na(AGE), median(AGE, na.rm = TRUE),
                      AGE), YOJ = ifelse(is.na(YOJ),
                                         median(YOJ, na.rm = TRUE), YOJ),
         CAR_AGE = ifelse(is.na(CAR_AGE), median(CAR_AGE, na.rm = TRUE),
                          CAR_AGE)) %>%
  mutate(INCOME = ifelse(is.na(INCOME), median(INCOME,
                                               na.rm = TRUE), INCOME),
         HOME_VAL = ifelse(is.na(HOME_VAL),
                           median(HOME_VAL, na.rm = TRUE), YOJ))

summary(df_insur_eval$CAR_AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   8.000   8.172  12.000  26.000
```

1.4 Exploratory Data Analysis

Summary statistics for the numeric variables:

```r
df_insur_train %>%
  select(TARGET_AMT, AGE, YOJ, INCOME, HOME_VAL, TRAVTIME, BLUEBOOK, TIF,
         OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, HOMEKIDS) %>%
  describe()
```

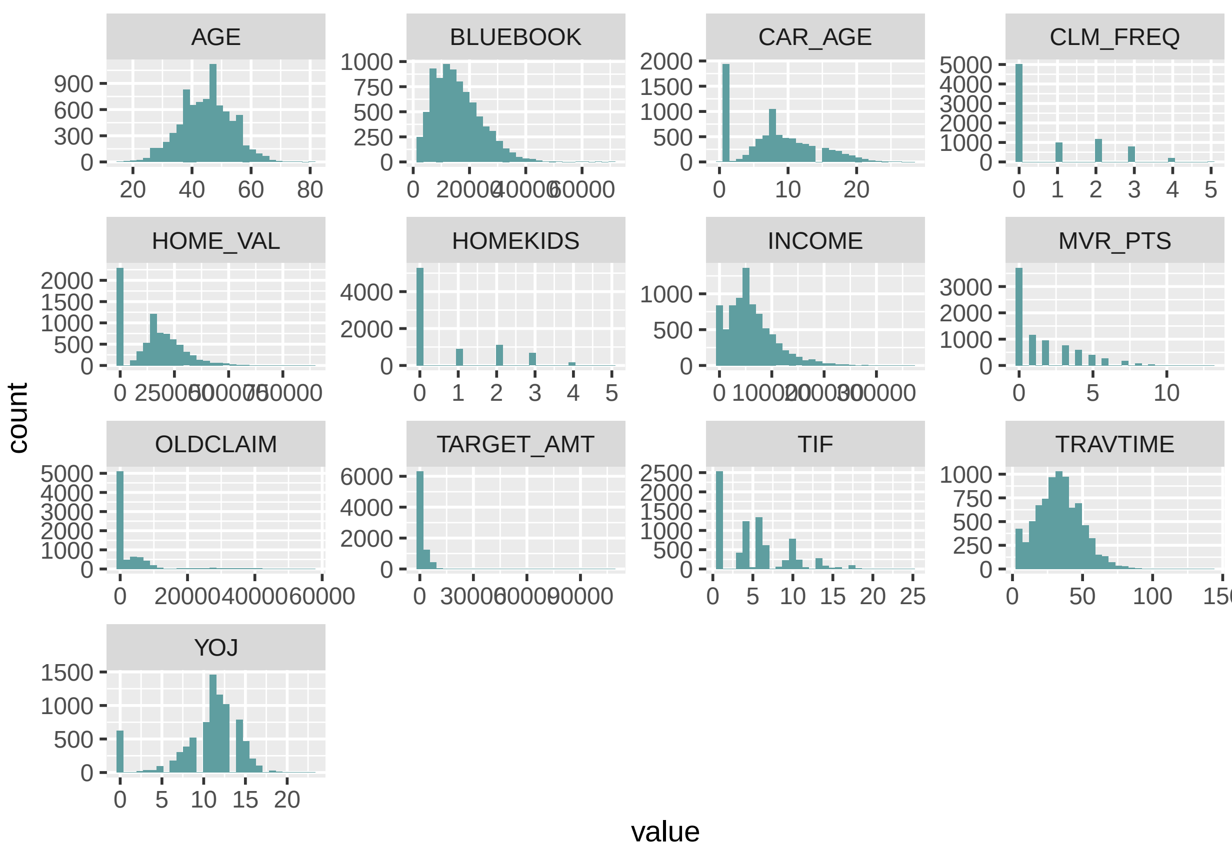
```
##            vars    n      mean        sd median   trimmed       mad  min
## TARGET_AMT    1 8161   1504.32   4704.03      0    593.71      0.00    0
## AGE           2 8161     44.79      8.62     45     44.83      8.90   16
## YOJ           3 8161     10.53      3.98     11     11.08      2.97    0
## INCOME        4 8161  61468.96  46291.83  54028  56557.35  38976.07    0
## HOME_VAL      5 8161 155225.07 125407.35 161160 145061.93 131525.89    0
## TRAVTIME      6 8161     33.49     15.91     33     33.00     16.31    5
## BLUEBOOK      7 8161  15709.90   8419.73  14440  15036.89   8450.82 1500
## TIF           8 8161      5.35      4.15      4      4.84      4.45    1
## OLDCLAIM      9 8161   4037.08   8777.14      0   1719.29      0.00    0
## CLM_FREQ     10 8161      0.80      1.16      0      0.59      0.00    0
## MVR_PTS      11 8161      1.70      2.15      1      1.31      1.48    0
## CAR_AGE      12 8161      8.31      5.52      8      7.96      5.93    0
## HOMEKIDS     13 8161      0.72      1.12      0      0.50      0.00    0
##                 max    range  skew kurtosis    se
## TARGET_AMT 107586.1 107586.1  8.71   112.29 52.07
## AGE            81.0     65.0 -0.03    -0.06  0.10
```

```
## YOJ            23.0     23.0 -1.26      1.45    0.04
## INCOME     367030.0 367030.0  1.24      2.45  512.43
## HOME_VAL   885282.0 885282.0  0.49      0.16 1388.20
## TRAVTIME      142.0    137.0  0.45      0.66    0.18
## BLUEBOOK    69740.0  68240.0  0.79      0.79   93.20
## TIF            25.0     24.0  0.89      0.42    0.05
## OLDCLAIM    57037.0  57037.0  3.12      9.86   97.16
## CLM_FREQ        5.0      5.0  1.21      0.28    0.01
## MVR_PTS        13.0     13.0  1.35      1.38    0.02
## CAR_AGE        28.0     28.0  0.30     -0.60    0.06
## HOMEKIDS        5.0      5.0  1.34      0.65    0.01
```

- The skewness and Kurtosis values for the outcome variable TARGET_AMT strongly suggests that the distribution is likely not normal.

```r
df_insur_train %>%
  select(TARGET_AMT, AGE, YOJ, INCOME, HOME_VAL, TRAVTIME, BLUEBOOK,
         TIF, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, HOMEKIDS) %>%
  gather() %>%
  ggplot(aes(x = value)) +
  geom_histogram(fill = "cadetblue") +
  facet_wrap(~key, scales = "free")
```

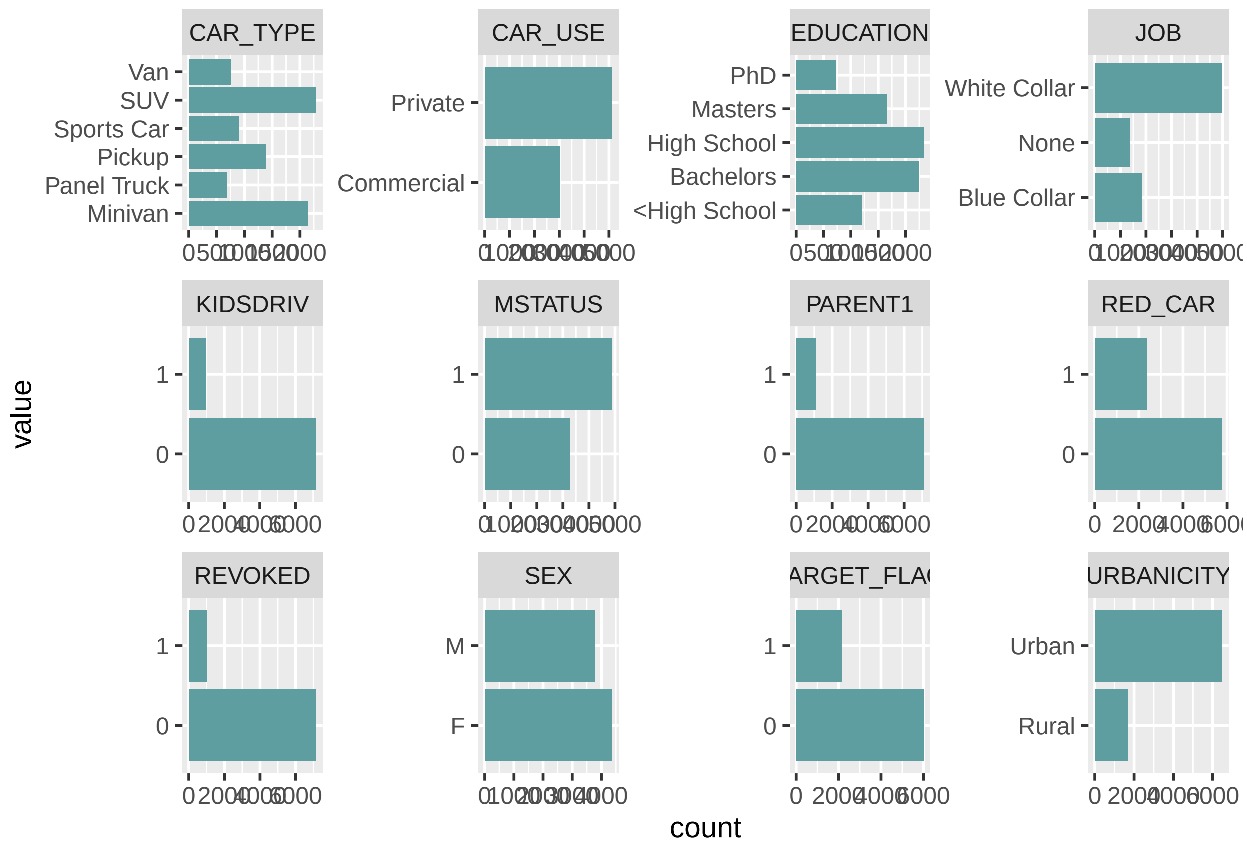
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- The histogram for TARGET_AMT, CAR_AGE, CLM_FREQ,HOME_VAL, INCOME, MRV_PTS, OLDCLAIM, and TIF are clearly not normally distributed and will need to be transformed if the residuals are not normally distributed.
- We will explore the proportions of the discrete variables.

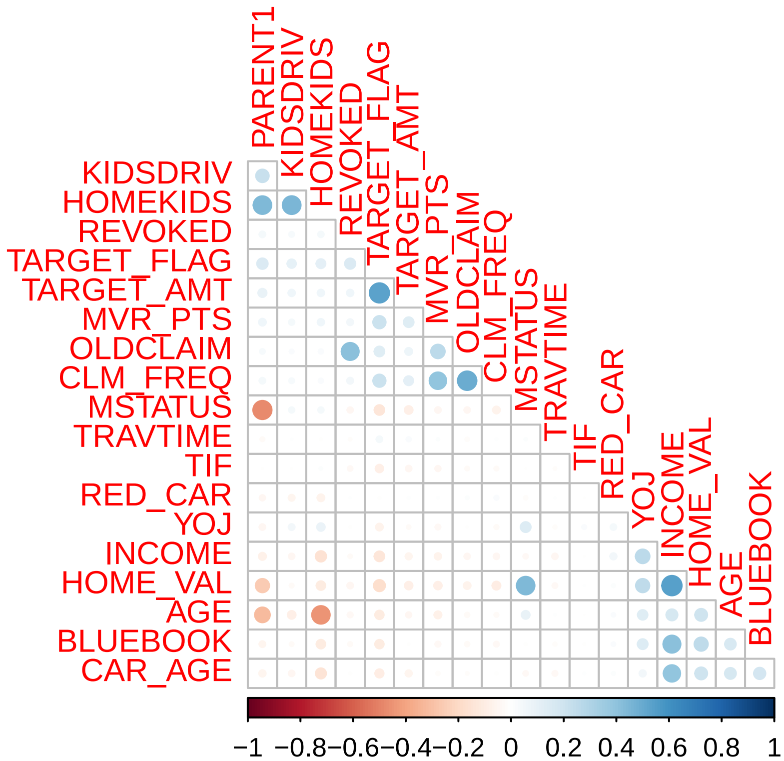
```
df_insur_train %>%
  select(TARGET_FLAG, KIDSDRIV, PARENT1, MSTATUS, SEX, EDUCATION,
         JOB, CAR_USE, CAR_TYPE, RED_CAR, REVOKED, URBANICITY) %>%
  gather() %>%
  ggplot(aes(x = value)) +
  geom_bar(fill = "cadetblue") +
  coord_flip()+
  facet_wrap(~key, scales = "free")
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```



- To check for collinearity through the correlation of the variables

```
mat <- df_insur_train %>%
  select(-CAR_TYPE, -CAR_USE, -EDUCATION, -JOB, -SEX, -URBANICITY) %>%
  mutate(PARENT1 = as.numeric(PARENT1), MSTATUS = as.numeric(MSTATUS),
         RED_CAR = as.numeric(RED_CAR), REVOKED = as.numeric(REVOKED)) %>%
  cor()
corrplot(mat, method = "circle", diag = FALSE, order ="hclust", type = "lower")
```



- We do not seem to have very much concern for high collinearity at this point.

2 BUILD & SELECT MODELS

2.1 Logistic Regression Models

2.1.1 Model with All Predictors - AIC 7416.5

- First, let's take a look at a binary logistic model with all variables included:

```
log_mod <- glm(TARGET_FLAG ~., data = df_insur_train[,-2],
               family = binomial(link = "logit"))
summary(log_mod)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
##     data = df_insur_train[, -2])
##
## Coefficients:
##                          Estimate    Std. Error z value            Pr(>|z|)
## (Intercept)          -2.3323791351  0.2756854844  -8.460 < 0.0000000000000002
## KIDSDRIV              0.6199891324  0.0956338656   6.483 0.000000000089949034
```

```
## AGE                  -0.0025290782  0.0039624049  -0.638            0.523299
## HOMEKIDS              0.0610346583  0.0363202203   1.680            0.092868
## YOJ                  -0.0108523166  0.0085466125  -1.270            0.204163
## INCOME               -0.0000044975  0.0000010535  -4.269 0.000019616658829300
## PARENT11              0.3238899472  0.1092512973   2.965            0.003030
## HOME_VAL             -0.0000012332  0.0000003376  -3.653            0.000260
## MSTATUS1             -0.5167577681  0.0832886764  -6.204 0.000000000548996678
## SEXM                  0.0707646899  0.1112038657   0.636            0.524548
## EDUCATIONHigh School -0.0453124396  0.0931034098  -0.487            0.626478
## EDUCATIONBachelors   -0.5324093228  0.1081694077  -4.922 0.000000856662454844
## EDUCATIONMasters     -0.4865479323  0.1404739469  -3.464            0.000533
## EDUCATIONPhD         -0.5073188801  0.1740096425  -2.915            0.003552
## JOBNone              -0.1190215906  0.1157288580  -1.028            0.303737
## JOBWhite Collar      -0.1562558912  0.0892152355  -1.751            0.079869
## TRAVTIME              0.0150807603  0.0018736651   8.049 0.000000000000000836
## CAR_USEPrivate       -0.7763930961  0.0849801166  -9.136 < 0.0000000000000002
## BLUEBOOK             -0.0000216475  0.0000052350  -4.135 0.000035468950972329
## TIF                  -0.0546223938  0.0073114823  -7.471 0.000000000000079728
## CAR_TYPEPanel Truck   0.5604601495  0.1581493487   3.544            0.000394
## CAR_TYPEPickup        0.5312823335  0.0999114706   5.318 0.000000105184819385
## CAR_TYPESports Car    0.9926036876  0.1292205386   7.681 0.000000000000015727
## CAR_TYPESUV           0.7502081312  0.1107145597   6.776 0.000000000012350025
## CAR_TYPEVan           0.6080520183  0.1254046825   4.849 0.000001242615379541
## RED_CAR1             -0.0210562962  0.0859125458  -0.245            0.806387
## OLDCLAIM             -0.0000142092  0.0000038840  -3.658            0.000254
## CLM_FREQ              0.1947632487  0.0284057980   6.856 0.000000000007058730
## REVOKED1              0.9052577132  0.0907523206   9.975 < 0.0000000000000002
## MVR_PTS               0.1192356659  0.0135679751   8.788 < 0.0000000000000002
## CAR_AGE              -0.0010329598  0.0075172906  -0.137            0.890706
## URBANICITYUrban       2.3223762321  0.1123207354  20.676 < 0.0000000000000002
##
## (Intercept)          ***
## KIDSDRIV             ***
## AGE
## HOMEKIDS             .
## YOJ
## INCOME               ***
## PARENT11             **
## HOME_VAL             ***
## MSTATUS1             ***
## SEXM
## EDUCATIONHigh School
## EDUCATIONBachelors   ***
## EDUCATIONMasters     ***
## EDUCATIONPhD         **
## JOBNone
## JOBWhite Collar      .
## TRAVTIME             ***
## CAR_USEPrivate       ***
## BLUEBOOK             ***
## TIF                  ***
## CAR_TYPEPanel Truck  ***
## CAR_TYPEPickup       ***
## CAR_TYPESports Car   ***
```

```
## CAR_TYPESUV          ***
## CAR_TYPEVan          ***
## RED_CAR1
## OLDCLAIM             ***
## CLM_FREQ             ***
## REVOKED1             ***
## MVR_PTS              ***
## CAR_AGE
## URBANICITYUrban      ***
## ---
## 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: 7352.5  on 8129  degrees of freedom
## AIC: 7416.5
##
## Number of Fisher Scoring iterations: 5
```

```
vif(log_mod)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## KIDSDRIV   1.325444  1        1.151279
## AGE        1.437446  1        1.198935
## HOMEKIDS   2.101442  1        1.449635
## YOJ        1.447790  1        1.203242
## INCOME     2.351147  1        1.533345
## PARENT1    1.942979  1        1.393908
## HOME_VAL   1.831162  1        1.353204
## MSTATUS    2.059943  1        1.435250
## SEX        3.677546  1        1.917693
## EDUCATION  3.369170  4        1.163966
## JOB        2.969760  2        1.312745
## TRAVTIME   1.038168  1        1.018905
## CAR_USE    2.117569  1        1.455187
## BLUEBOOK   2.178258  1        1.475892
## TIF        1.008117  1        1.004050
## CAR_TYPE   6.204570  5        1.200248
## RED_CAR    1.831573  1        1.353356
## OLDCLAIM   1.646459  1        1.283144
## CLM_FREQ   1.465650  1        1.210640
## REVOKED    1.313484  1        1.146073
## MVR_PTS    1.158854  1        1.076501
## CAR_AGE    2.011633  1        1.418321
## URBANICITY 1.133593  1        1.064703
```

- The full model above gives us an AIC of 7416.5, and indicates that using all the predictors does a better job predicting whether a person was in a car crash (TARGET_FLAG) than a null model with only the intercept (Residual deviance is less than the Null deviance).
- The degree of freedom adjusted variance inflation factors suggests that there is no concerning collinearity because all of the values are less than 3.

2.1.2 Model with Strongest Significant Predictors - AIC 8376.4

- Next, let's explore a model with the predictors with the lowest p-values. As shown above, there are 4 variable coefficients with significant p-values less than 0.0000000000000002 including CAR_USE, REVOKED, MVR_PTS, and URBANICITY.

```
log_mod_2 <- glm(TARGET_FLAG ~CAR_USE + REVOKED + MVR_PTS + URBANICITY  , data = df_insur_train[,-2],
               family = binomial(link = "logit"))
summary(log_mod_2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ CAR_USE + REVOKED + MVR_PTS + URBANICITY,
##     family = binomial(link = "logit"), data = df_insur_train[,
##         -2])
##
## Coefficients:
##                 Estimate Std. Error z value            Pr(>|z|)
## (Intercept)     -2.55313    0.10329  -24.72 <0.0000000000000002 ***
## CAR_USEPrivate  -0.67421    0.05434  -12.41 <0.0000000000000002 ***
## REVOKED1         0.82400    0.07361   11.19 <0.0000000000000002 ***
## MVR_PTS          0.17956    0.01178   15.25 <0.0000000000000002 ***
## URBANICITYUrban  1.69523    0.10220   16.59 <0.0000000000000002 ***
## ---
## 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: 8366.4  on 8156  degrees of freedom
## AIC: 8376.4
##
## Number of Fisher Scoring iterations: 5
```

- The model above gives us an AIC of 8376.4, indicating that our initial model was a better fit given it's lower AIC of 7416.5. Similarly we see that using these 4 predictors does a better job predicting whether a person was in a car crash (TARGET_FLAG) than a null model with only the intercept (Residual deviance is less than the Null deviance).

2.1.3 Backward Elimination Model - AIC 7408.4

- As the model with all the predictors included was a better fit according to the AIC, we will use backward elimination to create an additional model for comparison.

```
log_step <- step(log_mod, direction = "backward", test = "LRT")
```

```
## Start:  AIC=7416.54
## 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
```

```
##
##              Df Deviance    AIC    LRT             Pr(>Chi)
## - CAR_AGE     1   7352.6 7414.6   0.02            0.8907159
## - RED_CAR     1   7352.6 7414.6   0.06            0.8064289
## - SEX         1   7352.9 7414.9   0.40            0.5245599
## - AGE         1   7352.9 7414.9   0.41            0.5232743
## - JOB         2   7355.6 7415.6   3.09            0.2130171
## - YOJ         1   7354.1 7416.1   1.61            0.2042655
## <none>            7352.5 7416.5
## - HOMEKIDS    1   7355.3 7417.3   2.81            0.0936697 .
## - PARENT1     1   7361.3 7423.3   8.80            0.0030112 **
## - HOME_VAL    1   7365.9 7427.9  13.39            0.0002528 ***
## - OLDCLAIM    1   7366.2 7428.2  13.67            0.0002184 ***
## - BLUEBOOK    1   7369.9 7431.9  17.39 0.0000304880874862281 ***
## - INCOME      1   7371.1 7433.1  18.60 0.0000161498084255769 ***
## - EDUCATION   4   7389.7 7445.7  37.15 0.0000001679649376969 ***
## - MSTATUS     1   7390.6 7452.6  38.09 0.0000000006754024180 ***
## - KIDSDRIV    1   7394.3 7456.3  41.73 0.0000000001050153135 ***
## - CLM_FREQ    1   7398.9 7460.9  46.39 0.0000000000096993704 ***
## - TIF         1   7410.4 7472.4  57.88 0.0000000000000278171 ***
## - TRAVTIME    1   7417.5 7479.5  65.01 0.0000000000000007454 ***
## - MVR_PTS     1   7430.4 7492.4  77.85 < 0.00000000000000022 ***
## - CAR_TYPE    5   7441.4 7495.4  88.91 < 0.00000000000000022 ***
## - CAR_USE     1   7437.2 7499.2  84.65 < 0.00000000000000022 ***
## - REVOKED     1   7450.3 7512.3  97.79 < 0.00000000000000022 ***
## - URBANICITY  1   7967.7 8029.7 615.20 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=7414.56
## 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 + URBANICITY
##
##              Df Deviance    AIC    LRT             Pr(>Chi)
## - RED_CAR     1   7352.6 7412.6   0.06            0.8054636
## - SEX         1   7353.0 7413.0   0.41            0.5230998
## - AGE         1   7353.0 7413.0   0.41            0.5221311
## - JOB         2   7355.6 7413.6   3.09            0.2132397
## - YOJ         1   7354.2 7414.2   1.61            0.2048280
## <none>            7352.6 7414.6
## - HOMEKIDS    1   7355.4 7415.4   2.81            0.0936144 .
## - PARENT1     1   7361.4 7421.4   8.80            0.0030116 **
## - HOME_VAL    1   7365.9 7425.9  13.37            0.0002554 ***
## - OLDCLAIM    1   7366.2 7426.2  13.67            0.0002181 ***
## - BLUEBOOK    1   7369.9 7429.9  17.38 0.0000306760531236483 ***
## - INCOME      1   7371.2 7431.2  18.67 0.0000155367421105392 ***
## - MSTATUS     1   7390.7 7450.7  38.10 0.0000000006708113146 ***
## - KIDSDRIV    1   7394.3 7454.3  41.72 0.0000000001051681970 ***
## - EDUCATION   4   7401.4 7455.4  48.87 0.0000000006227139095 ***
## - CLM_FREQ    1   7398.9 7458.9  46.37 0.0000000000097707202 ***
## - TIF         1   7410.5 7470.5  57.90 0.0000000000000275347 ***
## - TRAVTIME    1   7417.6 7477.6  65.00 0.0000000000000007507 ***
```

```
## - MVR_PTS     1   7430.4 7490.4  77.86 < 0.00000000000000022 ***
## - CAR_TYPE    5   7441.6 7493.6  89.00 < 0.00000000000000022 ***
## - CAR_USE     1   7437.2 7497.2  84.65 < 0.00000000000000022 ***
## - REVOKED     1   7450.4 7510.4  97.80 < 0.00000000000000022 ***
## - URBANICITY  1   7967.8 8027.8 615.22 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=7412.62
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC    LRT             Pr(>Chi)
## - SEX         1   7353.0 7411.0   0.35            0.5543969
## - AGE         1   7353.0 7411.0   0.40            0.5248378
## - JOB         2   7355.7 7411.7   3.10            0.2120390
## - YOJ         1   7354.2 7412.2   1.61            0.2041423
## <none>            7352.6 7412.6
## - HOMEKIDS    1   7355.4 7413.4   2.80            0.0943466 .
## - PARENT1     1   7361.4 7419.4   8.82            0.0029784 **
## - HOME_VAL    1   7366.0 7424.0  13.33            0.0002607 ***
## - OLDCLAIM    1   7366.3 7424.3  13.68            0.0002165 ***
## - BLUEBOOK    1   7370.0 7428.0  17.34 0.0000313160289052534 ***
## - INCOME      1   7371.3 7429.3  18.67 0.0000155571751299656 ***
## - MSTATUS     1   7390.7 7448.7  38.09 0.0000000006750666184 ***
## - KIDSDRIV    1   7394.4 7452.4  41.81 0.0000000001006978286 ***
## - EDUCATION   4   7401.6 7453.6  48.95 0.0000000005986125019 ***
## - CLM_FREQ    1   7399.0 7457.0  46.35 0.0000000000099125938 ***
## - TIF         1   7410.5 7468.5  57.88 0.0000000000000278754 ***
## - TRAVTIME    1   7417.6 7475.6  64.99 0.0000000000000007521 ***
## - MVR_PTS     1   7430.5 7488.5  77.84 < 0.00000000000000022 ***
## - CAR_TYPE    5   7441.8 7491.8  89.14 < 0.00000000000000022 ***
## - CAR_USE     1   7437.3 7495.3  84.67 < 0.00000000000000022 ***
## - REVOKED     1   7450.4 7508.4  97.82 < 0.00000000000000022 ***
## - URBANICITY  1   7967.8 8025.8 615.17 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=7410.97
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC    LRT             Pr(>Chi)
## - AGE         1   7353.3 7409.3   0.33            0.5680102
## - JOB         2   7356.1 7410.1   3.12            0.2099388
## - YOJ         1   7354.6 7410.6   1.60            0.2055882
## <none>            7353.0 7411.0
## - HOMEKIDS    1   7355.8 7411.8   2.78            0.0952733 .
## - PARENT1     1   7361.8 7417.8   8.80            0.0030065 **
## - HOME_VAL    1   7366.3 7422.3  13.35            0.0002579 ***
```

```
## - OLDCLAIM    1   7366.6 7422.6  13.68            0.0002171 ***
## - INCOME      1   7371.7 7427.7  18.71 0.0000152245731286608 ***
## - BLUEBOOK    1   7377.0 7433.0  24.08 0.0000009252236446092 ***
## - MSTATUS     1   7391.0 7447.0  38.08 0.0000000006801331761 ***
## - KIDSDRIV    1   7394.6 7450.6  41.61 0.0000000001111456235 ***
## - EDUCATION   4   7401.9 7451.9  48.97 0.0000000005936321975 ***
## - CLM_FREQ    1   7399.4 7455.4  46.42 0.0000000000095617383 ***
## - TIF         1   7410.8 7466.8  57.87 0.0000000000000279462 ***
## - TRAVTIME    1   7418.0 7474.0  65.06 0.0000000000000007283 ***
## - MVR_PTS     1   7430.8 7486.8  77.79 < 0.00000000000000022 ***
## - CAR_USE     1   7437.8 7493.8  84.80 < 0.00000000000000022 ***
## - REVOKED     1   7451.0 7507.0  98.00 < 0.00000000000000022 ***
## - CAR_TYPE    5   7460.7 7508.7 107.75 < 0.00000000000000022 ***
## - URBANICITY  1   7968.5 8024.5 615.50 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=7409.29
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC    LRT              Pr(>Chi)
## - JOB         2   7356.4 7408.4   3.12             0.2106325
## - YOJ         1   7355.2 7409.2   1.90             0.1680004
## <none>            7353.3 7409.3
## - HOMEKIDS    1   7357.4 7411.4   4.08             0.0432958 *
## - PARENT1     1   7362.6 7416.6   9.26             0.0023407 **
## - OLDCLAIM    1   7366.9 7420.9  13.63             0.0002223 ***
## - HOME_VAL    1   7367.0 7421.0  13.73             0.0002110 ***
## - INCOME      1   7371.8 7425.8  18.54 0.0000166002232339155 ***
## - BLUEBOOK    1   7378.2 7432.2  24.89 0.0000006060959024278 ***
## - MSTATUS     1   7391.3 7445.3  38.03 0.0000000006979896189 ***
## - KIDSDRIV    1   7394.8 7448.8  41.52 0.0000000001164003665 ***
## - EDUCATION   4   7403.1 7451.1  49.85 0.0000000003880952473 ***
## - CLM_FREQ    1   7399.5 7453.5  46.24 0.0000000000104530499 ***
## - TIF         1   7411.1 7465.1  57.78 0.0000000000000293011 ***
## - TRAVTIME    1   7418.2 7472.2  64.93 0.0000000000000007763 ***
## - MVR_PTS     1   7431.4 7485.4  78.15 < 0.00000000000000022 ***
## - CAR_USE     1   7438.2 7492.2  84.90 < 0.00000000000000022 ***
## - REVOKED     1   7451.3 7505.3  97.99 < 0.00000000000000022 ***
## - CAR_TYPE    5   7460.7 7506.7 107.42 < 0.00000000000000022 ***
## - URBANICITY  1   7969.5 8023.5 616.22 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=7408.41
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY
##
##              Df Deviance    AIC    LRT              Pr(>Chi)
```

```
## - YOJ         1   7358.4 7408.4   1.95            0.1624689
## <none>            7356.4 7408.4
## - HOMEKIDS    1   7360.4 7410.4   3.97            0.0463124 *
## - PARENT1     1   7365.5 7415.5   9.13            0.0025154 **
## - HOME_VAL    1   7369.7 7419.7  13.27            0.0002702 ***
## - OLDCLAIM    1   7370.1 7420.1  13.70            0.0002145 ***
## - INCOME      1   7375.3 7425.3  18.88 0.0000139532328069684 ***
## - BLUEBOOK    1   7381.3 7431.3  24.86 0.0000006166321554982 ***
## - MSTATUS     1   7395.7 7445.7  39.25 0.0000000003719155558 ***
## - KIDSDRIV    1   7398.6 7448.6  42.21 0.0000000000819740730 ***
## - CLM_FREQ    1   7402.7 7452.7  46.32 0.0000000000100404207 ***
## - EDUCATION   4   7419.7 7463.7  63.32 0.0000000000005816989 ***
## - TIF         1   7414.5 7464.5  58.14 0.0000000000000244423 ***
## - TRAVTIME    1   7422.2 7472.2  65.76 0.0000000000000005098 ***
## - MVR_PTS     1   7434.4 7484.4  77.99 < 0.00000000000000022 ***
## - REVOKED     1   7454.1 7504.1  97.70 < 0.00000000000000022 ***
## - CAR_TYPE    5   7462.8 7504.8 106.41 < 0.00000000000000022 ***
## - CAR_USE     1   7493.6 7543.6 137.23 < 0.00000000000000022 ***
## - URBANICITY  1   7978.1 8028.1 621.72 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=7408.36
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + TIF +
##     CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY
##
##              Df Deviance    AIC    LRT             Pr(>Chi)
## <none>            7358.4 7408.4
## - HOMEKIDS    1   7361.8 7409.8   3.39             0.065420 .
## - PARENT1     1   7367.7 7415.7   9.33             0.002257 **
## - OLDCLAIM    1   7372.3 7420.3  13.95             0.000188 ***
## - HOME_VAL    1   7372.6 7420.6  14.26             0.000159 ***
## - INCOME      1   7381.2 7429.2  22.85 0.0000017500735577702 ***
## - BLUEBOOK    1   7384.0 7432.0  25.66 0.0000004071455097257 ***
## - MSTATUS     1   7398.7 7446.7  40.38 0.0000000002095503663 ***
## - KIDSDRIV    1   7400.7 7448.7  42.33 0.0000000000769242452 ***
## - CLM_FREQ    1   7404.8 7452.8  46.40 0.0000000000096360291 ***
## - EDUCATION   4   7420.4 7462.4  62.02 0.0000000000010935142 ***
## - TIF         1   7417.0 7465.0  58.66 0.0000000000000187074 ***
## - TRAVTIME    1   7424.0 7472.0  65.60 0.0000000000000005519 ***
## - MVR_PTS     1   7437.1 7485.1  78.70 < 0.00000000000000022 ***
## - REVOKED     1   7456.2 7504.2  97.79 < 0.00000000000000022 ***
## - CAR_TYPE    5   7466.7 7506.7 108.33 < 0.00000000000000022 ***
## - CAR_USE     1   7496.2 7544.2 137.86 < 0.00000000000000022 ***
## - URBANICITY  1   7978.2 8026.2 619.84 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(log_step)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 +
```

```
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY, family = binomial(link = "logit"), data = df_insur_train[,
##     -2])
##
## Coefficients:
##                          Estimate    Std. Error z value            Pr(>|z|)
## (Intercept)          -2.4943816201  0.1886680456 -13.221 < 0.0000000000000002
## KIDSDRIV              0.6152611479  0.0942321692   6.529  0.0000000000661971
## HOMEKIDS              0.0614085035  0.0332242481   1.848             0.06456
## INCOME               -0.0000046464  0.0000009802  -4.740  0.00000213294947135
## PARENT11              0.3310327292  0.1084643456   3.052             0.00227
## HOME_VAL             -0.0000012485  0.0000003307  -3.775             0.00016
## MSTATUS1             -0.5283508951  0.0827299259  -6.386  0.00000000016977608
## EDUCATIONHigh School -0.0669243945  0.0917168894  -0.730             0.46558
## EDUCATIONBachelors   -0.5741862161  0.0980996788  -5.853  0.00000000482523813
## EDUCATIONMasters     -0.5701234987  0.1100689992  -5.180  0.00000022225272394
## EDUCATIONPhD         -0.5889033474  0.1486433507  -3.962  0.00007436981773162
## TRAVTIME              0.0151261670  0.0018707989   8.085  0.00000000000000062
## CAR_USEPrivate       -0.8543204269  0.0733065136 -11.654 < 0.0000000000000002
## BLUEBOOK             -0.0000235185  0.0000046906  -5.014  0.00000053327776419
## TIF                  -0.0549343106  0.0073058183  -7.519  0.00000000000005509
## CAR_TYPEPanel Truck   0.5339392183  0.1425502938   3.746             0.00018
## CAR_TYPEPickup        0.4958169971  0.0979949672   5.060  0.00000042009954806
## CAR_TYPESports Car    0.9523593176  0.1059348717   8.990 < 0.0000000000000002
## CAR_TYPESUV           0.7101142283  0.0849534961   8.359 < 0.0000000000000002
## CAR_TYPEVan           0.5977553781  0.1196852078   4.994  0.00000059020044738
## OLDCLAIM             -0.0000143406  0.0000038809  -3.695             0.00022
## CLM_FREQ              0.1945707002  0.0283752870   6.857  0.00000000000702981
## REVOKED1              0.9046333945  0.0906987432   9.974 < 0.0000000000000002
## MVR_PTS               0.1195618867  0.0135333831   8.835 < 0.0000000000000002
## URBANICITYUrban       2.3236895514  0.1119522262  20.756 < 0.0000000000000002
##
## (Intercept)          ***
## KIDSDRIV             ***
## HOMEKIDS             .
## INCOME               ***
## PARENT11             **
## HOME_VAL             ***
## MSTATUS1             ***
## EDUCATIONHigh School
## EDUCATIONBachelors   ***
## EDUCATIONMasters     ***
## EDUCATIONPhD         ***
## TRAVTIME             ***
## CAR_USEPrivate       ***
## BLUEBOOK             ***
## TIF                  ***
## CAR_TYPEPanel Truck  ***
## CAR_TYPEPickup       ***
## CAR_TYPESports Car   ***
## CAR_TYPESUV          ***
## CAR_TYPEVan          ***
## OLDCLAIM             ***
```

```
## CLM_FREQ             ***
## REVOKED1             ***
## MVR_PTS              ***
## URBANICITYUrban      ***
## ---
## 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: 7358.4  on 8136  degrees of freedom
## AIC: 7408.4
##
## Number of Fisher Scoring iterations: 5
```

- This model above gives us an AIC of 7408.36, indicating that this model is a better fit than the others based on having the lowest AIC. Backward Elimination leaves us with 17 variables including KIDSDRIV, HOMEKIDS, INCOME, PARENT, HOME_VAL, MSTATUS, EDUCATION, TRAVTIME, CAR_USE, BLUEBOOK, TIF, CAR_TYPE, OLDCLAIM, CLM_FREQ, REVOKED, MVR_PTS, and URBANICITY. As with the other models, the null model is outperformed as shown by the lower residual deviance compared to the null deviance.

The variables that positively impact the log odds of having car crash are the following:

- Kids driving
- Having kids at home (although this is a marginally significant p-value)
- Being a parent(vs not being a a parent)
- Having a longer travel time
- Having a car type other than minivan(when compared to minivan)
- Having an increased claims frequency
- Having a revoked license
- Residing in an urban environment
- Having more points on the drivers license

The variables that negatively impact the log odds of having car crash are the following:

- Having a higher income
- Having a higher home value
- Being married
- Having a college of graduate level education as opposed to having less than a high school level education(there is no difference between having a high school diploma and not having one)
- Using the car for private as opposed to commercial use
- Having a higher Bluebook value for your vehicle
- Having a longer tenure as insurance client
- Having longer period of times between claims

2.1.4 Assessing Model Performance

- We have selected the backward elimination model as our final Binary Logistic Regression Model for predicting a car crash given its better AIC. First, we will predict the probabilities of a car crash using the final backward step-wise regression model from which we will then call the predicted car crash based on the probability of 0.5.

```r
df_insur_train$log_pred_prob <- predict(log_step,
                                        newdata = df_insur_train[,-c(1:2)],
                                        type = "response")
df_insur_train$log_pred <- ifelse(df_insur_train$log_pred_prob > 0.5, 1, 0)
```

- Next we will assess model performance by calculating the area under the curve (AUC) for this model.

```r
pROC::auc(df_insur_train$TARGET_FLAG, df_insur_train$log_pred)
```

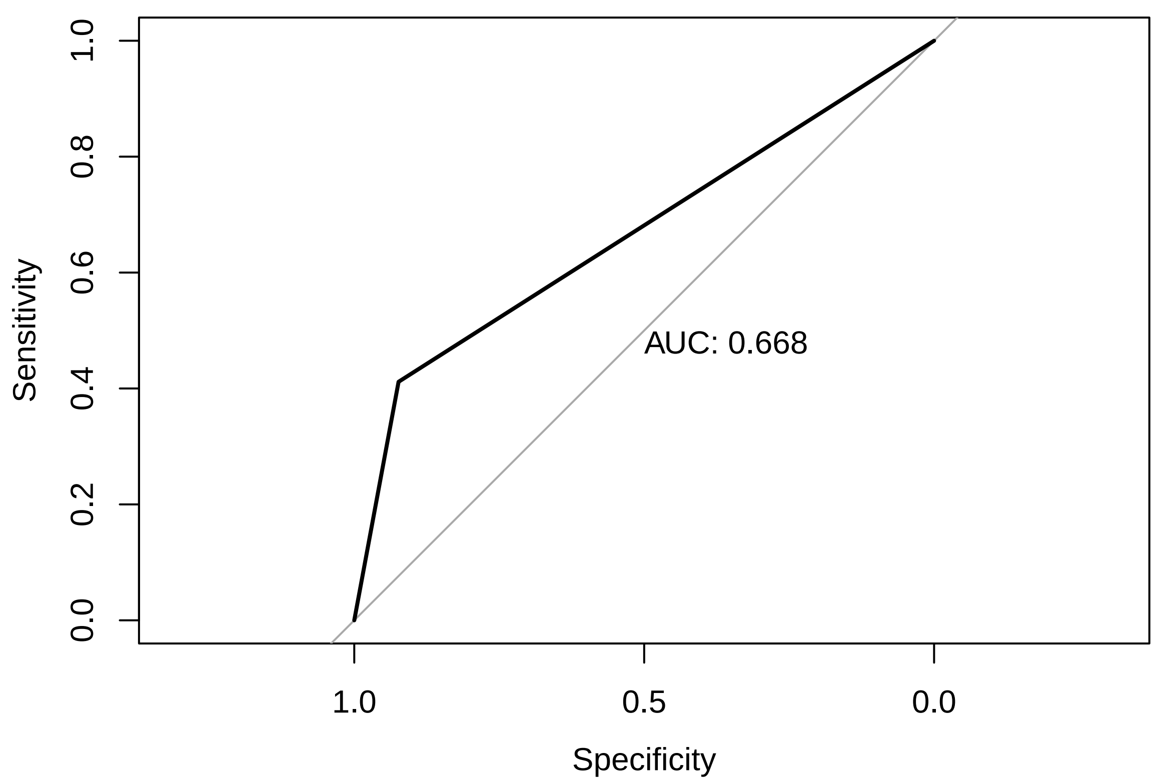
```
## Setting levels: control = 0, case = 1
```

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

```
## Area under the curve: 0.6676
```

```r
pROC::roc(df_insur_train$TARGET_FLAG~df_insur_train$log_pred,
          plot = TRUE, print.auc = TRUE)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



```
##
## Call:
## roc.formula(formula = df_insur_train$TARGET_FLAG ~ df_insur_train$log_pred,     plot = TRUE, print.a
##
## Data: df_insur_train$log_pred in 6008 controls (df_insur_train$TARGET_FLAG 0) < 2153 cases (df_insur
## Area under the curve: 0.6676
```

- The AUC of the model of .67 indicates that the model is only fair at predicting whether or not an insurance client will have a car crash.
- We can get a clearer sense of how the model under-performed by looking at a confusion matrix.

```
confusionMatrix(as.factor(df_insur_train$log_pred),
                as.factor(df_insur_train$TARGET_FLAG), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5549 1267
##          1  459  886
##
##                Accuracy : 0.7885
##                  95% CI : (0.7795, 0.7973)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.381
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.4115
##             Specificity : 0.9236
##          Pos Pred Value : 0.6587
##          Neg Pred Value : 0.8141
##              Prevalence : 0.2638
##          Detection Rate : 0.1086
##    Detection Prevalence : 0.1648
##       Balanced Accuracy : 0.6676
##
##        'Positive' Class : 1
##
```

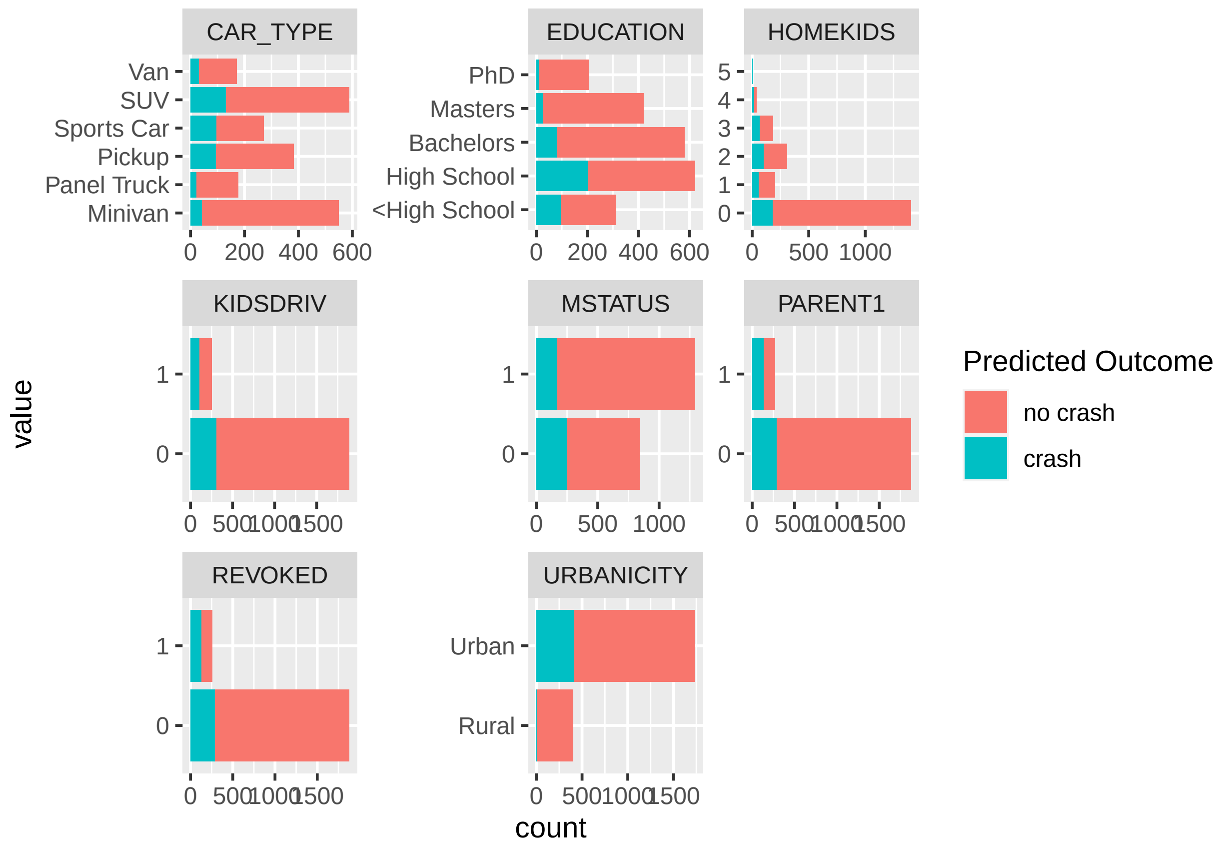
- After fitting the final logistic model to the train data the accuracy obtained is 78.9%, but the sensitivity is extremely low at only 41% thus the balance accuracy is the same as the AUC at 66.8%.
- It is worth noting that with such low sensitivity we can expect predictions to grossly under perform when predicting car crashes.

2.1.5 PREDICTING CAR CRASHES

- With the final logistic model, we will predict car crashes for the Evaluation data

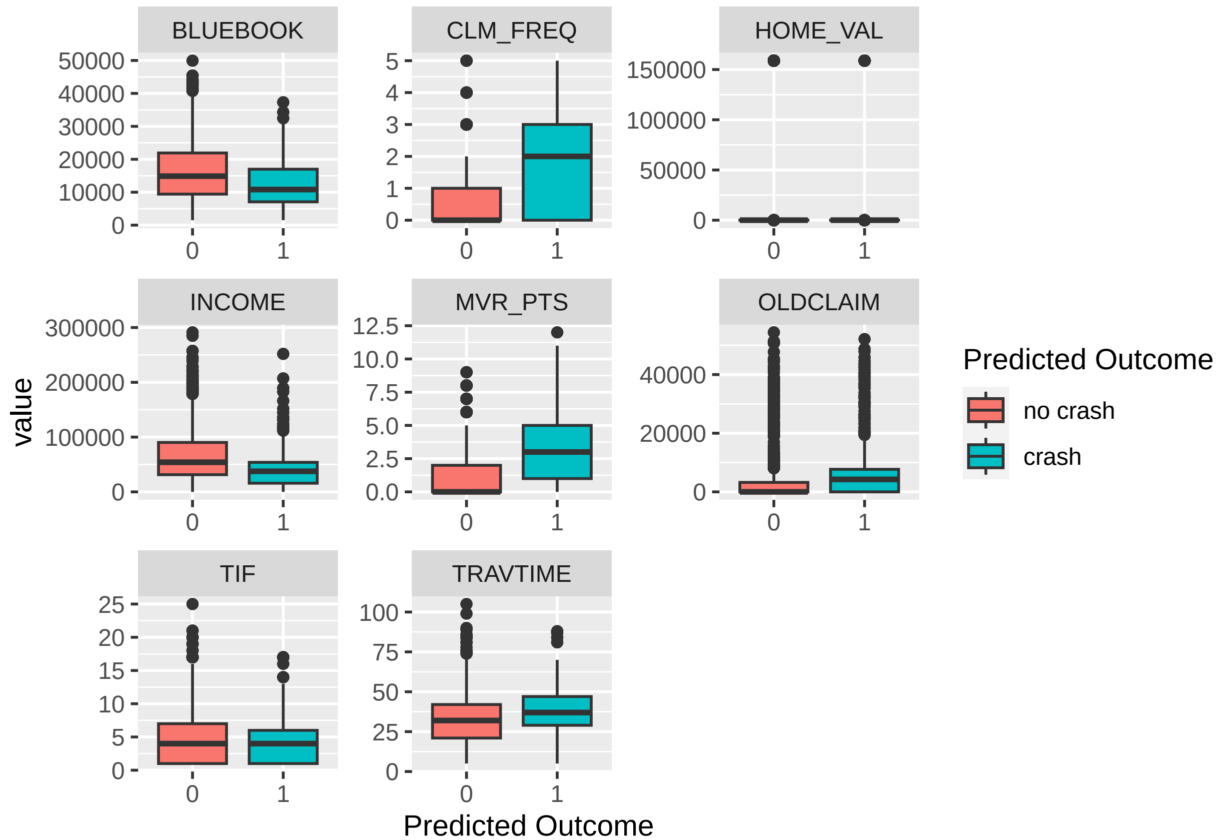
```
df_insur_eval$log_pred_prob <- predict(log_step,
                                       newdata = df_insur_eval[,-c(1:2)],
                                       type = "response")
df_insur_eval$log_pred <- ifelse(df_insur_eval$log_pred_prob > 0.5, 1, 0)
```

```
df_insur_eval %>%
  select(log_pred, KIDSDRIV, PARENT1, MSTATUS, EDUCATION, CAR_TYPE,
         REVOKED, URBANICITY, HOMEKIDS) %>%
  mutate_if(is.character, as.factor) %>%
  mutate_if(is.numeric, as.factor) %>%
  pivot_longer(-log_pred, names_to = "key", values_to = "value") %>%
  ggplot(aes(x = value, fill = log_pred)) +
  geom_bar() +
  scale_fill_discrete(labels = c("no crash", "crash"),
                      name = "Predicted Outcome") +
  coord_flip()+
  facet_wrap(~key, scales = "free")
```




```
df_insur_eval %>%
  select(log_pred, INCOME, HOME_VAL, TRAVTIME, TIF, OLDCLAIM,
         CLM_FREQ, MVR_PTS, BLUEBOOK) %>%
  pivot_longer(-log_pred, names_to = "key", values_to = "value") %>%
  ggplot(aes(y = value, x = as.factor(log_pred), fill = as.factor(log_pred))) +
  geom_boxplot() +
  scale_fill_discrete(labels = c("no crash", "crash"),
```

```
                    name = "Predicted Outcome") +
  xlab("Predicted Outcome")+
  facet_wrap(~key, scales = "free")
```



Assessing the predicted car crashes for the evaluation dataset, seems to largely reflect what was put into the model. Areas with stronger predictions were:

- Being a parent(vs not being a a parent)
- Having a longer travel time
- Having a car type other than minivan
- Having an increased claims frequency
- Having a revoked license
- Residing in an urban environment
- Having a lower Bluebook value for your vehicle

We do not see any change in the predicted car crashes with respect to the variable home values.

2.2 Multiple Linear Regression Models

2.2.1 Model with All Predictors - Adj.R-Squared 0.06476

- We will now be using Multiple Linear Regression to predict the cost if the person crashed their car (TARGET_AMT).

```r
mlr_mod <- lm(TARGET_AMT ~., data = df_insur_train[,-c(1,26:27)])
summary(mlr_mod)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = df_insur_train[, -c(1, 26:27)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5429  -1676   -767    317 104026
##
## Coefficients:
##                          Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)           379.8401734  459.5204283   0.827             0.408487
## KIDSDRIV              615.5559574  176.9377477   3.479             0.000506 ***
## AGE                     3.1116701    7.0161165   0.444             0.657414
## HOMEKIDS               70.0561331   64.3115145   1.089             0.276043
## YOJ                    -2.8171120   15.0824703  -0.187             0.851837
## INCOME                 -0.0054748    0.0017633  -3.105             0.001910 **
## PARENT11              526.9606719  202.4757919   2.603             0.009269 **
## HOME_VAL               -0.0004650    0.0005867  -0.792             0.428105
## MSTATUS1             -593.8842359  144.7638666  -4.102     0.00004128162293 ***
## SEXM                  344.1841461  183.0561065   1.880             0.060115 .
## EDUCATIONHigh School -128.7632680  168.9920488  -0.762             0.446113
## EDUCATIONBachelors   -375.4356181  190.2169545  -1.974             0.048447 *
## EDUCATIONMasters     -182.7961728  243.1457660  -0.752             0.452195
## EDUCATIONPhD         -165.1444043  296.7057633  -0.557             0.577821
## JOBNone              -212.5592004  207.7885821  -1.023             0.306358
## JOBWhite Collar      -206.5883013  161.9372944  -1.276             0.202087
## TRAVTIME               12.5849943    3.2229398   3.905     0.00009505912672 ***
## CAR_USEPrivate       -783.6326681  153.3427959  -5.110     0.00000032891363 ***
## BLUEBOOK                0.0139585    0.0086261   1.618             0.105666
## TIF                   -47.9581483   12.1832423  -3.936     0.00008340206639 ***
## CAR_TYPEPanel Truck   268.7765484  272.3486582   0.987             0.323729
## CAR_TYPEPickup        362.0271711  170.1993788   2.127             0.033444 *
## CAR_TYPESports Car    998.8533347  217.9020230   4.584     0.00000463073866 ***
## CAR_TYPESUV           732.0762393  179.3895411   4.081     0.00004528488721 ***
## CAR_TYPEVan           520.0497573  211.9636445   2.453             0.014169 *
## RED_CAR1              -56.2546948  149.1559536  -0.377             0.706069
## OLDCLAIM               -0.0111005    0.0074381  -1.492             0.135636
## CLM_FREQ              145.7559191   55.0675771   2.647             0.008140 **
## REVOKED1              574.2591546  173.5236173   3.309             0.000939 ***
## MVR_PTS               182.9110450   25.8904680   7.065     0.00000000000174 ***
## CAR_AGE               -26.9888035   12.8048265  -2.108             0.035087 *
## URBANICITYUrban      1543.4894649  136.9233069  11.273 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4549 on 8129 degrees of freedom
## Multiple R-squared:  0.06831,    Adjusted R-squared:  0.06476
## F-statistic: 19.23 on 31 and 8129 DF,  p-value: < 0.00000000000000022
```

```r
vif(mlr_mod)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## KIDSDRIV   1.305652  1        1.142651
## AGE        1.443709  1        1.201545
## HOMEKIDS   2.032282  1        1.425581
## YOJ        1.419854  1        1.191576
## INCOME     2.627261  1        1.620883
## PARENT1    1.851968  1        1.360870
## HOME_VAL   2.134691  1        1.461058
## MSTATUS    1.983930  1        1.408521
## SEX        3.286398  1        1.812842
## EDUCATION  3.394326  4        1.165049
## JOB        2.872971  2        1.301916
## TRAVTIME   1.036526  1        1.018099
## CAR_USE    2.164241  1        1.471136
## BLUEBOOK   2.079947  1        1.442202
## TIF        1.006338  1        1.003164
## CAR_TYPE   5.269027  5        1.180791
## RED_CAR    1.811504  1        1.345921
## OLDCLAIM   1.680564  1        1.296366
## CLM_FREQ   1.604631  1        1.266740
## REVOKED    1.276685  1        1.129905
## MVR_PTS    1.218472  1        1.103844
## CAR_AGE    1.969678  1        1.403452
## URBANICITY 1.202770  1        1.096709
```

- We can see the model with all the predictors, while overall significant, does a poor job in predicting cost as it can only account for 6.5% of the variability in the response variable TARGET_AMT. There are only 17 of 31 significant variable coefficients.
- The degree of freedom adjusted variance inflation factors suggests that there is no concerning collinearity because all of the values are less than 3.

2.2.2 Square Root Transformed Model - Adj.R-Squared 0.1698

- Let's see if a square root transformation of the numeric variables will make a better model

```r
df_train_sqrt <- (df_insur_train) %>%
  mutate(TARGET_AMT = sqrt(TARGET_AMT)) %>%
  mutate(KIDSDRIV = sqrt(KIDSDRIV)) %>%
  mutate(AGE = sqrt(AGE)) %>%
  mutate(HOMEKIDS = sqrt(HOMEKIDS)) %>%
  mutate(YOJ = sqrt(YOJ)) %>%
  mutate(INCOME = sqrt(INCOME)) %>%
  mutate(HOME_VAL = sqrt(HOME_VAL)) %>%
  mutate(TRAVTIME = sqrt(TRAVTIME)) %>%
  mutate(BLUEBOOK = sqrt(BLUEBOOK)) %>%
  mutate(TIF = sqrt(TIF)) %>%
  mutate(OLDCLAIM = sqrt(OLDCLAIM)) %>%
  mutate(CLM_FREQ = sqrt(CLM_FREQ)) %>%
  mutate(MVR_PTS = sqrt(MVR_PTS)) %>%
  mutate(CAR_AGE = sqrt(CAR_AGE))
```

```r
mlr_mod_sqrt <- lm(TARGET_AMT ~., data = df_train_sqrt[,-c(1,26:27)])
summary(mlr_mod_sqrt)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = df_train_sqrt[, -c(1, 26:27)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.462 -19.405  -7.954   9.724 293.953
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)          13.853795   5.252354   2.638             0.008365 **
## KIDSDRIV              6.414206   1.251113   5.127 0.000000301500709800 ***
## AGE                  -0.296283   0.645783  -0.459             0.646393
## HOMEKIDS              1.324440   0.774672   1.710             0.087363 .
## YOJ                  -0.125101   0.459887  -0.272             0.785609
## INCOME               -0.030274   0.005855  -5.171 0.000000238414922618 ***
## PARENT11              4.782634   1.453082   3.291             0.001001 **
## HOME_VAL             -0.005890   0.002174  -2.709             0.006758 **
## MSTATUS1             -5.715852   1.031786  -5.540 0.000000031234162843 ***
## SEXM                  1.732023   1.248930   1.387             0.165539
## EDUCATIONHigh School -0.010245   1.168480  -0.009             0.993004
## EDUCATIONBachelors   -4.806392   1.337684  -3.593             0.000329 ***
## EDUCATIONMasters     -3.432957   1.646090  -2.086             0.037053 *
## EDUCATIONPhD         -4.177555   1.984449  -2.105             0.035309 *
## JOBNone              -4.060981   1.587132  -2.559             0.010525 *
## JOBWhite Collar      -2.479942   1.117390  -2.219             0.026487 *
## TRAVTIME              1.633837   0.240946   6.781 0.000000000012775051 ***
## CAR_USEPrivate       -8.463489   1.055692  -8.017 0.000000000000001234 ***
## BLUEBOOK             -0.016241   0.014099  -1.152             0.249396
## TIF                  -2.596442   0.379572  -6.840 0.000000000008466679 ***
## CAR_TYPEPanel Truck   3.691483   1.826061   2.022             0.043255 *
## CAR_TYPEPickup        4.763567   1.172265   4.064 0.000048785006704621 ***
## CAR_TYPESports Car   10.141194   1.499602   6.763 0.000000000014492062 ***
## CAR_TYPESUV           7.580302   1.223825   6.194 0.000000000615298550 ***
## CAR_TYPEVan           5.355476   1.459132   3.670             0.000244 ***
## RED_CAR1             -0.267797   1.026393  -0.261             0.794168
## OLDCLAIM             -0.032479   0.011974  -2.712             0.006695 **
## CLM_FREQ              4.857657   0.886176   5.482 0.000000043413235383 ***
## REVOKED1              9.351468   1.181889   7.912 0.000000000000002859 ***
## MVR_PTS               3.377621   0.413587   8.167 0.000000000000000365 ***
## CAR_AGE              -0.507229   0.425990  -1.191             0.233804
## URBANICITYUrban      18.720732   0.946170  19.786 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.3 on 8129 degrees of freedom
## Multiple R-squared:  0.173,  Adjusted R-squared:  0.1698
## F-statistic: 54.84 on 31 and 8129 DF,  p-value: < 0.00000000000000022
```

- We can see transforming the numeric data using square root transformation did improve the model while remaining overall significant and increasing the adjusted r-squared from .0644 to .1698 and the number of significant variable coefficients from 17 to 23.

2.2.3 Backward Elimination - Adj.R-Squared .17

- Our next model will use the square root transformed dataset and backward elimination.

```
mlr_step <- step(mlr_mod_sqrt, direction = "backward", test = "F")
```

```
## Start:  AIC=56239.62
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df Sum of Sq     RSS   AIC  F value                Pr(>F)
## - RED_CAR     1        67 7964890 56238   0.0681             0.7941679
## - YOJ         1        73 7964896 56238   0.0740             0.7856088
## - AGE         1       206 7965029 56238   0.2105             0.6463926
## - BLUEBOOK    1      1300 7966123 56239   1.3269             0.2493962
## - CAR_AGE     1      1389 7966212 56239   1.4178             0.2338035
## - SEX         1      1884 7966708 56240   1.9232             0.1655391
## <none>                    7964823 56240
## - HOMEKIDS    1      2864 7967687 56241   2.9230             0.0873633 .
## - JOB         2      7558 7972382 56243   3.8571             0.0211676 *
## - HOME_VAL    1      7192 7972015 56245   7.3400             0.0067578 **
## - OLDCLAIM    1      7208 7972031 56245   7.3568             0.0066949 **
## - PARENT1     1     10614 7975438 56248  10.8331             0.0010012 **
## - EDUCATION   4     22213 7987036 56254   5.6676             0.0001493 ***
## - KIDSDRIV    1     25753 7990576 56264  26.2841 0.0000003015007098002 ***
## - INCOME      1     26199 7991022 56264  26.7391 0.0000002384149226180 ***
## - CLM_FREQ    1     29441 7994264 56268  30.0479 0.0000000434132353830 ***
## - MSTATUS     1     30069 7994892 56268  30.6890 0.0000000312341628429 ***
## - TRAVTIME    1     45052 8009876 56284  45.9810 0.0000000000127750508 ***
## - TIF         1     45847 8010670 56284  46.7917 0.0000000000084666792 ***
## - CAR_TYPE    5     62344 8027167 56293  12.7258 0.0000000000024134978 ***
## - REVOKED     1     61340 8026163 56300  62.6046 0.0000000000000028594 ***
## - CAR_USE     1     62974 8027798 56302  64.2724 0.0000000000000012343 ***
## - MVR_PTS     1     65347 8030171 56304  66.6943 0.0000000000000003647 ***
## - URBANICITY  1    383572 8348395 56621 391.4780 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=56237.69
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY
##
##              Df Sum of Sq     RSS   AIC  F value                Pr(>F)
## - YOJ         1        74 7964963 56236   0.0751             0.7840900
## - AGE         1       201 7965091 56236   0.2054             0.6504387
## - BLUEBOOK    1      1286 7966176 56237   1.3125             0.2519827
```

```
## - CAR_AGE     1      1394 7966284 56237   1.4227            0.2329927
## <none>                    7964890 56238
## - SEX         1      2027 7966917 56238   2.0689            0.1503636
## - HOMEKIDS    1      2858 7967748 56239   2.9175            0.0876622 .
## - JOB         2      7568 7972458 56241   3.8626            0.0210509 *
## - HOME_VAL    1      7159 7972049 56243   7.3074            0.0068815 **
## - OLDCLAIM    1      7217 7972107 56243   7.3665            0.0066589 **
## - PARENT1     1     10630 7975520 56247  10.8502            0.0009921 ***
## - EDUCATION   4     22266 7987156 56252   5.6818            0.0001455 ***
## - KIDSDRIV    1     25823 7990713 56262  26.3582 0.0000002901833481936 ***
## - INCOME      1     26189 7991079 56262  26.7319 0.0000002392999730515 ***
## - CLM_FREQ    1     29434 7994324 56266  30.0445 0.0000000434891177571 ***
## - MSTATUS     1     30063 7994953 56266  30.6859 0.0000000312829032034 ***
## - TRAVTIME    1     45016 8009906 56282  45.9491 0.0000000000129840127 ***
## - TIF         1     45833 8010723 56283  46.7829 0.0000000000085042965 ***
## - CAR_TYPE    5     62447 8027337 56291  12.7483 0.0000000000022882696 ***
## - REVOKED     1     61346 8026235 56298  62.6172 0.0000000000000028413 ***
## - CAR_USE     1     62959 8027849 56300  64.2639 0.0000000000000012396 ***
## - MVR_PTS     1     65347 8030237 56302  66.7014 0.0000000000000003634 ***
## - URBANICITY  1    383505 8348395 56619 391.4554 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=56235.76
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df Sum of Sq     RSS   AIC  F value               Pr(>F)
## - AGE         1       232 7965195 56234   0.2366            0.6266976
## - BLUEBOOK    1      1286 7966249 56235   1.3124            0.2519914
## - CAR_AGE     1      1395 7966359 56235   1.4242            0.2327548
## <none>                    7964963 56236
## - SEX         1      2020 7966983 56236   2.0619            0.1510553
## - HOMEKIDS    1      2788 7967752 56237   2.8463            0.0916226 .
## - JOB         2      7543 7972507 56239   3.8503            0.0213116 *
## - HOME_VAL    1      7115 7972079 56241   7.2635            0.0070515 **
## - OLDCLAIM    1      7246 7972210 56241   7.3973            0.0065460 **
## - PARENT1     1     10660 7975623 56245  10.8819            0.0009752 ***
## - EDUCATION   4     22201 7987164 56250   5.6659            0.0001498 ***
## - KIDSDRIV    1     25931 7990894 56260  26.4712  0.000000273742834028 ***
## - CLM_FREQ    1     29467 7994431 56264  30.0816  0.000000042668414519 ***
## - INCOME      1     29735 7994699 56264  30.3553  0.000000037072324760 ***
## - MSTATUS     1     30487 7995450 56265  31.1223  0.000000025004706910 ***
## - TRAVTIME    1     45004 8009967 56280  45.9418  0.000000000013031901 ***
## - TIF         1     45887 8010851 56281  46.8441  0.000000000008244414 ***
## - CAR_TYPE    5     62543 8027507 56290  12.7694  0.000000000002176563 ***
## - REVOKED     1     61399 8026362 56296  62.6789  0.000000000000002754 ***
## - CAR_USE     1     63030 8027993 56298  64.3438  0.000000000000001191 ***
## - MVR_PTS     1     65469 8030433 56301  66.8339  0.000000000000000340 ***
## - URBANICITY  1    383487 8348451 56618 391.4812 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Step:  AIC=56234
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df Sum of Sq     RSS   AIC  F value                Pr(>F)
## - CAR_AGE     1      1393 7966589 56233   1.4226             0.2330133
## - BLUEBOOK    1      1456 7966652 56233   1.4869             0.2227296
## - SEX         1      1913 7967108 56234   1.9532             0.1622826
## <none>                    7965195 56234
## - HOMEKIDS    1      4200 7969395 56236   4.2875             0.0384250 *
## - JOB         2      7555 7972750 56238   3.8564             0.0211820 *
## - OLDCLAIM    1      7251 7972446 56239   7.4028             0.0065261 **
## - HOME_VAL    1      7279 7972474 56239   7.4314             0.0064232 **
## - PARENT1     1     10782 7975977 56243  11.0075             0.0009114 ***
## - EDUCATION   4     22332 7987527 56249   5.6999             0.0001407 ***
## - KIDSDRIV    1     26015 7991210 56259  26.5593 0.0000002615772006256 ***
## - CLM_FREQ    1     29429 7994624 56262  30.0448 0.0000000434816429346 ***
## - INCOME      1     29739 7994935 56262  30.3622 0.0000000369401155785 ***
## - MSTATUS     1     30781 7995976 56263  31.4254 0.0000000214029508401 ***
## - TRAVTIME    1     44974 8010169 56278  45.9154 0.0000000000132071553 ***
## - TIF         1     45910 8011105 56279  46.8716 0.0000000000081298474 ***
## - CAR_TYPE    5     62322 8027517 56288  12.7254 0.0000000000024156422 ***
## - REVOKED     1     61510 8026705 56295  62.7984 0.0000000000000025933 ***
## - CAR_USE     1     63161 8028356 56296  64.4836 0.0000000000000011097 ***
## - MVR_PTS     1     65896 8031092 56299  67.2765 0.0000000000000002721 ***
## - URBANICITY  1    383749 8348945 56616 391.7858 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=56233.43
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY
##
##              Df Sum of Sq     RSS   AIC  F value                Pr(>F)
## - BLUEBOOK    1      1440 7968029 56233   1.4704             0.2253227
## <none>                    7966589 56233
## - SEX         1      1970 7968559 56233   2.0113             0.1561746
## - HOMEKIDS    1      4301 7970890 56236   4.3910             0.0361598 *
## - JOB         2      7612 7974201 56237   3.8855             0.0205758 *
## - HOME_VAL    1      7121 7973710 56239   7.2701             0.0070255 **
## - OLDCLAIM    1      7225 7973814 56239   7.3764             0.0066225 **
## - PARENT1     1     10713 7977302 56242  10.9370             0.0009467 ***
## - KIDSDRIV    1     25951 7992540 56258  26.4932 0.0000002706527177044 ***
## - EDUCATION   4     33600 8000189 56260   8.5756 0.0000006671031739280 ***
## - CLM_FREQ    1     29413 7996001 56262  30.0272 0.0000000438775939518 ***
## - INCOME      1     29977 7996565 56262  30.6029 0.0000000326456470939 ***
## - MSTATUS     1     30979 7997568 56263  31.6263 0.0000000193068124134 ***
## - TRAVTIME    1     44945 8011534 56277  45.8843 0.0000000000134178566 ***
## - TIF         1     45879 8012468 56278  46.8377 0.0000000000082710017 ***
```

```
## - CAR_TYPE    5     62736 8029324 56287  12.8092 0.0000000000019810168 ***
## - REVOKED     1     61421 8028009 56294  62.7037 0.0000000000000027199 ***
## - CAR_USE     1     63237 8029825 56296  64.5576 0.0000000000000010691 ***
## - MVR_PTS     1     65611 8032200 56298  66.9818 0.0000000000000003156 ***
## - URBANICITY  1    383803 8350392 56615 391.8203 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=56232.9
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + TIF +
##     CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY
##
##              Df Sum of Sq     RSS   AIC  F value                Pr(>F)
## <none>                    7968029 56233
## - SEX         1      4134 7972163 56235   4.2205             0.0399693 *
## - HOMEKIDS    1      4606 7972635 56236   4.7020             0.0301559 *
## - JOB         2      7453 7975482 56237   3.8040             0.0223207 *
## - HOME_VAL    1      7192 7975221 56238   7.3420             0.0067503 **
## - OLDCLAIM    1      7307 7975336 56238   7.4596             0.0063235 **
## - PARENT1     1     10634 7978663 56242  10.8558             0.0009891 ***
## - KIDSDRIV    1     25626 7993655 56257  26.1597 0.0000003214961541581 ***
## - EDUCATION   4     34337 8002366 56260   8.7630 0.0000004687495182697 ***
## - CLM_FREQ    1     29670 7997699 56261  30.2877 0.0000000383810359849 ***
## - MSTATUS     1     31094 7999123 56263  31.7413 0.0000000182006605574 ***
## - INCOME      1     32808 8000837 56264  33.4913 0.0000000074236023003 ***
## - TRAVTIME    1     44843 8012872 56277  45.7774 0.0000000000141656281 ***
## - TIF         1     45850 8013879 56278  46.8052 0.0000000000084084597 ***
## - REVOKED     1     61588 8029617 56294  62.8711 0.0000000000000024999 ***
## - CAR_USE     1     62953 8030982 56295  64.2644 0.0000000000000012392 ***
## - CAR_TYPE    5     73566 8041595 56298  15.0196 0.0000000000000104037 ***
## - MVR_PTS     1     65976 8034005 56298  67.3503 0.0000000000000002622 ***
## - URBANICITY  1    383603 8351632 56615 391.5934 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mlr_final <- lm(TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL +
    MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + TIF +
    CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY,
              data = df_train_sqrt[,-c(1, 26:27)])
```

```r
summary(mlr_final)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY, data = df_train_sqrt[, -c(1, 26:27)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.881 -19.468  -7.962   9.931 294.422
```

```
##
## Coefficients:
##                        Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            8.616117   2.731610   3.154            0.001615 **
## KIDSDRIV               6.258667   1.223674   5.115 0.000000321496154159 ***
## HOMEKIDS               1.515918   0.699090   2.168            0.030156 *
## INCOME                -0.031913   0.005514  -5.787 0.000000007423602300 ***
## PARENT11               4.783065   1.451695   3.295            0.000989 ***
## HOME_VAL              -0.005873   0.002167  -2.710            0.006750 **
## MSTATUS1              -5.787889   1.027324  -5.634 0.000000018200660557 ***
## SEXM                   2.062629   1.004008   2.054            0.039969 *
## EDUCATIONHigh School -0.125606   1.163795  -0.108            0.914055
## EDUCATIONBachelors    -5.414097   1.245981  -4.345 0.000014080780459686 ***
## EDUCATIONMasters      -4.437174   1.429185  -3.105            0.001911 **
## EDUCATIONPhD          -5.201762   1.797541  -2.894            0.003816 **
## JOBNone               -3.918855   1.530107  -2.561            0.010450 *
## JOBWhite Collar       -2.477286   1.117029  -2.218            0.026600 *
## TRAVTIME               1.629848   0.240891   6.766 0.000000000014165628 ***
## CAR_USEPrivate        -8.459111   1.055211  -8.017 0.000000000000001239 ***
## TIF                   -2.596280   0.379494  -6.841 0.000000000008408460 ***
## CAR_TYPEPanel Truck    2.897241   1.683812   1.721            0.085353 .
## CAR_TYPEPickup         4.947122   1.162295   4.256 0.000021013365578203 ***
## CAR_TYPESports Car    10.714682   1.406493   7.618 0.000000000000028657 ***
## CAR_TYPESUV            8.087984   1.137194   7.112 0.000000000001238681 ***
## CAR_TYPEVan            5.018210   1.420752   3.532            0.000415 ***
## OLDCLAIM              -0.032692   0.011970  -2.731            0.006323 **
## CLM_FREQ               4.875196   0.885848   5.503 0.000000038381035985 ***
## REVOKED1               9.368501   1.181530   7.929 0.000000000000002500 ***
## MVR_PTS                3.389298   0.412991   8.207 0.000000000000000262 ***
## URBANICITYUrban       18.718122   0.945899  19.789 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.3 on 8134 degrees of freedom
## Multiple R-squared:  0.1726, Adjusted R-squared:   0.17
## F-statistic: 65.27 on 26 and 8134 DF,  p-value: < 0.00000000000000022
```

- We are selecting this backward elimination model on square root transformed data as our final Multiple Linear Regression Model as it has the greatest effect size (.17), is less complex given the smaller degrees of freedom (26), and has 24 of 26 significant variable coefficients.

The variables that positively impact the average cost of having car crash are the following:

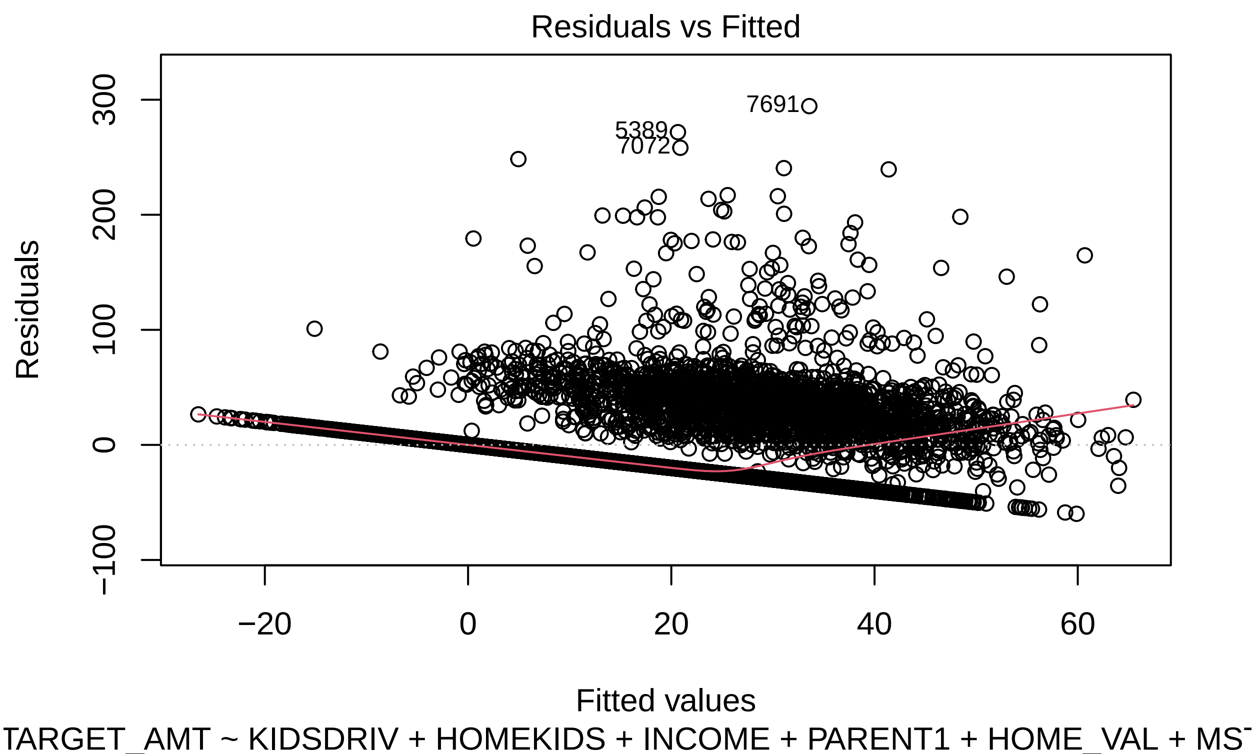
- Kids driving
- Being male
- Being a parent(vs not being a a parent)
- Having a longer travel time
- Having a car type other than minivan(when compared to minivan)
- Having an increased claims frequency
- Having white collar job
- Having a revoked license
- Residing in an urban environment
- Having higher points on drivers license

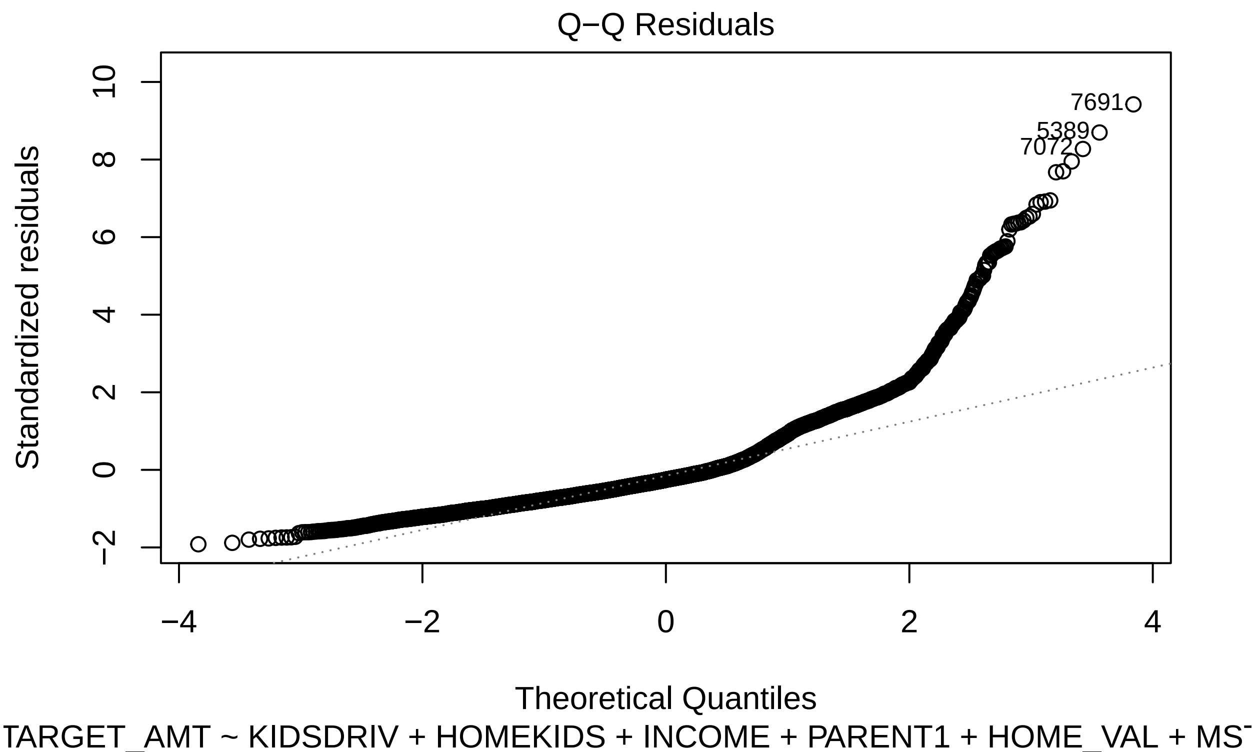
The variable that negatively impact the average cost of having crash are the following:

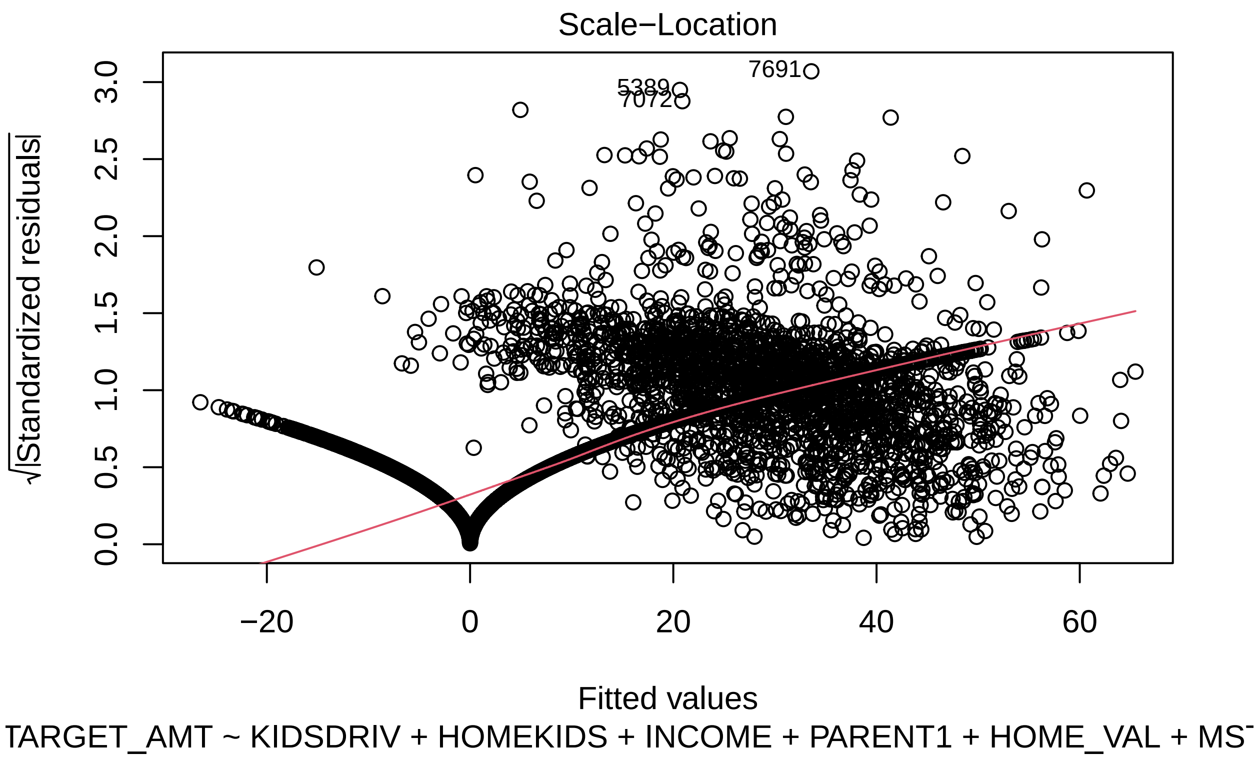
- Having a higher income
- Being married
- Education beyond high school
- Using the car for private as opposed to commercial use
- Having a longer tenure as insurance client
- Having an older car

2.2.4 Test Model Assumptions

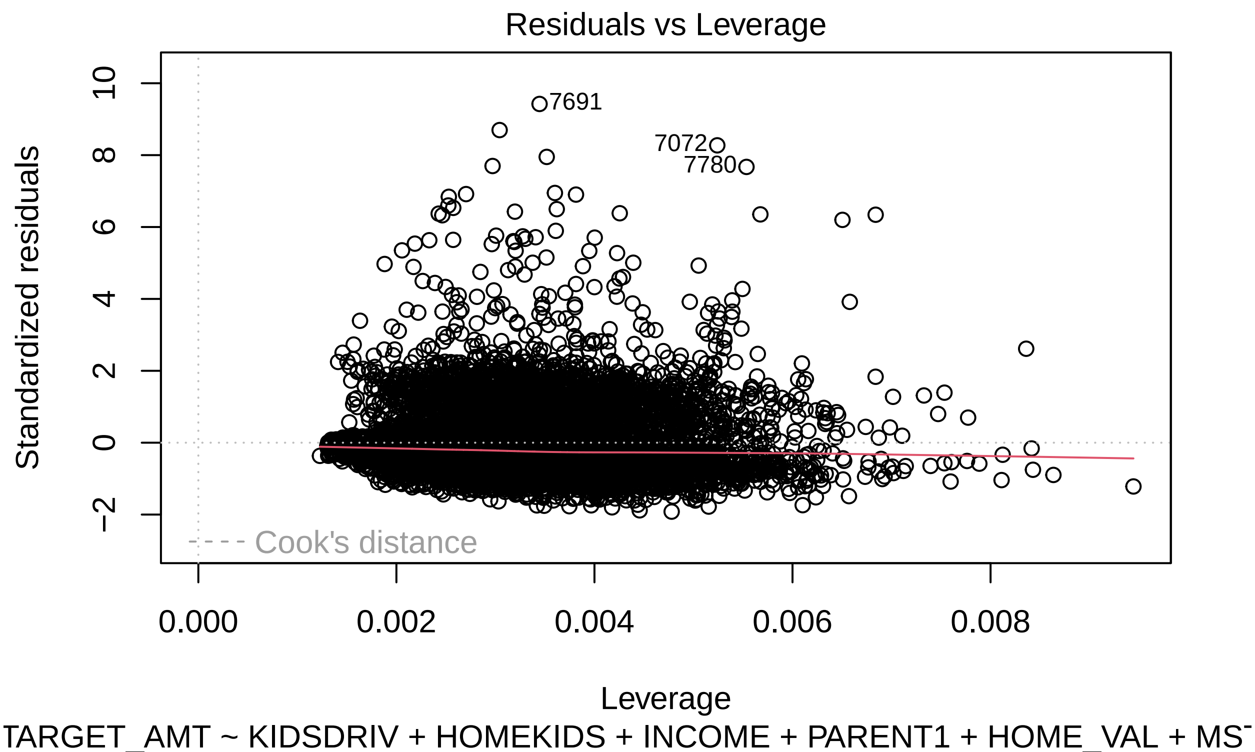
```
plot(mlr_final)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL + MS
7691
5389
7072
10
8
6
4
2
0
−2
−4
−2
0
2
4



Scale−Location
√|Standardized residuals|
Fitted values
7691
5389
7072
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−20
0
20
40
60
TARGET_AMT ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + HOME_VAL + MST



1. Linearity - the first plot shows that the relationship between target amount and the predictor variables in the final model is linear, so the assumption is net.

2. Normality - the second plot shows that our assumption of approximate normal distribution of the residuals may not be met due to the tails especially the right.

3. Equality of Variances - the third plot that there is unequal variance, however, the relationship is largely homoscedastic.

4. Leverage / High Influence - the fourth plot shows that there are a few outliers with very high claim

Thus we should be cautious of this model given these issues with our model assumptions

2.2.5 Assessing Model Performance

- We are going to first predict the amount of the crash using the final model, we will then calculate the RMSE using the predictions.

```
df_train_sqrt$mlr_pred <- predict(mlr_final,
                                  newdata = df_train_sqrt[,-c(1:2, 26:27)],
                                  type = "response")
RMSE(df_train_sqrt$mlr_pred, df_train_sqrt$TARGET_AMT)
```

```
## [1] 31.24667
```

```r
summary(mlr_final)$adj.r.squared
```

```
## [1] 0.1699741
```

- The RMSE for this model suggest an average deviation in the square root transformed predicted claim amount from the true claim amount of 31.2, which squared is 973.44. This suggests that the model is not doing a particularly good job at predicting accurate claim amounts. This is not surprising given that the R squared of the final model could only explain 17% of the total variation in the claim amount.

2.2.6 PREDICTING AMOUNT OF CLAIM

- With the final Multiple Linear Regression model, we will predict the amount of the claim for car crashes for the Evaluation dataset after performing the square root transformations.

```r
df_eval_sqrt <- (df_insur_eval) %>%
  mutate(TARGET_AMT = sqrt(TARGET_AMT)) %>%
  mutate(KIDSDRIV = sqrt(KIDSDRIV)) %>%
  mutate(AGE = sqrt(AGE)) %>%
  mutate(HOMEKIDS = sqrt(HOMEKIDS)) %>%
  mutate(YOJ = sqrt(YOJ)) %>%
  mutate(INCOME = sqrt(INCOME)) %>%
  mutate(HOME_VAL = sqrt(HOME_VAL)) %>%
  mutate(TRAVTIME = sqrt(TRAVTIME)) %>%
  mutate(BLUEBOOK = sqrt(BLUEBOOK)) %>%
  mutate(TIF = sqrt(TIF)) %>%
  mutate(OLDCLAIM = sqrt(OLDCLAIM)) %>%
  mutate(CLM_FREQ = sqrt(CLM_FREQ)) %>%
  mutate(MVR_PTS = sqrt(MVR_PTS)) %>%
  mutate(CAR_AGE = sqrt(CAR_AGE))
```

```r
df_eval_sqrt$mlr_pred <- predict(mlr_final,
                                  newdata = df_eval_sqrt[,-c(1:2, 26:27)],
                                  type = "response")

## Update TARGET_FLAG with the predicted values from our binary logistic regression
df_eval_sqrt$TARGET_FLAG <- df_insur_eval$log_pred

## Convert MLR prediction into original unit by squaring the values
df_eval_sqrt$TARGET_AMT <- ifelse(df_eval_sqrt$TARGET_FLAG ==1, df_eval_sqrt$mlr_pred^2, 0)
```

```r
coef(mlr_final)
```

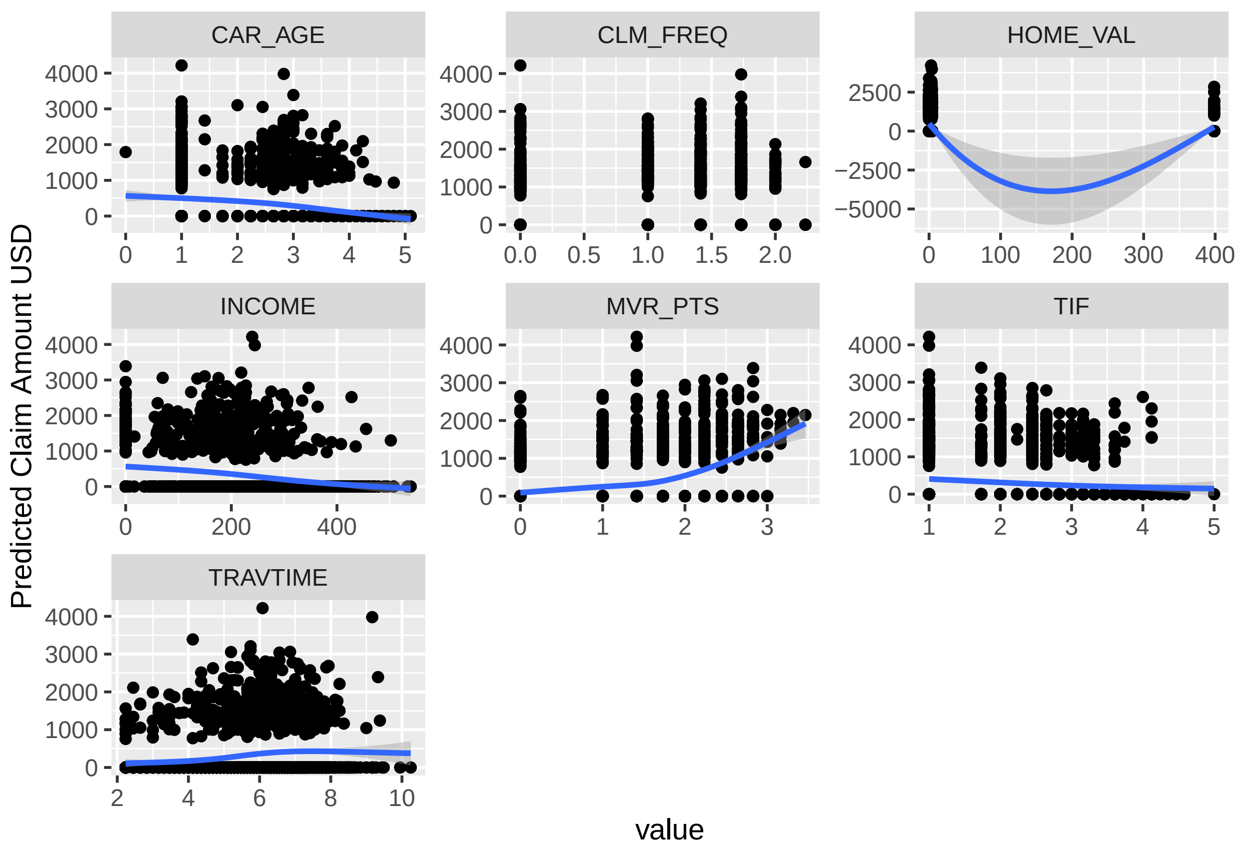
```
##          (Intercept)             KIDSDRIV             HOMEKIDS
##          8.616117096          6.258666667          1.515917951
##               INCOME             PARENT11             HOME_VAL
##         -0.031912502          4.783064529         -0.005872758
##             MSTATUS1                 SEXM EDUCATIONHigh School
##         -5.787888864          2.062629449         -0.125606246
##   EDUCATIONBachelors      EDUCATIONMasters         EDUCATIONPhD
##         -5.414097421         -4.437173781         -5.201762417
##              JOBNone      JOBWhite Collar             TRAVTIME
```

```
##         -3.918854875         -2.477286041          1.629847640
##        CAR_USEPrivate                  TIF  CAR_TYPEPanel Truck
##         -8.459110951         -2.596280182          2.897241323
##        CAR_TYPEPickup   CAR_TYPESports Car          CAR_TYPESUV
##          4.947121997         10.714681674          8.087983835
##          CAR_TYPEVan             OLDCLAIM             CLM_FREQ
##          5.018209946         -0.032692478          4.875196135
##             REVOKED1              MVR_PTS       URBANICITYUrban
##          9.368500916          3.389298381         18.718122409
```

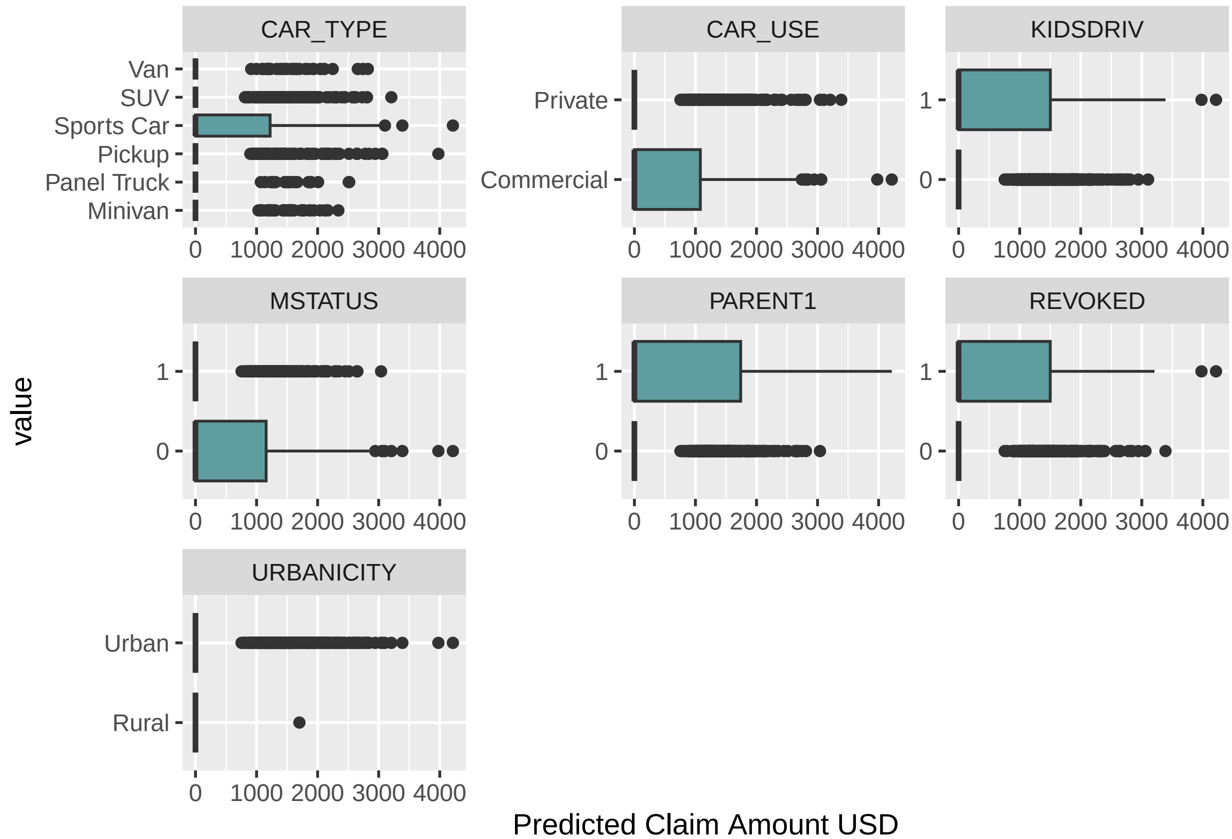
```r
df_eval_sqrt %>%
  select(TARGET_AMT, INCOME, HOME_VAL, TRAVTIME, TIF, CLM_FREQ, MVR_PTS,
         CAR_AGE) %>%
  mutate_if(is.character, as.numeric) %>%
  pivot_longer(-TARGET_AMT, names_to = "key", values_to = "value") %>%
  ggplot(aes(x = value, y = TARGET_AMT)) +
  geom_point() +
  geom_smooth() +
  ylab("Predicted Claim Amount USD")+
  facet_wrap(~key, scales = "free")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Computation failed in `stat_smooth()`
## Caused by error in `smooth.construct.cr.smooth.spec()`:
## ! x has insufficient unique values to support 10 knots: reduce k.
```



```r
df_eval_sqrt %>%
  select(TARGET_AMT, KIDSDRIV, MSTATUS, CAR_TYPE, REVOKED, URBANICITY, CAR_USE,
         PARENT1) %>%
  mutate_if(is.numeric, as.character) %>%
  pivot_longer(-TARGET_AMT, names_to = "key", values_to = "value") %>%
  mutate(TARGET_AMT = as.numeric(TARGET_AMT)) %>%
  ggplot(aes(x = TARGET_AMT, y = value)) +
  geom_boxplot(fill = "cadetblue") +
  xlab("Predicted Claim Amount USD") +
  facet_wrap(~key, scales = "free")
```




```r
table(df_eval_sqrt$TARGET_FLAG)
```

```
##
##    0    1
## 1727  414
```

```r
sum(df_eval_sqrt$TARGET_AMT)
```

```
## [1] 663689.4
```

2.3 Conclusion

- Using our binary logistic & multiple linear regression models, we predict that 414 of 1727 cases will have a car crash, which will amount to $663,690. However, given the under performance of the models on the training data and potential test assumption violations, we would be very cautious in using these predictions until additional variables & transformations could better improve the models.